Lecture Notes in Earth System Sciences

Fernando Sansò
Michael G. Sideris *Editors*

Geoid Determination

Theory and Methods

Springer

Lecture Notes in Earth System Sciences

For further volumes:
http://www.springer.com/series/10529

Fernando Sansò • Michael G. Sideris
Editors

Geoid Determination

Theory and Methods

Springer

Editors
Fernando Sansò
School of Civil and
Environmental Engineering
Politecnico Milano
Como
Italy

Michael G. Sideris
Dept. Geomatics Engineering
University of Calgary
Calgary Alberta
Canada

ISBN 978-3-540-74699-7 ISBN 978-3-540-74700-0 (eBook)
DOI 10.1007/978-3-540-74700-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940200

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

"to the International Association of Geodesy
in the 150th Anniversary of its foundation"

Foreword

"... the present book is intended to be theoretical in the sense in which the word is used in the term theoretical physics" from the preface of Physical Geodesy

by W.A. Heiskanen and H. Moritz

In the year 1994 the International Geoid Service, on behalf of the International Association of Geodesy, has organized and given in Milan the first course of the International School for the Determination and Use of the Geoid. The purpose was to gain momentum in spreading worldwide that part of the geodetic scientific culture which is known as physical geodesy, namely the theory of the determination of the potential of the gravity field of the earth.

Since 1994 other nine courses of the school have been run in Rio de Janeiro, Milan, Johor, Thessaloniki, Budapest, Kopenhagen, Como, La Plata, St. Petersburg. A large number of students from all over the world, in fact 307, have attended the school and after that they have actively joined the international geodetic community, so that we can say that the concept has proved to be fruitful. The courses have been organized in a quite regular fashion with 1 day of introductory theory and 4 days of explanations and labs exercises to get trained in the use of the software relevant to different specific items.

For the purpose of effectiveness the school was endowed with lecture notes where both theory and applications were supplied. They constitute the first core of the present book.

The teachers configuration had a turnover in the years, yet all the authors of the book have been teachers at some of the courses.

The full group of teachers has included: O. Andersen, R. Barzaghi, R. Forsberg, G. Fotopoulos, W. Kearsley, N. Pavlis, R. Rapp, F. Sansò, P. Schwintzer, M. Sideris, C.C. Tscherning, I. Tziavos, H.G. Wenzel.

The geoid, which is also mentioned in the title of the book, is plainly an equipotential surface of the gravity field of the earth, identified by a conventional

value of the potential, such that it runs close to the surface of the ocean, within meters, but then well inside the continental masses specially in mountainous areas.

As such the geoid is a geometrical entity, usually described by the height of its points over the earth ellipsoid, the so called geoid undulation.

This in *turn* has become nowadays an important piece of knowledge for a number of scientific and technological applications; here we mention only two, namely the analysis of the oceanic flow which is related to the average sea surface height with respect to the geoid and the transformation of ellipsoidal heights, determined by ubiquitous GNSS techniques, into the more physically meaningful orthometric heights, i.e. the heights above the geoid.

From the point of view of the determination of the gravity field, knowing the geoid and the mass distribution above it is a sufficient information to compute the gravity potential and all derived quantities throughout the whole space, outside the geoid itself.

The problem of how to deal with the mass distribution above the geoid has historically produced two different lines of thought in Geodesy.

One dates back to Helmert (1884), who further developed the ingenious ideas of Stokes (1849), assuming that the mass distribution is know. These ideas are still pursued by a number of modern authors among which we mention only B. Heck, Z. Martinec and L. Sjoberg. The second line of thought, known as the theory, can be traced back to the seminal monography of Molodensky-Eremeev-Yourkina (1962). In this case it is the surface of the earth and not the geoid to be directly determined.

This is based on the calculation of the separation of the earth surface to a much closer one, the telluroid, actually determined from surface data only by applying a rigorous linearization of the so called geodetic boundary value problem. In fact it is show that the determination of the geometric quantity "separation between earth surface and telluroid", the so called height anomaly, has once more to be done by simultaneously solving for the potential of the gravity field.

The Molodensky concept is basically that the gravity field outside the masses can be fully computed from data taken on the surface only. From the modern mathematical point of view this is an early formulation of a so called free boundary, boundary value problem. It is after the determination of the surface of the earth has been achieved, that one can then put the problem of its downward continuation inside the masses until the geoid is derived.

In this way the problem is split basically into two steps. The first is the direct determination of the earth surface through the solution of a free boundary value problem, which is a well posed problem even in its general non linear formulation as shown by authors like L. Hormander, P. Holota and F. Sansò. The second step is then the approximate solution of an improperly posed problem, requiring the knowledge of the mass distribution too. Along this line many modern authors have been working among which we want to mention only H-Moritz and T. Krarup, who not only developed the modern mathematical foundations of this theory but also provided a quite original approach to the computation of approximate solutions, borrowing methods from the theory of random fields, known in physical geodesy under the name of collocation.

It is on this foundation that the book builds the modern approach to the determination of the geoid and more generally to the solution of the main problems of physical geodesy.

In this respect we could say that the book is not comprehensive, since the Helmert line is not covered in the text.

On the other hand this book, in continuation with the most classical text of physical geodesy by W. Heiskanen and H. Moritz, covers and develops the material of the other text of "Advanced Physical Geodesy", by H. Moritz. With the purpose of addressing students with a basic background in mathematics and physics, the book in Part I builds its own tools and theory from ground level. In order to avoid interrupting the logical line of thinking, when appropriate some more technical mathematical proofs are delayed to appendices at the end of each chapter.

In the second part various methods are illustrated with reference to specific applications and with fully developed examples, together with the explanation of the current solution to the relevant numerical problems.

In part three the modern mathematical foundation of Molodensky's theory and the most recent theoretical achievements, at least for the linearized formulation, are presented for students at advanced level that want to go deeper into the subject. Once more all the material is created starting from ground without presupposing a higher level of mathematics.

During the long period of preparation of the book many events happened which are changing the environment of geodetic theory and methods. Just to mention two of them, on the one hand a global gravity model has been established, EGM08, with a resolution of about 10 km on the earth surface and with an unprecedented accuracy; on the other hand dedicated gravity satellite missions, like GRACE and GOCE are flying, that are still at work providing new data sets requiring a significant development of the tools for the combination of different gravity models.

Some of these items are fully included into the book, for instance the calculation and use of EGM08 model; some others are only contingently touched. As a matter of fact we are leaving now in an epoch where limits of the present theory and methods start showing here and there.

Indeed research in physical geodesy is, hopefully, a never ending story so this is a challenge for future work and future books.

Fernando Sansò
Honorary President of the
International Association of Geodesy

Acknowledgements

We would like to acknowledge the past presidents of the International Association of Geodesy, who fostered the Geoid School in the last years. We would especially like to acknowledge the support of the past President Klaus Peter Schwarz († 2012).

We also wish to mention the following teachers who contributed to the school, delivering several versions of the courses: R.H. Rapp, R. Forsberg, W. Kearsley, H.G. Wenzel († 1999), P. Schwintzer († 2004) and R. Barzaghi. Though they have not participated directly in the drafting of this book, they have contributed indirectly by means of their scientific knowledge and lecturing skills.

Finally a personal sincere thank is due by the writer to Cristina Giannetto for the competent typing of the manuscript of Part I and III and to Fausto Sacerdote and Christian Tschering for careful proof reading.

Contents

List of Contributors

O.B. Andersen Geodetic Department, DTU-Space, Elektrovej, DK-2800, Denmark, oa@space.dtu.dk

Georgia Fotopoulos The University of Texas at Dallas, Geosciences, Richardson, Texas, USA, foto@utdallas.edu

Nikolaos K. Pavlis National Geospatial-Intelligence Agency (NGA), Springfield, Virginia, USA, Nikolaos.K.Pavlis@nga.mil

Fernando Sansò Politecnico di Milano, DIIAR – Milano, Italy, fernando.sanso@polimi.it

Michael G. Sideris University of Calgary, Geomatics Engineering, Calgary, Canada, sideris@ucalgary.ca

C.C. Tscherning Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen Ø, Denmark, cct@gfy.ku.dk

Ilias N. Tziavos Aristotle University of Thessaloniki, Department of Geodesy and Surveying, Thessaloniki, Greece, tziavos@olimpia.topo.auth.gr

Part I
Theory
Fernando Sansò

Chapter 1
The Forward Modelling of the Gravity Field

1.1 Outline of the Chapter

The chapter has the purpose of presenting all the main *characters* of the book and some tools to handle them, and to understand their mathematical properties. We start with the gravitation law in Sect. 1.2, we clarify what is a gravitation field, in particular for an extended body, and we prove that this is a conservative or potential field (Sect. 1.3), i.e., the vector field of gravitational accelerations can be expressed as the gradient of a potential. Switching from an inertial system to one attached to the body of the earth, a proof mass rigidly attached to it will experience the centrifugal acceleration which is also a field that can be expressed as the gradient of a potential. By adding gravitational and centrifugal acceleration vectors, or their potentials (Sect. 1.3), we define the gravity field, which is the object of our study.

In order to understand the mathematical properties of the gravitational part of the gravity potential, we need theorems of vector calculus which are standard in mathematical physics. These are the Gauss theorem, the Dirichlet and the Green identities (Sects. 1.5 and 1.3).

They are used to build the Poisson equation, that relates in differential terms the potential to the mass density, and we find for the first time that the gravitational potential is harmonic outside the masses and regular (i.e., tending to 0) at infinity.

In Sect. 1.7 we introduce the concepts of plumblines and equipotential (or horizontal) surfaces and we study in a quite elementary way their relation to the vertical variation of the gravity vector.

Strictly speaking, this last item, which has been a long lasting object of researches in geodesy, might not be necessary in view of Molodensky's principle that the knowledge of the exterior gravity field can be fully achieved by observations taken exclusively outside the masses. Yet there are points in our theory, where certain approximation procedures can be facilitated by the knowledge of the equations contained in this section.

We also meet in it, for the first time, the definition of geoid and orthometric height, namely the height of any point on the geoid, computed along the vertical.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_1,

Section 1.8 has the aim of learning how to express the Laplace operator in orthogonal coordinates, in particular in spherical and ellipsoidal coordinates, which are so relevant to geodesy. Any reader acquainted with differential and tensor calculus can simply skip it.

Section 1.9 is devoted to the introduction of the so-called *normal field*. This is a model of the gravity field $\boldsymbol{\gamma}$ and of its potential U, which by means of the choice of four constants (the equatorial semi-axis a, the eccentricity e, the angular velocity ω and the value U_0 of U on the reference ellipsoid) and by fixing five geometrical parameters (the position of the center of the ellipsoid and the direction of its polar axis) approximates at once the true gravity field with a relative accuracy somewhere between 10^{-4} and 10^{-5}.

The explicit form of the normal potential is derived in the book by exploiting standard methods of differential equations without any recourse to the theory of analytical functions, as it is usually done in textbooks of theoretical geodesy. This item, which is typically not well-known by students in geodesy, is only shortly touched in Sect. 3.6, where we study global models.

Once the normal potential U is available, it is obvious to define the anomalous potential T as the difference between the actual gravity potential W and U. This is done in Sect. 1.10, where several anomalous quantities are introduced too, such as the height anomaly and the geoid undulation, the gravity disturbance, the free air gravity anomaly and the deflection of the vertical.

Finally, in Sect. 1.11 all the main types of height systems in use in geodesy are recalled. Among them, dynamic and orthometric heights are by definition intrinsic, in the sense that they are defined only on the basis of the physical position of the point P with respect to the earth body and to its true gravity field. On the contrary, ellipsoidal heights and normal heights require the definition of a reference ellipsoid, of its position in space as well as of the normal potential U attached to it.

Since the position of the ellipsoid $\mathcal{E}$ in space is only implicitly defined through conventions and observations, it is not perfectly fixed with respect to the earth body and even more it undergoes variations in time. Therefore it is only natural to study how the various height coordinates that depend on $\mathcal{E}$ change as a consequence of small rototranslations of $\mathcal{E}$ and of its normal potential.

The problem is solved for ellipsoidal heights in a linearized form, while normal heights are basically proved to be invariant at least to the first order.

1.2 Newton's Gravitation Law

In the year 1686 I. Newton, in his "Philosophiae naturalis principia mathematica", formulated one of the basic laws of physics, astronomy and geodesy, namely his celebrated law of gravitational attraction:

- Any two point masses, M_P, M_Q, in an inertial system, attract each other with a force proportional to the values of the masses, and inversely proportional to the square of the distance

$$F = G\frac{M_P M_Q}{\ell_{PQ}^2}\ ; \tag{1.1}$$

the proportionality constant G is known as the *universal gravitational constant* and it has a value which is approximately

$$G = 6{,}672.59 \cdot 10^{-14}\,\mathrm{m^3\,s^{-2}\,kg^{-1}}\ ; \tag{1.2}$$

such a value is known in these years with an accuracy of $\pm 0.30 \cdot 10^{-14}$ $\mathrm{m^3\,s^{-2}\,kg^{-1}}$,

- The direction of the gravitational force exerted by M_Q on M_P is along the line joining M_P and M_Q, it is directed from M_P towards M_Q, so that, in vector form, (1.1) reads

$$\mathbf{F}_{QP} = -GM_P M_Q \frac{\mathbf{r}_{QP}}{\ell_{QP}^3}. \tag{1.3}$$

This law, together with the second law of dynamics, is the basis of Celestial Mechanics, which has obtained so many experimental confirmations in the centuries, that it is considered as an untouchable foundation of physics (Todhunter, 1873). Even the general relativity theory has provided a generalization of it, rather than a disproof (Fischbach et al., 1999).

As a matter of fact, the law has been re-discussed in the history of science. Particularly in recent years the hypothesis has been put forward that the gravitational force could include a term depending as a negative exponential on the distance; therefore this term would not affect the dynamics of bodies very distant from one another, like planets and stars, although, it was guessed, it could be seen in the gravitational interaction between earth and artificial satellites. The hypothesis has not been confirmed by experiments and Newton's law still has to be considered true as it is, at least as far as not too massive bodies are considered (e.g., giant stars or black holes) nor objects moving with a speed comparable to the velocity of light, because in these cases relativistic effects become important.

1.3 The Newtonian Gravitational Attraction of Bodies

As we said, the first formulation of Newton's law refers to point masses. But how can we use it to compute the gravitational attraction of extended bodies, that are part of our common experience? First we note that two masses M_{Q_1}, M_{Q_2} would act on a proof mass m at a point P with a force given by the vector sum of $\mathbf{F}_{Q_1 P}$ and $\mathbf{F}_{Q_2 P}$, namely

Table 1.1 Measurement units: factor is the ratio between MKS and CGS units

Quantity	Symbol	MKS (name)	CGS (name)	Factor
Mass	M	kg (kilogram)	gr (gram)	10^3
Length	L	m (meter)	cm (centimeter)	10^2
Time	T	s (second)	s (second)	10^0
Velocity	$V = LT^{-1}$	ms^{-1}	$cm\,s^{-1}$	10^2
Acceleration	$A = LT^{-2}$	ms^{-2}	$cm\,s^{-2}$ (Gal)	10^2
Force	$F = MA$	$kg\,ms^{-2}$ (Newton)	$gr\,cm\,s^{-2}$ (dyne)	10^5
Energy/work	$E = FL$	N· m	dyne · cm	10^7

$$\mathbf{F}_{Q_1Q_2}(P) = Gm\left(-M_{Q_1}\frac{\mathbf{r}_{Q_1P}}{\ell^3_{Q_1P}} - M_{Q_2}\frac{\mathbf{r}_{Q_2P}}{\ell^3_{Q_2P}}\right). \tag{1.4}$$

From (1.4) we learn two things: first that gravitational forces add like vectors, according to Leonardo da Vinci's parallelogram rule; second that, since the force $\mathbf{F}_{Q_1Q_2}(P)$ is proportional to the proof mass, one can divide both members of (1.4) by m and obtain a "field" $\mathbf{g}_{Q_1Q_2}(P)$ of forces, per unit of proof mass, generated by M_{Q_1} and M_{Q_2}. Such a field has the dimension of an acceleration and thus it is expressed in Newton per kilogram or Gal units

$$\begin{cases} 1\,\text{Gal} = 1\,\text{cm}\,\text{s}^{-2} = 10^{-2}\,\text{N}\,\text{kg}^{-1} \\ 1\,\text{N} = 1\text{Newton} = 1\,\text{kg}\,\text{ms}^{-2}\,; \end{cases} \tag{1.5}$$

in this respect see the Table 1.1 above.

For instance, the order of magnitude of the actual earth gravitational acceleration, on its surface, is about

$$g_{\text{earth}} \sim 10^3\,\text{Gal} = 10^6\,\text{mGal}. \tag{1.6}$$

Generalizing, we arrive at expressing the gravitational field of N point masses $(M_{Q_1}, M_{Q_2} \dots M_{Q_N})$, placed at points Q_i, $i = 1, 2 \dots N$, by the formula

$$\mathbf{g}(P) = -G\sum_{i=1}^{N} M_{Q_i}\frac{\mathbf{r}_{Q_iP}}{\ell^3_{Q_iP}}\,; \tag{1.7}$$

this represents the force exerted by $(M_{Q_1} \dots M_{Q_N})$ on a unit proof mass, placed at P.

Now, by taking masses continuously distributed along a line L, on a surface S or on a body B (cf. Fig. 1.1) one gets the integral formulation of (1.7), namely

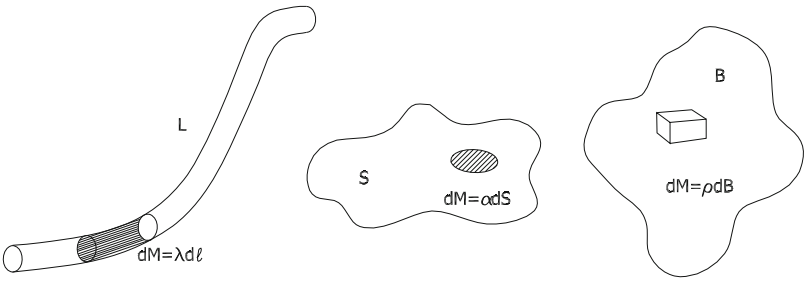


Fig. 1.1 Continuous mass distributions for a line L, a surface S, a body B, and the respective line, surface and body densities λ, α, ρ

$$\mathbf{g}(P) = -G \int_L \frac{\mathbf{r}_{QP}}{\ell_{QP}^3} \lambda(Q) dL_Q \tag{1.8}$$

$$\mathbf{g}(P) = -G \int_S \frac{\mathbf{r}_{QP}}{\ell_{QP}^3} \alpha(Q) dS_Q \tag{1.9}$$

$$\mathbf{g}(P) = -G \int_B \frac{\mathbf{r}_{QP}}{\ell_{QP}^3} \rho(Q) dB_Q \tag{1.10}$$

The functions $\lambda(Q), \alpha(Q), \rho(Q)$ are respectively the line, surface and volume densities of the mass distribution (Farr et al., 2007). More complicated distributions, like double layers, are also used in potential theory.

By their very nature the density functions λ, α, ρ can only be positive as they come from ratios of positive masses to positive line, surface or volume elements. However, it has to be noted that $\mathbf{g}$ depends linearly on such densities, e.g., on ρ. Many times it is, then, convenient to use some average value $\overline{\rho}$ to compute a first approximate value for $\mathbf{g}$ and then compute, as a perturbation, the small contribution to $\mathbf{g}$ due to the variations of the density $\delta\rho = \rho - \overline{\rho}$; indeed in this case $\delta\rho$ can be either positive or negative and still the Newtonian integral (1.10) retains its meaning.

One fundamental concept, in the theory of gravitation, is the gravitational potential $V(P)$ (see Todhunter 1873; Heiskanen and Moritz 1967, Chap. 1); this is by definition a scalar function such that

$$\mathbf{g}(P) = \nabla V(P) \tag{1.11}$$

$$(\nabla = \mathbf{e}_x \frac{\partial}{\partial x} + \mathbf{e}_y \frac{\partial}{\partial y} + \mathbf{e}_z \frac{\partial}{\partial z} = \text{gradient operator}$$

represented in Cartesian coordinates; $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$

are unit vectors parallel to the axes).

That such a function always exists for a given gravitational field, comes from the remark that already for a point mass M one has[1]

$$\mathbf{g}(P) = -GM\frac{\mathbf{r}_{QP}}{\ell_{QP}^3} = \nabla_P\left(\frac{GM}{r_{QP}}\right), \tag{1.13}$$

as it can be directly verified.

By using (1.13), for instance in (1.10) we find that, given suitable regularity conditions on ρ (e.g., it has to be measurable and bounded) and on the domain B (e.g., the volume of B has to be finite), we can write

$$\mathbf{g}(P) = G\int_B\left(\nabla_P\frac{1}{r_{PQ}}\right)\rho(Q)dB_Q = \nabla_P\left(G\int_B\frac{\rho(Q)}{r_{PQ}}dB_Q\right)$$

which proves that

$$V(P) = G\int_B\frac{\rho(Q)}{\ell_{PQ}}dB_Q. \tag{1.14}$$

Since there are functions $V(P)$ such that

$$\nabla_P V(P) \equiv 0$$

over all the space, namely the constant functions, it is clear that $V(P)$ is not uniquely defined by (1.11); nevertheless if we add the condition that

$$V(P) \to 0,\ r_P \to \infty$$

at infinity, we definitely get only one $V(P)$ satisfying (1.11), and this has to be of the form (1.14).

That the function (1.14) goes to zero at infinity is easy to see if we assume that $\rho(Q)$ is a bounded function and B is a bounded set, as we can safely claim to be true for the earth. In fact it is clear that if B is contained in a ball B_0 of radius R_0, then, when $r >> R_0$

$$\frac{1}{\ell_{PQ}} = \frac{1}{r_P} + O\left(\frac{1}{r_P^2}\right),$$

so that from (1.14) we see that

[1]Note: we shall use in an equivalent way the two notations

$$\ell_{QP} = r_{QP} = |\mathbf{r}_P - \mathbf{r}_Q|\,; \tag{1.12}$$

in general we shall prefer r_{QP} when some differential operator has to be applied to this function.

$$V(P) = \frac{G}{r_P}\int_B \rho(Q)dB + O\left(\frac{1}{r_P^2}\right) \tag{1.15}$$
$$= \frac{GM}{r_P} + O\left(\frac{1}{r_P^2}\right).$$

where M is the total mass generating V.

Since it will be useful in this chapter, we refine here the relation (1.15), although the argument will be taken up again in Chap. 2. In fact, note that if (r_P, r_Q) are the radial distances of (P, Q) from the origin and if ψ_{PQ} is the angle between them, i.e.,

$$r_P = |\mathbf{r}_P|, r_Q = |\mathbf{r}_Q|, \cos\psi_{PQ} = \frac{\mathbf{r}_P \cdot \mathbf{r}_Q}{r_P r_Q},$$

then

$$\ell_{PQ} = \sqrt{r_P^2 + r_Q^2 - 2r_P r_Q \cos\psi_{PQ}}$$

and when $r_P >> R_0 > r_Q$ we have

$$\frac{1}{\ell_{PQ}} = \frac{1}{r_P}\left\{1 + \frac{r_Q}{r_P}\cos\psi_{PQ}\right\} + O\left(\frac{1}{r_P^3}\right)$$
$$= \frac{1}{r_P} + \frac{\mathbf{r}_P \cdot \mathbf{r}_Q}{r_P^3} + O\left(\frac{1}{r_P^3}\right).$$

By using this relation in (1.14) and recalling that by definition the barycenter of the mass distribution is

$$\mathbf{b} = \frac{1}{M}\int_B \rho(Q)\mathbf{r}_Q dB_Q,$$

we finally find the sought asymptotic relation

$$V(P) = \frac{GM}{r_P} + \frac{GM\mathbf{r}_P}{r_P^3}\cdot\mathbf{b} + O\left(\frac{1}{r_P^3}\right) \tag{1.16}$$

valid for all Newtonian potentials.

Similar reasonings hold for (1.8) and (1.9), namely we can define a line and a surface potential, with analogous properties,

$$V(P) = G\int_L \frac{\lambda(Q)}{r_{PQ}}d\ell_Q,\ V(P) = G\int_S \frac{\alpha(Q)}{r_{PQ}}dS_Q$$

if (L, λ) and (S, α) are bounded.

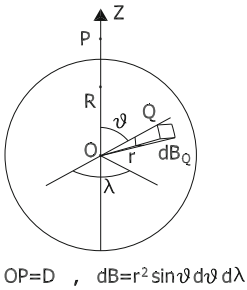


Fig. 1.2 The spherical coordinates for the computation of the potential $V(P)$

Let us try here to see how we could compute the gravitational potential of spherical bodies.

Example 1. Given a ball of radius R and constant density ρ, we want to compute the corresponding Newtonian potential $V(P)$.

First we take the axis Z to go from the center of the ball O, towards the computation point P (see Fig. 1.2), assumed to be at distance D from 0.

By using a spherical coordinate system (see Fig. 1.2) and considering that

$$|\mathbf{r}_P - \mathbf{r}_Q| = \sqrt{r_P^2 + r_Q^2 - 2\mathbf{r}_P \cdot \mathbf{r}_Q},\ \ell_{QP} = \sqrt{r^2 + D^2 - 2rD\cos\vartheta}$$

we have from (1.10)

$$\begin{aligned} V(P) &= G\rho \int_0^R dr r^2 \int_0^{2\pi} d\lambda \int_0^{\pi} \frac{\sin\vartheta}{\sqrt{r^2 + D^2 - 2rD\cos\vartheta}} d\vartheta \\ &= 2\pi G\rho \int_0^R dr r^2 \int_{-1}^{1} \frac{dt}{\sqrt{r^2 + D^2 - 2rDt}} \\ &= 2\pi G\rho \int_0^R dr r^2 \left[\frac{|D + r| - |D - r|}{rD}\right], \end{aligned} \tag{1.17}$$

where we have used $t = \cos\vartheta$.

Now if in (1.15) the point P lies outside the sphere ($r_P = D > R$), we have

$$V_{\text{ext}}(P) = \pi G\rho \frac{4}{3}\frac{R^3}{D} = \left(\frac{4}{3}\pi R^3\rho\right)\frac{G}{D} = \frac{GM}{D} = \frac{GM}{r_P}. \tag{1.18}$$

This is nothing but the statement, well-known since the times of Newton, that the exterior potential of a homogeneous sphere is equal to that of a point with the same mass placed at its center.

On the contrary in the interior, $D < R$, we get

$$V_{\text{int}}(P) = G\rho 2\pi \left(R^2 - \frac{1}{3}D^2\right) = G\rho 2\pi \left(R^2 - \frac{1}{3}r_P^2\right). \tag{1.19}$$

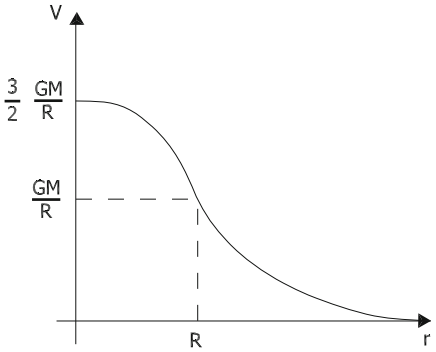


Fig. 1.3 The potential of a homogeneous sphere

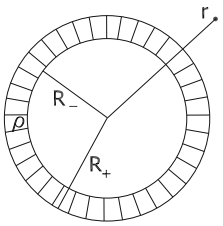


Fig. 1.4 A layer of density ρ as difference of two concentric spheres

A radial section of the potential is shown in Fig. 1.3 where we can read that the gravity modulus $\left(\dfrac{\partial V}{\partial D}\right)$, is zero at the center of the sphere, as expected for symmetry reasons.

Example 2. By subtracting the potential of two concentric balls (see Fig. 1.4) we can get the potential of the layer with inner radius R_-, outer radius R_+ and constant density ρ.

As we can see from (1.18) we get

$$r > R_+, \qquad V_{\text{ext}} = \frac{GM_+}{r} - \frac{GM_-}{r} = \frac{GM}{r} \tag{1.20}$$

where

$$M = M_+ - M_- = \rho \frac{4}{3}\pi (R_+^3 - R_-^3)$$

is the total mass of the layer.

When we penetrate into the layer, on the contrary, we get from (1.19)

$$R_- < r < R_+, \qquad V_{\text{int}} = 2\pi G\rho \left(R_+^2 - \frac{1}{3}r^2 - \frac{2}{3}\frac{R_-^3}{r} \right), \tag{1.21}$$

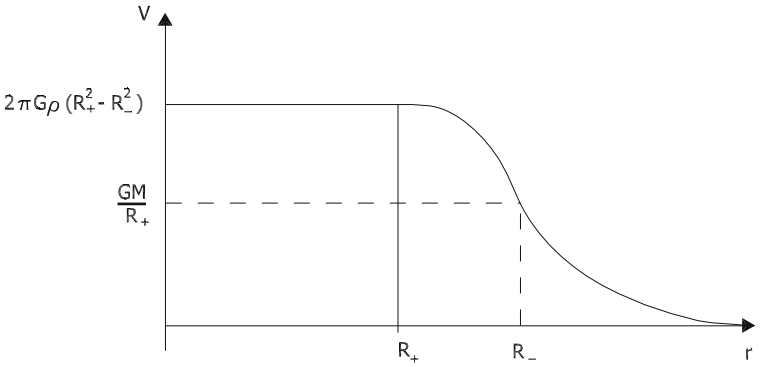


Fig. 1.5 An outlook of the potential of a layer

and finally, when we are inside the hollow,

$$r < R_-, \qquad V_{\text{hollow}} = 2\pi G\rho(R_+^2 - R_-^2). \tag{1.22}$$

It is easy to see that such a potential is continuous and, being constant into the hollow, it generates no attraction there (Fig. 1.5).

Finally if we take $R_+ = \overline{r} + d\overline{r}$, $R_- = \overline{r}$ in (1.20) we see that an infinitesimal layer with density $\rho(\overline{r})$, possibly varying with $\overline{r}$, will generate outside ($r > \overline{r}$) a potential

$$r > \overline{r}, \qquad dV_{\text{out}} = 4\pi G\rho(\overline{r})\overline{r}^2\frac{d\overline{r}}{r} \tag{1.23}$$

and inside a potential

$$r < \overline{r}, \qquad dV_{\text{int}} = 4\pi G\rho(\overline{r})\overline{r}d\overline{r}. \tag{1.24}$$

With the help of (1.23) and (1.24) we find the general expression of a sphere with layered density, i.e., $\rho = \rho(\overline{r})$, as

$$r < R, \qquad V(r) = 4\pi G\left(\frac{\int_0^r \rho(\overline{r})\overline{r}^2 d\overline{r}}{r} + \int_r^R \rho(\overline{r})\overline{r}d\overline{r}\right) \tag{1.25}$$

and again

$$r > R, \qquad V(r) = \frac{4\pi G\int_0^R \rho(\overline{r})\overline{r}^2 d\overline{r}}{r} = \frac{GM}{r} \tag{1.26}$$

outside the masses.

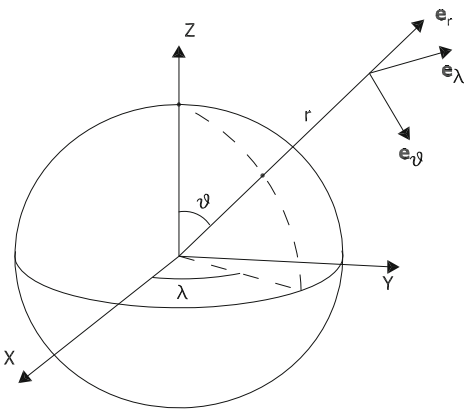


Fig. 1.6 Spherical coordinates and spherical triad

Remark 1. If one wants to derive the gravitational acceleration for the bodies described in the examples above, one has to apply the gradient operator to V. Given the spherical symmetry of those examples, it is convenient to express the gradient operator in spherical coordinates, i.e., with coordinates (ϑ, λ, r) (see Fig. 1.6) (cf. Freeden and Schreiner 2009)

$$\nabla = \frac{1}{r}\mathbf{e}_\vartheta \frac{\partial}{\partial \vartheta} + \frac{1}{r \sin \vartheta}\mathbf{e}_\lambda \frac{\partial}{\partial \lambda} + \mathbf{e}_r \frac{\partial}{\partial r}. \tag{1.27}$$

If we apply such an operator to (1.25) and (1.26) we get

$$\mathbf{g}(r) = g_r \mathbf{e}_r = \begin{cases} -4\pi G \frac{\int_0^r \rho(r) r^2 dr}{r^2}\mathbf{e}_r & (r < R) \\ -\frac{GM}{r^2}\mathbf{e}_r & (r > R). \end{cases} \tag{1.28}$$

As we see, in this spherically layered setting the gravitational vector always points to the origin.

At this point we send the reader to the exercises at the end of the chapter, where potential and attraction for a number of bodies with constant density are presented.

Such expressions, particularly that of prisms, can be used to build models of gravitational attraction for bodies that can be approximated by a combination of such elementary forms. The reader is invited to try to prove the validity of these formulas, with the help of integration tables.

We close the section by recalling the most common measure units, already reported in Table 1.1, related to gravitation and motion, i.e., to mechanics. Let us remember that all units are expressed in terms of the primitive quantities mass, length and time.

Basically there are two systems in use: one is the so-called international system (IS) also called (MKS) from meter, kilogram and second; the other one is the

so-called (CGS) system for centimeter, gram and second. Throughout the book we will use the second, for reasons of geodetic tradition.

In addition, in geodesy we use a special unit for the Newtonian potential, namely the geo-potential-unit

$$1g.p.u = EM^{-1} = AL = 10^5\,\mathrm{Gal\,cm} = 1\,\mathrm{k\,Gal\,m}$$

We also remind that in general when subunits are needed, we use prefixes like deci-(d $= 10^{-1}$), centi-(c $= 10^{-2}$), milli-(m $= 10^{-3}$), micro-($\mu = 10^{-6}$), nano-(n $= 10^{-9}$), pico-(p $= 10^{-12}$). Beyond the familiar examples with units of lenght and time, we quote here mGal (milliGal), μGal (microGal), nGal (nanoGal). Multiples of units are as usual denoted as kilo-(k $= 10^3$), mega-(M $= 10^6$), and so forth.

1.4 The Gravity Field

The theory of gravitation presented in the previous sections is valid in an inertial reference system.

However, when we want to study the forces, acting on material bodies, based on a platform like the earth, one has immediately to realize that an earth fixed reference system cannot be considered as inertial. In fact, the earth is moving, with respect to an inertial system at least with two important non-linear motions: one is the revolution of the earth around the sun, the other is the revolution of the earth around its own rotation axis.

To simplify matters, and with an approximation level more than sufficient for the purpose of this book, we shall consider these two rotations as uniform, namely as having a constant angular velocity, in modulus as well as for the direction, with respect to both an inertial system and the earth body itself. The rotation around the sun can be neglected, in this context, because, although its value is quite large (of the order of $0.6\,\mathrm{Gal} = 6 \cdot 10^{-3}\,\mathrm{m\,s^{-2}}$) the acceleration of a point on the earth surface is about the same, with a maximum variation of the order of 0.025 mGal; this is an expression of the fact that any (small) body is attracted by the sun with the same acceleration as the whole earth. So, if we use as reference system a Cartesian triad centered somehow to the earth and with the Z_I axis along its rotation axis, while X_I and Y_I are always pointing in the same direction with respect to fixed stars, we realize a system which is quasi-inertial, i.e., Newton's law holds in it with quite a good approximation.

However, if we now switch from this system to another one earth-fixed, we can take the same origin and the same Z axis, $Z = Z_I$, because this is the rotation axis, and we shall see (X, Y) uniformly rotating with respect to (X_I, Y_I), with an angular velocity roughly equal to

$$\omega = 0.729 \cdot 10^{-4}\,\mathrm{s^{-1}}. \tag{1.29}$$

This modifies the fundamental law of dynamics of a point, of mass m and coordinate vector $\mathbf{x}_I$, from

$$\begin{aligned} m\ddot{\mathbf{x}}_I &= \mathbf{F} + m\mathbf{g}_N \\ \mathbf{g}_N &= \text{Newtonian gravitational force acting on } m, \\ \mathbf{F} &= \text{other forces acting on } m \end{aligned} \tag{1.30}$$

to the Coriolis law, in terms of the earth-fixed coordinate vector $\mathbf{x}$, (cf. Arnold 1978)

$$\begin{aligned} & m\left[\ddot{\mathbf{x}} + 2\boldsymbol{\omega} \wedge \dot{\mathbf{x}} + \dot{\boldsymbol{\omega}} \wedge \mathbf{x} - \omega^2(I - P_Z)\mathbf{x}\right] \\ &= \mathbf{F} + m\mathbf{g}_N \\ & \boldsymbol{\omega} = \omega\mathbf{e}_Z = \text{angular rotation vector} \\ & P_Z = \text{orthogonal projection on the } Z \text{ axis}, \end{aligned} \tag{1.31}$$

where $\mathbf{a} \wedge \mathbf{b}$ denotes the vector product of $\mathbf{a}$ and $\mathbf{b}$.

If we consider that, due to our hypothesis of uniform rotation, $\dot{\boldsymbol{\omega}} = 0$, we can write (1.31) in the form

$$\mathbf{F} = -m[-\ddot{\mathbf{x}} - 2\boldsymbol{\omega} \wedge \dot{\mathbf{x}} + \omega^2(I - P_Z)\mathbf{x} + \mathbf{g}_N]\,. \tag{1.32}$$

So if we have to apply a force $\mathbf{F}$ to the point mass m, to keep it clamped to the earth (i.e., such that $\dot{\mathbf{x}} = 0, \ddot{\mathbf{x}} = 0$) we see that

$$\begin{aligned} \mathbf{F} &= -m[\mathbf{g}_N + \omega^2(I - P_Z)\mathbf{x}] \\ &= -m[\mathbf{g}_N + \omega^2(x\mathbf{e}_x + y\mathbf{e}_y)]. \end{aligned} \tag{1.33}$$

In other words with point masses fixed to the earth, we feel an acceleration field $\left(-\frac{1}{m}\mathbf{F}\right)$ which is given by

$$\mathbf{g} = \mathbf{g}_N + \mathbf{g}_c = \mathbf{g}_N + \omega^2(x\mathbf{e}_x + y\mathbf{e}_y)\,; \tag{1.34}$$

This is by definition the field of the gravity vector, which is composed by the Newtonian gravitation $\mathbf{g}_N$ and the centrifugal acceleration $\mathbf{g}_c$. Note that in definition (1.34) it is essential that the Z axis be parallel to the rotation axis and that this one is considered to be fixed in the earth body.

Remark 2. We know that in reality the instantaneous north pole (i.e., the intersection of the rotation axis with the earth surface) can move by several meters along the surface in one year. However, with the value of ω of (1.29) we see that the maximum variation of the centrifugal acceleration when a point is diplaced a distance d from the rotation axis, e.g., $d = 10\,\text{m}$, is approximately

$$\delta g_c = \omega^2 d \cong 0.5 \cdot 10^{-8}\,\mathrm{s}^{-2} 10\,\mathrm{m} = 5\,\mu\mathrm{Gal},$$

which is certainly a negligible quantity for our purposes.

We note that also the centrifugal acceleration can be expressed as the gradient of a potential

$$\mathbf{g}_c = \omega^2(x\mathbf{e}_x + y\mathbf{e}_y) = \nabla\frac{1}{2}\omega^2(x^2 + y^2) = \nabla V_c. \tag{1.35}$$

V_c is called the *centrifugal potential*.
From (1.34) we see that we can write

$$\mathbf{g} = \mathbf{g}_N + \mathbf{g}_c = \nabla(V + V_c) = \nabla W. \tag{1.36}$$

The potential

$$W = V + V_c \tag{1.37}$$

is called the *gravity potential*.

The modulus of the gravity vector $g = |\mathbf{g}|$, also called gravity, is a quantity that is directly observable, for instance by measuring the acceleration of a free-falling proof mass along a pipe where vacuum has been made. This is, at least in principle, the idea of an absolute measurement of g, which can be done with an accuracy down to the 1 μGal level.

More common is the relative measurement of gravity, i.e., the difference of gravity values between two points, which can also be performed with an accuracy of a few μGals. We shall not dwell on this problem, that can be more thoroughly studied for instance in Torge (2001), but we just underline that gravity at the μGal level is fairly unstable, reflecting phenomena of a nature which is not of interest in this book. So we shall consider gravity signals to become relevant only when they reach some level between 10^{-2} and 10^{-1} mGal; just to fix the ideas let us assume this threshold to be conventionally equal to 0.03 mGal.

1.5 Gauss, Poisson, Laplace

In this section we aim to prove that there is a fundamental differential equation which is satisfied by the Newtonian gravitational potential V, namely the Poisson equation

$$\nabla \cdot \mathbf{g}_N = \nabla \cdot \nabla V = \Delta V = -4\pi G\rho, \tag{1.38}$$

where the Laplace operator $\nabla \cdot \nabla = \Delta$ is represented, in terms of Cartesian coordinates, by

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{1.39}$$

In (1.38) $\rho = \rho(P)$ is the density of masses which, in our case, are confined by the topographic surface of the earth. In fact, although the atmosphere as a whole is not a light body, yet its density is much lower than that of the solid or liquid earth ($\rho \cong 10^{-3}\,\mathrm{g\,cm^{-3}}$ for the atmosphere as opposed to $\rho \cong 1\,\mathrm{g\,cm^{-3}}$ of water and $\rho \cong 2.67\,\mathrm{g\,cm^{-3}}$ of the earth upper layer) and, even more important, such density is basically spherically layered so that its effect on the earth surface is practically none, according to our Example 2.

Indeed the difference in the gravitational potential due to the presence of the atmosphere between the earth surface and the level of satellites is well-visible, though it can be accounted for by simple corrective terms. For a more detailed report on the subject see Sjöberg (2000). So we can ideally think that ρ is zero outside S. Accordingly, if we call B the volume occupied by the masses, Ω the space exterior to S, so that S is the frontier of both B and Ω, we can split (1.38) as a matter of fact into two equations,

$$\Delta V = 0, \qquad \text{in}\,\Omega \tag{1.40}$$

$$\Delta V = -4\pi G\rho, \quad \text{in}\,B, \tag{1.41}$$

the first one being usually named the Laplace equation, while to the second is more properly reserved the name of Poisson's equation.

At first sight it might seem futile to study the differential equations that V has to satisfy since we have a definite analytical expression for it, as it is the Newton integral (1.10). However, it is precisely the contribution of Physical Geodesy to Geophysics in general, to show how one can determine V from the Laplace equation (1.40), from observations performed on S or in Ω (e.g., satellite observations) and maybe from some knowledge of ρ in the uppermost layer of the earth, namely the crust. In other words we aim at determining V without a detailed knowledge of ρ, so that the outcome of physical geodesy puts constraints on the theory of the earth constitution and its dynamics, rather than viceversa.

A direct computation of the Laplacian of the gravity potential W, shows that (1.38) has to be changed into

$$\Delta W = \nabla \cdot \mathbf{g} = -4\pi G\rho + 2\omega^2. \tag{1.42}$$

Such an equation however is not enough to identify W. One has always to add the definition (1.37), because the same term $2\omega^2$ could be generated by other functions different from $V_c = \frac{1}{2}\omega^2(x^2 + y^2)$, for instance from the function $\frac{1}{3}\omega^2 r^2$. So (1.42)

has always to be accompanied by the specification that $W = V + V_c$, with V a regular Newtonian potential.

Before we prove (1.40) and (1.41), we need a well-known theorem of vector analysis, namely Gauss' theorem (Freeden and Schreiner 2009; Hotine 1969).

Gauss' theorem: Let B be a bounded set, with a boundary S satisfying some smoothness condition, for instance that it is possible to define an outer normal $\mathbf{n}$ at every point P of S and that $\mathbf{n}(P)$ is continuous on S. Let $\mathbf{v}$ be a vector field with first derivatives integrable in B; then, by calling $\mathbf{n}$ the outer normal of S, we have

$$\int_B (\nabla \cdot \mathbf{v}) dB = \int_S \mathbf{v} \cdot \mathbf{n} dS. \tag{1.43}$$

We don't prove the theorem here, but we rather observe that (1.43) implies as well the identity

$$\int_B \nabla f dB = \int_S \mathbf{n} f dS. \tag{1.44}$$

In fact, let's take the scalar product of (1.44) with a constant vector $\mathbf{c}$ and note that

$$\mathbf{c} \cdot \int_B \nabla f dB = \int_B \mathbf{c} \cdot \nabla f dB = \int_B \nabla \cdot (\mathbf{c} f) dB =$$
$$= \int_S \mathbf{n} \cdot \mathbf{c} f dS = \mathbf{c} \cdot \int_S \mathbf{n} f dS.$$

Now we can turn to (1.40). Take the simple potential of a point mass

$$V = \frac{1}{r}, \ \mathbf{g} = \nabla V = -\frac{\mathbf{r}}{r^3};$$

a direct computation shows that when $r \neq 0$

$$\nabla \cdot \mathbf{g} = \Delta V = -\frac{\partial}{\partial x}\left(\frac{x}{r^3}\right) - \frac{\partial}{\partial y}\left(\frac{y}{r^3}\right) - \frac{\partial}{\partial z}\left(\frac{z}{r^3}\right) = 0. \tag{1.45}$$

Since

$$\frac{1}{r} = \frac{1}{|\mathbf{r}_P|}, \ \frac{1}{r_{PQ}} = \frac{1}{|\mathbf{r}_P - \mathbf{r}_Q|}, \tag{1.46}$$

we immediately see that

$$\Delta_P \left(\frac{1}{r_{PQ}}\right) = 0 \quad P \neq Q. \tag{1.47}$$

Already this shows that, when $P \in \Omega$

$$P \in \Omega, \quad \Delta_P V(P) = \Delta_P \int_B G \frac{\rho(Q)}{r_{PQ}} dB_Q = 0. \tag{1.48}$$

A function which satisfies the Laplace equation in some open set is called harmonic in this set. Any Newtonian potential, generated by masses contained in B, is a harmonic function in Ω.

But how to deal with $\Delta_P V(P)$ when P is placed inside B, where the condition $P \neq Q$ is not satisfied? To answer we first compute the

$$\Delta \frac{1}{r_{PQ}} = \nabla \cdot \left(-\frac{\mathbf{r}_{PQ}}{r_{PQ}^3} \right) \tag{1.49}$$

without the restriction $r \neq 0$.

As proved in Sect. A.1, it turns out that

$$\nabla \cdot \left(-\frac{\mathbf{r}_{PQ}}{r_{PQ}^3} \right) = -4\pi \delta(P, Q), \tag{1.50}$$

where $\delta(P, Q)$ is the famous Dirac's function with a pole in Q (cf. Taylor 1958; Yosida 1978). This means that for every continuous $f(Q)$ the identity holds

$$\int \delta(P, Q) f(Q) dB_Q \equiv f(P), \tag{1.51}$$

the integral being extended to the whole space, or, equivalently, to any neighborhood of P.

Accordingly

$$\begin{aligned} \Delta V(P) &= G \int_B \Delta_P \left(\frac{1}{r_{PQ}} \right) \rho(\Omega) dB_Q \\ &= -4\pi G \int_B \delta(P, Q) \rho(Q) dB_Q = -4\pi G \rho(P) \end{aligned} \tag{1.52}$$

and (1.41) is proved.

So far we have considered the case of a potential generated by a volume mass distribution; however in the sequel it will be useful to consider single layer potentials like

$$V(P) = G \int_S \frac{\alpha(Q)}{\ell_{PQ}} dS_Q. \tag{1.53}$$

Without going into details we recall that (Mikhlin, 1964, 1957) if $\alpha(Q)$ is a function integrable on S, then $V(P)$ is a function everywhere continuous. Naturally, outside $S, V(Q)$ is very regular, for instance indefinitely differentiable. Yet across S the derivatives of V have quite a peculiar behaviour. In fact, let us call $\left(\frac{\partial V}{\partial n}\right)_+, \left(\frac{\partial V}{\partial n}\right)_-$ the normal derivatives of V, taken respectively on the outer and on the inner face of S; then such derivatives satisfy the jump relation

$$\left(\frac{\partial V}{\partial n}\right)_+ - \left(\frac{\partial V}{\partial n}\right)_- = -4\pi G\alpha, \tag{1.54}$$

as proved in Sect. A.2.

1.6 Dirichlet, Green

We prove in this section a number of integral identities, which derive basically from Gauss' theorem, that will be used in the sequel (Freeden and Schreiner 2009).

We start with a first identity which comes from

$$\nabla \cdot (v\nabla u) = v\Delta u + \nabla v \cdot \nabla u, \tag{1.55}$$

for suitably smooth u and v.

If we integrate this equation over B we get the first Green identity (Miranda 1970; Heiskanen and Moritz 1967; Kellog 1953)

$$\int_S v\frac{\partial u}{\partial n}dS = \int_B (v\Delta u)dB + \int_B \nabla v \cdot \nabla u dB. \tag{1.56}$$

If we interchange u and v in (1.56) and subtract the two relations, we get

$$\int_S \left(u\frac{\partial v}{\partial u} - v\frac{\partial u}{\partial u}\right) dS = \int_B (u\Delta v - v\Delta u)dB, \tag{1.57}$$

which is also known as *second Green's identity*. A particular case of (1.56) is when u is harmonic in B and $v = u$. In this case we obtain

$$\int_B |\nabla u|^2 dB = \int_S u\frac{\partial u}{\partial n}dS \tag{1.58}$$

which is known as the *Dirichlet identity*.

In particular (1.58) shows that if $u = 0$ on S and u is harmonic in B, then $|\nabla u| = 0$ in B, i.e., $u =$ constant, and therefore $u \equiv 0$ in B because u is already zero on S.

If both u and v are harmonic, from (1.57) we find

$$\int_S u\frac{\partial v}{\partial n}dS = \int_S v\frac{\partial u}{\partial n}dS. \tag{1.59}$$

Now let u satisfy the Poisson equation $\Delta u = -4\pi G\rho$, and take $v = \frac{1}{\ell_{PQ}}$, $P \in B$, in (1.57).

Recalling (1.50) we obtain

$$P \in B, \quad \int_S \left\{ u(Q)\frac{\partial}{\partial n}\frac{1}{\ell_{PQ}} - \frac{1}{\ell_{PQ}}\frac{\partial u(Q)}{\partial n} \right\} dS_Q$$
$$= -4\pi u(P) + 4\pi G \int_B \frac{\rho(Q)}{\ell_{PQ}} dB_Q.$$

Re-arranging we find the third Green equation

$$u(P) = G\int_B \frac{\rho(Q)}{\ell_{PQ}}dB_Q + \frac{1}{4\pi}\int_S \left\{ \frac{1}{\ell_{PQ}}\frac{\partial u(Q)}{\partial n} - u(Q)\frac{\partial}{\partial n}\frac{1}{\ell_{PQ}} \right\} dS_Q. \tag{1.60}$$

Similar considerations are valid for the outer domain Ω with the only proviso that now the normal to S pointing out of Ω is $-\mathbf{n}$. In particular, form (1.60), for the gravitational potential which is harmonic in Ω, we get

$$P \in \Omega, \quad u(P) = \frac{1}{4\pi}\int_S \left\{ u(Q)\frac{\partial}{\partial n}\frac{1}{\ell_{PQ}} - \frac{\partial u}{\partial n}\frac{1}{\ell_{PQ}} \right\} dS_Q. \tag{1.61}$$

The identity (1.61) has the merit to show that if we know u and $\dfrac{\partial u}{\partial n}$ on S, then we know the harmonic function $u(P)$ everywhere. That only one of the two functions is needed to determine u is shown in Part III, Proposition 12, where one sees for example that it is possible to find a Green function $G(P, Q)$ such that

$$P \in \Omega, \quad u(P) = -\frac{1}{4\pi}\int_S \frac{\partial}{\partial n_Q}G(P, Q)u(Q)dS_Q. \tag{1.62}$$

That one can get $u(P)$ from $\frac{\partial u}{\partial n}\big|_S$ can be shown by carefully taking the limit when P approaches S in (1.61), thus obtaining an integral equation, whose solution determines u (Mikhlin, 1957). Since this goes beyond the purpose of these notes we don't pursue this reasoning.

1.7 Elements of Geometry of the Gravity Field and Related Definitions

To represent a conservative field (i.e., one that is gradient of some potential) in geometric terms, it is customary to use a family of lines, called *force lines* of the field, and a family of surfaces, called *equipotential surfaces*.

Plumb-lines or lines of the vertical. These are the force lines of the gravity field, $\mathbf{g}(P)$, and by definition they are at every point tangent to the vector $\mathbf{g}$. Usually we define a unit vector $\mathbf{n}(P)$ which is directed upward, namely

$$\mathbf{n}(P) = -\frac{\mathbf{g}(P)}{g(P)}\,; \tag{1.63}$$

this is the vector of the direction of the vertical. Then by definition the equations satisfied by plumb-lines are, in vector form,

$$\begin{aligned} &\frac{d\mathbf{r}}{ds} = \frac{dx}{ds}\mathbf{e}_x + \frac{dy}{ds}\mathbf{e}_y + \frac{dz}{ds}\mathbf{e}_z = \mathbf{n}(P) \\ &ds = \sqrt{dx^2 + dy^2 + dz^2}\,; \end{aligned} \tag{1.64}$$

such a system of differential equations generates a family of lines, one passing through each point P_0 in space where

$$|\mathbf{g}(P_0)| \neq 0, \tag{1.65}$$

because at points where $g(P) = 0, \mathbf{n}(P_0)$ is not defined. Such a condition is certainly always satisfied in the outer space Ω, close to the surface S, and in the first layers of the earth body B. Furthermore, since $\mathbf{g}(P)$ is smooth enough, at least everywhere continuous up to the first derivatives, the plumb-lines are regular lines too.

Equipotential surfaces of the gravity field. These are the surfaces for which

$$W(P) = \overline{W} = \text{constant}. \tag{1.66}$$

Since $\mathbf{n}$ is parallel to $\mathbf{g} = \nabla W$, $\mathbf{n}$ is also orthogonal to the surfaces on which W is constant, therefore plumb-lines, which are tangent to $\mathbf{n}$, always cross orthogonally the equipotential surfaces, i.e., an equipotential surface is always tangent to the horizontal plane at each of its points (Fig. 1.7).

It is interesting in general to note that equipotential surfaces are closed when they lie in the surrounding of the earth surface, but they become unbounded and quite complicated if we move deeper in open space, as it will be illustrated by a simple case in Sect. 1.9.

We mention only at this point that a deep analysis has been done of several problems concerning the geometry of plumb-lines and equipotentials with various tensors related to the gravity field; on this subject we have at least to quote two famous books, namely Mathematical Geodesy by Hotine (1969) and Intrinsic Geodesy, by Marussi (1985). On similar items one can consult (Grafarend 1975, 1986) too.

In this context we just prove a formula connecting the principal curvature of plumb-lines to the horizontal gradient of $g(P) = |\mathbf{g}(P)|$ and another formula

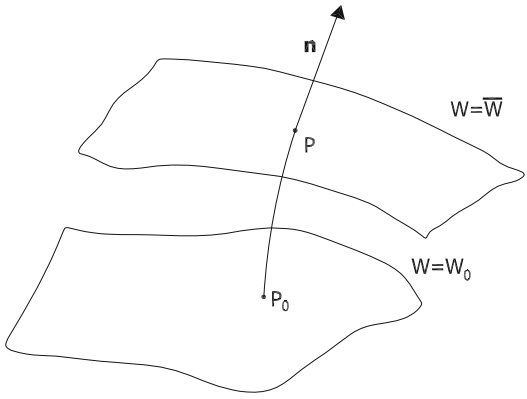


Fig. 1.7 Two equipotential surfaces ($W = \overline{W}$ and $W = W_0$) and a crossing plumb-line; P_0 is the "projection" along the plumb-line of P on $W = W_0$

relating the vertical gradient of $g(P)$ to the mean curvature of equipotential surfaces.

Plumb-line curvature. The horizontal gradient of g, that we denote by ∇_h, is by definition given by

$$\begin{cases} \nabla_h g(P) = \nabla g - \mathbf{n}(\mathbf{n} \cdot \nabla g) = (I - P_n)\nabla g \\ P_n = \text{orthogonal projection on the vertical } \mathbf{n}. \end{cases} \tag{1.67}$$

On the other hand we have (using the shorthand notation $\partial_i = \frac{\partial}{\partial x_i}$)

$$\begin{aligned} \nabla g &= \left\{ \frac{1}{2g} \partial_i g^2 \right\} = \left\{ \frac{1}{2g} \partial_i \Sigma_k (\partial_k W)^2 \right\} \\ &= \left\{ \frac{1}{g} \Sigma_k (\partial_{ki} W) \partial_k W \right\} = \frac{1}{g} \underline{\mathbf{W}}\mathbf{g} = -\underline{\mathbf{W}}\mathbf{n}, \end{aligned} \tag{1.68}$$

where we have introduced the matrix $\underline{\mathbf{W}}$, also called the *Marussi tensor*, of the second derivatives of the potential W (cf. Marussi 1985).

So, by using (1.68) in (1.67) we see that

$$\nabla_h g = -(I - P_n)\underline{\mathbf{W}}\mathbf{n}. \tag{1.69}$$

Now recall that by definition of curvature of any line with tangent vector $\boldsymbol{\tau}$, we have

$$\frac{d\boldsymbol{\tau}}{ds} = \mathbf{c}, \tag{1.70}$$

where $\mathbf{c}$ is a vector orthogonal to $\boldsymbol{\tau}$, $|\mathbf{c}|^{-1} = \mathcal{R}$ the curvature radius of the line and ds is just the line element.

To apply (1.70) to $\mathbf{n}$, for a shift ds along a plumb-line, we first write, recalling also (1.63),

$$d\mathbf{n} = d\left(-\frac{\mathbf{g}}{g}\right) = -\frac{1}{g}d\mathbf{g} + \frac{1}{g^2}dg\mathbf{g} \tag{1.71}$$
$$= -\frac{1}{g}(d\mathbf{g} + dg\mathbf{n}).$$

But, when we move the point P through a distance ds along the vertical $\mathbf{n}$,

$$d\mathbf{g} = \underline{\mathbf{W}}(ds\mathbf{n}) \tag{1.72}$$

and, according to (1.68),

$$dg = ds\mathbf{n} \cdot \nabla g = -ds\mathbf{n} \cdot \underline{\mathbf{W}}\mathbf{n}. \tag{1.73}$$

Summarizing (1.72) and (1.73) in (1.71) we find

$$d\mathbf{n} = -\frac{1}{g}[\underline{\mathbf{W}}\mathbf{n} - \mathbf{n}(\mathbf{n} \cdot \underline{\mathbf{W}}\mathbf{n})]ds \tag{1.74}$$
$$= \frac{1}{g}[-(I - P_n)\underline{\mathbf{W}}\mathbf{n}]ds,$$

i.e., comparing with (1.69)

$$\frac{d\mathbf{n}}{ds} = \mathbf{c} = \frac{1}{g}\nabla_h g. \tag{1.75}$$

Equation (1.75) tells us also that the horizontal gradient of g is just g itself multiplied by the principal curvature vector of the plumb-line.

Vertical gradient of gravity. We want to prove its relation to the mean curvature of the equipotential surface.

Let us first remember that if we take a point P on any smooth surface and we cut the surface with planes containing its normal, $\mathbf{n}(P)$ at P, we get sections (so-called *normal sections*) with varying curvatures.

Among them, two particular normal sections will have the minimum and maximum curvatures, $c_1 = \mathcal{R}_1^{-1}, c_2 = \mathcal{R}_2^{-1}$ (cf. Hotine 1969). These two sections are orthogonal to one another, so that an area element on the surface can be written as (see Fig. 1.8)

$$dS = dL_1 dL_2 = \mathcal{R}_1 d\vartheta_1 \mathcal{R}_2 d\vartheta_2. \tag{1.76}$$

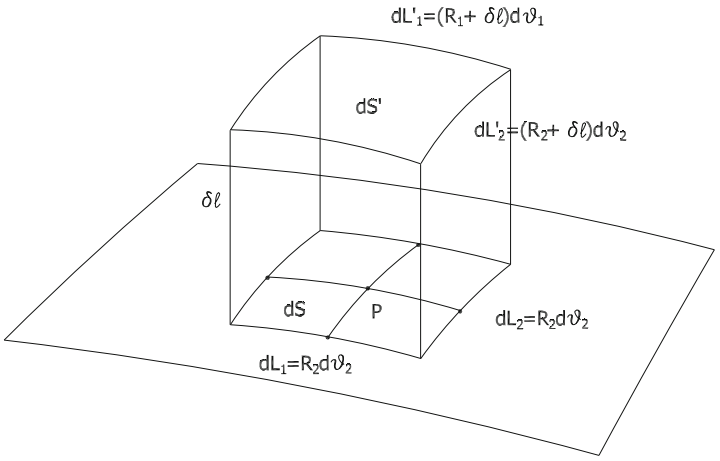


Fig. 1.8 A tube of flux of **g**, with base dS and vertical walls of height $\delta\ell$

If we shift upward dS by a quantity $\delta\ell$ along the plumb-lines we get a volume element with the top area given by

$$\begin{aligned} dS' = dL'_1 dL'_2 &= dL_1 dL_2 + \delta\ell d\vartheta_1 \mathcal{R}_2 d\vartheta_2 + \delta\ell d\vartheta_2 \mathcal{R}_1 d\vartheta_1 \\ &+ O(\delta\ell^2) = dS + \delta\ell \frac{dS}{\mathcal{R}_1} + \delta\ell \frac{dS}{\mathcal{R}_2} \\ &+ O(\delta\ell^2) = dS + \delta\ell dS 2C + O(\delta\ell^2), \end{aligned} \tag{1.77}$$

where we have put

$$\begin{aligned} C &= \frac{1}{2}\left(\frac{1}{\mathcal{R}_1} + \frac{1}{\mathcal{R}_2}\right) = \frac{1}{2}(c_1 + c_2) \\ &= \text{mean curvature of the surface at } P. \end{aligned} \tag{1.78}$$

Now let us write the flux of **g** through this volume element. Considering that the normal to dS' is $\mathbf{n}'$, the normal to dS is $-\mathbf{n}$ and that the lateral walls are parallel to **n**, so that **g** has no flux through them, we can write, by using Gauss' theorem and (1.42),

$$\begin{aligned} &\mathbf{g}' \cdot \mathbf{n}' dS' - \mathbf{g} \cdot \mathbf{n} dS = -g' dS' + g dS \\ &= (-4\pi G\rho + 2\omega^2) dS \delta\ell + O(\delta\ell^2). \end{aligned} \tag{1.79}$$

This can be rearranged as

$$-\frac{g' - g}{\delta\ell}\frac{dS'}{dS} - g\left(\frac{dS' - dS}{\delta\ell dS}\right) = -4\pi G\rho + 2\omega^2 + O(\delta\ell) \tag{1.80}$$

and, with the help of (1.77), we get to the limit for $\delta\ell \to 0$

$$-\frac{\partial g}{\partial \ell} - 2Cg = -4\pi G\rho + 2\omega^2$$

or

$$\frac{\partial g}{\partial \ell} = -2Cg + 4\pi G\rho - 2\omega^2, \tag{1.81}$$

that is the sought relation (Heiskanen and Moritz 1967).

Gravity gradient. We note that by combining (1.81) with (1.75) we get the beautiful equation of the gradient of $g(P)$,

$$\nabla g = -(2Cg - 4\pi G\rho + 2\omega^2)\mathbf{n} + g\mathbf{c} \tag{1.82}$$

relating directly the curvatures of equipotential surface and plumb-line in P with the variation of the modulus of gravity (Heiskanen and Moritz 1967; Hotine 1969; Marussi 1985).

Natural coordinates. Since, as we mentioned, equipotential surfaces in the surrounding of the earth surface are closed, people started to consider the possibility of using $W(P)$ as a natural (i.e., physical) coordinate for the point P.

A lot on this item can be found in geodetic literature, but we send the interested reader to the two classical books Hotine (1969) and Marussi (1985) or the works of Grafarend (1975, 1986).

Since by changing $\overline{W}$, the surface $S_{\overline{W}} = \{P;\ W(P) = \overline{W}\}$ moves up and down, it was only natural to consider $W(P)$ as a kind of "height" coordinate of P. Since any point P in a three-dimensional space needs at least three coordinates to be univocally identified, we have to look for another couple of coordinates that could fix P on the surface $S_{\overline{W}}$. For this purpose it is traditional to use the so-called Gauss mapping, i.e., a pair of angles that do define the direction of the vertical, $\mathbf{n}(P)$, in space. This requires that the correspondence between $\mathbf{n}$ and P (on $S_{\overline{W}}$) be one to one. In practice this is the case if the equipotential surfaces are convex and we shall accept that this hypothesis is verified for the earth without any further discussion. A counterexample could be found in Krarup (2006).

Typically, the angles used to identify $\mathbf{n}$ are the so-called astro-geodetic longitude and latitude defined as follows.

We use an earth-fixed Cartesian triad with the Z axis coinciding with the rotation axis and the origin placed at the barycenter of the mass distribution described by the density $\rho(Q)$. Recalling the definition of barycenter $\mathbf{b}$, we will have in this case

$$\mathbf{b} = \frac{1}{M}\int_B \mathbf{r}_Q\ \rho(Q)dB_Q \equiv 0, \tag{1.83}$$

with M the total mass of the earth.

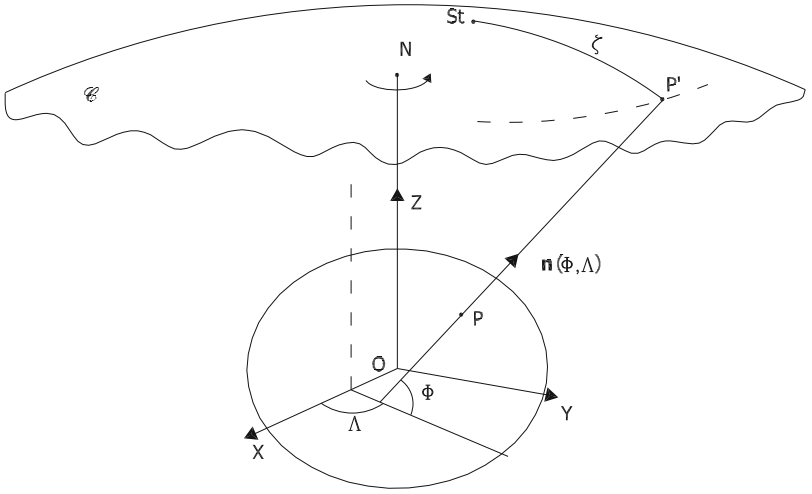


Fig. 1.9 The definition of Φ, Λ; P' projection of P on C along $\mathbf{n}$, $/St$ a star

The (X, Y) plane is called the equatorial plane, and the X axis is chosen on it by some conventional rule, for instance be requiring that the (X, Z) plane passes through some given point or that it is parallel to the vector $\mathbf{n}$ at some given point on the earth surface.

Then by definition the latitude Φ is the inclination of $\mathbf{n}$ with respect to the equatorial plane, i.e (cf. Fig. 1.9).

$$\sin \Phi_P = \mathbf{n}(P) \cdot \mathbf{e}_z \, ; \tag{1.84}$$

The longitude Λ is the dihedral angle between a plane parallel to both $\mathbf{e}_Z$ and $\mathbf{n}$ and the origin plane (X, Z). In practice Λ can be measured as an angle in the equatorial plane (cf. Fig. 1.9).

Note that in general the line through P containing $\mathbf{n}$ needs not to cross the equatorial plane at the origin O or to cross any of the axes, because the irregularities of the gravity field cause $\mathbf{n}$ not to follow any particular symmetry rule.

Note also that, in principle, Φ, Λ can be determined by astronomical observations. In fact the direction $\mathbf{n}$ has a trace on the celestial sphere, C in Fig. 1.9, that is rotating uniformly (in our simplistic model) around the north pole N.

So, by observing the spherical angle between some stars, like St in Fig. 1.9, of known celestial coordinates and knowing the time of the observation (so that we know the angle between the plane (ONX) and the reference meridian on C, fixed with respect to stars) we can infer both Φ and Λ. Whence the name of astro-geodetic coordinates.

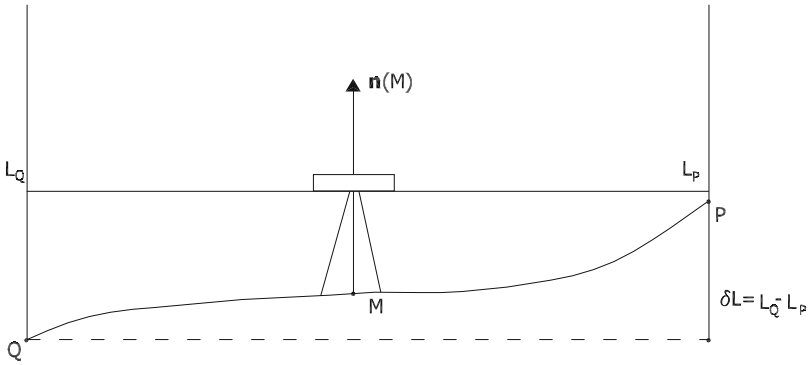


Fig. 1.10 Spirit leveling measurement, $L_Q - L_P = d\mathbf{r}_{QP} \cdot \mathbf{n}(M)$

Let us observe too that, given the definition of Φ, Λ, the vector $\mathbf{n}$ has components in (X, Y, Z)

$$\mathbf{n} = \begin{vmatrix} \cos\Phi\cos\Lambda \\ \cos\Phi\sin\Lambda \\ \sin\Phi \end{vmatrix}. \tag{1.85}$$

For a nicer discussion of terrestrial and celestial reference frames, their reciprocal relation and relevant coordinates, see for instance Vaníček and Krakiwsky (1986).

A first definition of geoid, and orthometric heights. As already claimed $W(P)$ can be used as a height coordinate. Yet $W(P)$ at present cannot be observed directly, though there are hopes that this will become feasible, with proper accuracy, in future, by measuring the frequency of an atomic clock.

Nevertheless the increment of W passing from a point P to a point Q can be easily determined by combining gravity measurements and spirit leveling. In fact assume the two points Q and P to be close enough to one another, say a distance of 100 m apart, so that we can consider the base vector

$$d\mathbf{r}_{QP} = \mathbf{r}_P - \mathbf{r}_Q$$

as infinitesimal, compared to the radius of the Earth. Let M be the midpoint of the segment $\overline{QP}$ and put

$$\delta L_{QP} = L_Q - L_P = d\mathbf{r}_{QP} \cdot \mathbf{n}(M). \tag{1.86}$$

This number is exactly what is observed by a single leveling measurement (cf. Fig. 1.10), that we shall call the *leveling increment*.

Since $\mathbf{n}(M) = -\frac{\mathbf{g}(M)}{g(M)}$, if we know $g(M)$, we can put

$$-g(M)\delta L_{QP} = d\mathbf{r}_{QP} \cdot \mathbf{g}(M) = W(P) - W(Q). \tag{1.87}$$

Adding many small increments of this kind, along a leveling line $\mathcal{L}$, between two points A and B, we get

$$W(B) - W(A) = \int_{\mathcal{L}} -g(Q)dL, \tag{1.88}$$

namely the potential increment between the extremes.

This calls for the use of one particular equipotential surface as reference, and we shall call it the geoid $\mathcal{G}$. Such surface $\mathcal{G}$ can be defined either through a conventional value W_0, and then we are left with the problem of finding some physical point lying on $\mathcal{G}$, or by requiring that $\mathcal{G}$ passes through some physical point and then we have the problem of determining the value of W at that point. Consequently one can determine the potential difference for any other point P by connecting it to some point of $\mathcal{G}$. If we assume that W_0 is the value of the potential on $\mathcal{G}$ then we shall be able to determine

$$C(P) = W_0 - W(P)\,; \tag{1.89}$$

$C(P)$ is called geopotential number of P.

Sometimes, in order to have a height with the more intuitive metric properties of being dimensionally a length, one defines a *dynamic height* of a point P by dividing $C(P)$ by some conventional value of gravity $\overline{g}$

$$H_{\mathrm{dyn}}(P) = \frac{C(P)}{\overline{g}}. \tag{1.90}$$

Completely different in nature is the definition of the so-called *orthometric height*; this is in fact the length of the plumb-line arc between the point P and its projection P_0 on $\mathcal{G}$, counted positively upward (see Fig. 1.11).

As intuitive as it is, yet the orthometric height is a quantity that cannot be easily related analytically to observables, in particular considering that since $\mathcal{G}$ is always chosen so as to be close to the mean surface of the oceans, then it is most of the times buried in the masses, in correspondence to continental areas. As we shall see later on, H can be approximately determined only if we assume to know as well the density of mass above $\mathcal{G}$. We warn the reader however that several nations have switched from using orthometric heights to other height systems that don't require, according to Molodensky's theory, any knowledge of mass density.

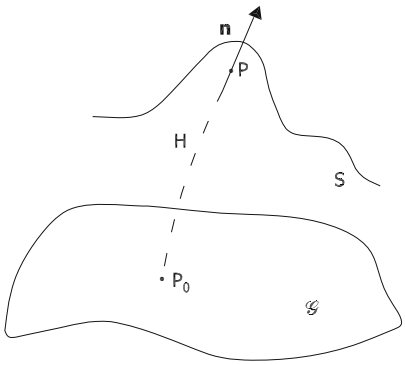


Fig. 1.11 The definition of orthometric height

1.8 The Laplace Operator in Curvilinear Coordinates

We shall soon need the expression of the Laplace operator in spherical and in ellipsoidal coordinates. In order to find them, we tackle first the problem of expressing the operator ∇ in all type of orthogonal coordinates; subsequently we shall compute $\nabla \cdot \nabla = \Delta$.

As proved in Sect. A.3, if one calls $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$ three orthogonal curvilinear coordinates and one puts

$$\mathbf{h}_j = \frac{\partial}{\partial \xi_j} \mathbf{r}(\boldsymbol{\xi}), \; h_j = |\mathbf{h}_j| \tag{1.91}$$

then the following formula holds

$$\boldsymbol{\nabla} = \sum_{j=1}^{3} \frac{\mathbf{h}_j}{h_j^2} \frac{\partial}{\partial \xi_j}. \tag{1.92}$$

Moreover, after introducing the quantity $H = h_1 h_2 h_3$, one finds the basic formula

$$\Delta = \frac{1}{H} \sum_{j=1}^{3} \frac{\partial}{\partial \xi_j} \left[\frac{H}{h_j^2} \frac{\partial}{\partial \xi_j} \right] \tag{1.93}$$

The expression (1.93), the proof of which is given in the Sect. A.3, is particularly manageable to be used in the two examples we have in mind.

Example 3. Take as (ξ_1, ξ_2, ξ_3) the spherical coordinates (r, ϑ, λ), so that

$$\begin{vmatrix} x \\ y \\ z \end{vmatrix} = \begin{vmatrix} r\sin\vartheta\cos\lambda \\ r\sin\vartheta\sin\lambda \\ r\cos\vartheta \end{vmatrix}. \tag{1.94}$$

From

$$\begin{aligned} d\mathbf{r} \equiv \begin{vmatrix} dx \\ dy \\ dz \end{vmatrix} &= \begin{vmatrix} \sin\vartheta\cos\lambda \\ \sin\vartheta\sin\lambda \\ \cos\vartheta \end{vmatrix} dr + \begin{vmatrix} r\cos\vartheta\cos\lambda \\ r\cos\vartheta\sin\lambda \\ -r\sin\vartheta \end{vmatrix} d\vartheta \\ &+ \begin{vmatrix} -r\sin\vartheta\sin\lambda \\ r\sin\vartheta\cos\lambda \\ 0 \end{vmatrix} d\lambda \equiv \mathbf{h}_r dr + \mathbf{h}_\vartheta d\vartheta + \mathbf{h}_\lambda d\lambda \end{aligned} \tag{1.95}$$

we find the three vectors $\mathbf{h}_r, \mathbf{h}_\vartheta, \mathbf{h}_\lambda$ and we can verify directly that they are orthogonal.

Furthermore we get

$$h_r = |\mathbf{h}_r| = 1,\ h_\vartheta = |\mathbf{h}_\vartheta| = r,\ h_\lambda = |\mathbf{h}_\lambda| = r\sin\vartheta \tag{1.96}$$

which implies the well-known metric relation in spherical coordinates

$$|d\mathbf{r}|^2 = dr^2 + r^2 d\vartheta^2 + r^2\sin^2\vartheta d\lambda^2. \tag{1.97}$$

Since then

$$H = r^2\sin\vartheta,$$

we get

$$\begin{aligned} H\Delta &= \partial_r(r^2\sin\vartheta)\partial_r + \partial_\vartheta(\sin\vartheta)\partial_\vartheta + \partial_\lambda\left(\frac{1}{\sin\vartheta}\right)\partial_\lambda \\ &= \sin\vartheta\left[\partial_r(r^2)\partial_r + \operatorname{ctg}\vartheta\,\partial_\vartheta + \partial_\vartheta^2 + \frac{1}{\sin^2\vartheta}\partial_\lambda^2\right] \end{aligned}$$

so that the Laplace equation takes the usual form

$$\Delta u = \frac{1}{r^2}(r^2\partial_r^2 u + 2r\partial_r u + \partial_\vartheta^2 u + \operatorname{ctg}\vartheta\,\partial_\vartheta u + \frac{1}{\sin^2\vartheta}\partial_\lambda^2)u = 0 \tag{1.98}$$

or

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\left(\frac{\partial^2 u}{\partial\vartheta^2} + \operatorname{ctg}\vartheta\frac{\partial u}{\partial\vartheta} + \frac{1}{\sin^2\vartheta}\frac{\partial^2 u}{\partial\lambda^2}\right) = 0. \tag{1.99}$$

Remark 3. Let us observe that in (1.99) one can separate the action of the radial differentiation and that of the angular derivatives. If we put (Heiskanen and Moritz 1967 and Freeden and Schreiner 2009)

$$\begin{aligned}\nabla &= \mathbf{e}_r \frac{\partial}{\partial r} + \frac{1}{r}\mathbf{e}_\vartheta \frac{\partial}{\partial \vartheta} + \frac{1}{r\sin\vartheta}\mathbf{e}_\lambda \frac{\partial}{\partial \lambda} \\ &= \mathbf{e}_r \frac{\partial}{\partial r} + \frac{1}{r}\nabla_\sigma\end{aligned} \tag{1.100}$$

one finds again

$$\begin{aligned}\Delta = \nabla \cdot \nabla &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}(\nabla_\sigma \cdot \nabla_\sigma) \\ &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\Delta_\sigma.\end{aligned} \tag{1.101}$$

In fact one can use the identities

$$\frac{\partial}{\partial r}\mathbf{e}_\vartheta = 0,\ \frac{\partial}{\partial r}\mathbf{e}_\lambda = 0,\ \frac{\partial}{\partial \vartheta}\mathbf{e}_r = \mathbf{e}_\vartheta,\ \frac{1}{\sin\vartheta}\frac{\partial}{\partial \lambda}\mathbf{e}_r = \mathbf{e}_\lambda$$

to prove that

$$\mathbf{e}_r \frac{\partial}{\partial r} \cdot \left(\frac{1}{r}\nabla_\sigma\right) = -\frac{1}{r^2}\mathbf{e}_r \cdot \nabla_\sigma + \frac{1}{r}\mathbf{e}_r \cdot \left(\frac{\partial}{\partial r}\nabla_\sigma\right) = 0$$

and

$$\nabla_\sigma \cdot \left(\mathbf{e}_r \frac{\partial}{\partial r}\right) = (\nabla_\sigma \cdot \mathbf{e}_r)\frac{\partial}{\partial r} + \mathbf{e}_r \cdot \nabla_\sigma \frac{\partial}{\partial r} = 2\frac{\partial}{\partial r}$$

from which (1.101) easily follows.

The operator

$$\Delta_\sigma = \frac{\partial^2}{\partial \vartheta^2} + \operatorname{ctg}\vartheta \frac{\partial}{\partial \vartheta} + \frac{1}{\sin^2\vartheta}\frac{\partial^2}{\partial \lambda^2} \tag{1.102}$$

is called the *Laplace-Beltrami operator*.

Example 4. Since it is known that the geoid is very-well approximated by an ellipsoid of revolution, we are interested in studying the Laplace operator in a form adapted to such an ellipsoid.

We introduce then the reduced ellipsoidal coordinates $(q, \overline{\vartheta}, \lambda)$ or (q, β, λ), where $\overline{\vartheta}$ is called the *reduced ellipsoidal co-latitude* and $\beta = \frac{\pi}{2} - \overline{\vartheta}$ the *reduced ellipsoidal latitude*,

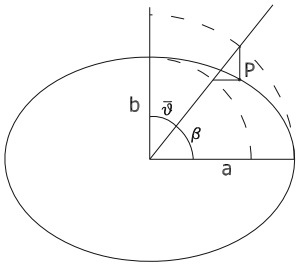


Fig. 1.12 The oblate ellipsoid with semi-axes a, b and the reduced latitude β of the point P

$$\left|\begin{matrix} x \\ y \\ z \end{matrix}\right| = \left|\begin{matrix} \sqrt{q^2+E^2}\sin\overline{\vartheta}\cos\lambda \\ \sqrt{q^2+E^2}\sin\overline{\vartheta}\sin\lambda \\ q\cos\overline{\vartheta} \end{matrix}\right|, \tag{1.103}$$

where

$$E^2 = a^2 - b^2 \tag{1.104}$$

is the squared linear eccentricity, and q ranges from b to $+\infty$.

From (1.103) one immediately realizes that the surfaces q = constant, have equations

$$\frac{x^2+y^2}{q^2+E^2} + \frac{z^2}{q^2} = 1, \tag{1.105}$$

namely they are ellipsoids of revolution. In particular if we take $q = b$ in (1.105) we get

$$\frac{x^2+y^2}{a^2} + \frac{z^2}{b^2} = 1$$

and in this sense we see that our coordinate system is adapted to an oblate ellipsoid of revolution with semi-minor axis b (polar) and semi-major axis a (equatorial), as shown in Fig. 1.12.

In this case we find, with

$$m = \sqrt{q^2+E^2},\ p = \sqrt{q^2+E^2\cos^2\vartheta},$$

$$\mathbf{h}_q = \left|\begin{matrix} \frac{q}{m}\sin\overline{\vartheta}\cos\lambda \\ \frac{q}{m}\sin\overline{\vartheta}\sin\lambda \\ \cos\overline{\vartheta} \end{matrix}\right|$$

$$\mathbf{h}_\vartheta = \begin{vmatrix} m\cos\overline{\vartheta}\cos\lambda \\ m\cos\overline{\vartheta}\sin\lambda \\ -q\sin\overline{\vartheta} \end{vmatrix} \tag{1.106}$$

$$\mathbf{h}_\lambda = \begin{vmatrix} -m\sin\overline{\vartheta}\sin\lambda \\ m\sin\overline{\vartheta}\cos\lambda \\ 0 \end{vmatrix}.$$

It is easy to verify directly that

$$\mathbf{h}_q \cdot \mathbf{h}_\vartheta = 0,\ \mathbf{h}_q \cdot \mathbf{h}_\lambda = 0,\ \mathbf{h}_\vartheta \cdot \mathbf{h}_\lambda = 0,$$

so that (1.241) applies.

In this case we have

$$h_q = \frac{p}{m},\ h_\vartheta = p,\ h_\lambda = m\sin\overline{\vartheta},\ H = p^2\sin\overline{\vartheta}. \tag{1.107}$$

Therefore a direct computation gives

$$\begin{aligned} H\Delta &= \frac{\partial}{\partial q}m^2\sin\overline{\vartheta}\frac{\partial}{\partial q} + \frac{\partial}{\partial\overline{\vartheta}}(\sin\overline{\vartheta})\frac{\partial}{\partial\overline{\vartheta}} + \frac{\partial}{\partial\lambda}\frac{p^2}{m^2\sin\overline{\vartheta}}\frac{\partial}{\partial\lambda} \\ &= \sin\overline{\vartheta}\left\{\frac{\partial}{\partial q}(q^2+E^2)\frac{\partial}{\partial q} + \frac{\partial^2}{\partial\overline{\vartheta}^2} + \mathrm{ctg}\overline{\vartheta}\frac{\partial}{\partial\overline{\vartheta}} + \frac{q^2+E^2\cos^2\overline{\vartheta}}{(q^2+E^2)\sin^2\overline{\vartheta}}\frac{\partial^2}{\partial\lambda^2}\right\} \end{aligned}$$

and finally the Laplace equation writes

$$\begin{aligned} &(q^2+E^2)\frac{\partial^2 u}{\partial q^2} + 2q\frac{\partial u}{\partial q} + \frac{\partial^2 u}{\partial\overline{\vartheta}^2} + \mathrm{ctg}\overline{\vartheta}\frac{\partial u}{\partial\overline{\vartheta}} \\ &+\frac{q^2+E^2\cos^2\overline{\vartheta}}{(q^2+E^2)\sin^2\overline{\vartheta}}\frac{\partial^2 h}{\partial\lambda^2} = 0. \end{aligned} \tag{1.108}$$

It will be useful in future to realize that by exploiting the identity

$$\frac{q^2+E^2\cos^2\overline{\vartheta}}{(q^2+E^2)\sin^2\overline{\vartheta}} = \frac{1}{\sin^2\overline{\vartheta}} - \frac{E^2}{q^2+E^2}, \tag{1.109}$$

(1.108) can be written as

$$(q^2 + E^2)\frac{\partial^2 u}{\partial q^2} + 2q\frac{\partial u}{\partial q} + \overline{\Delta}_\sigma u - \frac{E^2}{q^2 + E^2}\frac{\partial^2 u}{\partial \lambda^2} = 0 \tag{1.110}$$

with $\overline{\Delta}_\sigma$, the Laplace-Beltrami operator in ellipsoidal angular coordinates,

$$\overline{\Delta}_\sigma = \frac{\partial^2}{\partial \overline{\vartheta}^2} + \operatorname{ctg}\overline{\vartheta}\frac{\partial}{\partial \overline{\vartheta}} + \frac{1}{\sin^2 \overline{\vartheta}}\frac{\partial^2}{\partial \lambda^2}. \tag{1.111}$$

1.9 Simple Mathematical Models of the Gravity Field

After the Newton *Principia*, for about 150 years scientists have studied the problem of giving a convenient mathematical model to perform in an easy and direct way computations of quantities related to the gravity field like potential differences, gravity values, vectors of the vertical **n** and so on.

This research was conducted to a fully satisfactory point at the end of the nineteenth century by Pizzetti (cf. Pizzetti 1894) and further systematized by Somigliana (cf. Somigliana 1929) at the beginning of the twentieth century with the definition of the so-called *normal gravity potential* and *normal gravity field*.

At first sight one might think that a reasonable approximation of W can be obtained by taking just the spherical term

$$V_S = \frac{GM}{r}. \tag{1.112}$$

Indeed V_S will be used later on in suitable approximation procedures, called *spherical approximations*, but only carefully controlling the error introduced by taking $W \sim V_S$.

In fact, even if in (1.112) we use a perfect value for the mass of the earth, we see that $W - V_S$ still contains the centrifugal potential, so that this function is not harmonic in Ω and even more the difference can become very large if we move far enough from the surface, along the equatorial plane. We shall use (1.112) only with a careful control of the errors, which have a relative magnitude of $\sim 10^{-3}$, and only close to the surface of the earth. We might think then that a better approximation is given by

$$W_S = \frac{GM}{r} + \frac{1}{2}\omega^2(x^2 + y^2)\,; \tag{1.113}$$

this potential in fact contains at least the centrifugal effects, so that $W - W_S$ is a Newtonian potential harmonic in Ω. Yet, if one takes an equipotential of W_S, (i.e., $W_S = W_{S_0}$), close to the earth sphere, for which we fix a conventional radius

of $R = 6,371\,\text{km}$, a simple computation shows that its flattening, defined as

$$f = \frac{a-b}{a}, \tag{1.114}$$

with a the equatorial radius and b the polar radius, has a value approximately equal to

$$f \sim \frac{1}{2}\frac{\omega^2 R^3}{GM}. \tag{1.115}$$

In fact, after putting

$$\frac{GM}{a} + \frac{1}{2}\omega^2 a^2 = W_{S_0}$$
$$\frac{GM}{b} = W_{S_0}$$

one derives (1.115) considering that terms containing ω^2 are just smaller perturbations of the others.

The value of the parameter

$$\mu = \frac{\omega^2 R^3}{GM} \sim 3.4 \cdot 10^{-3}, \tag{1.116}$$

known also as *Clairant constant* (cf. Heiskanen and Moritz 1967), used in (1.115), yields a value of f which is about one half of the true one which, already from the end of the eighteenth century, was known to be $f \sim \mu \sim 3.4 \cdot 10^{-3}$ (cf. Todhunter 1873).

This is because the model (1.113) is basically that of a rigid layered sphere, with the addition of the centrifugal potential, while the real physical body of the earth, as it is non-rigid, reacts to self-gravitation and centrifugal force by displacing the masses from poles to the equator, thus increasing the flattening, as a matter of fact more or less doubling the value (1.115). So we use here the model (1.113) only to give a representation of its equipotential surfaces, because they give a qualitative understanding of the complex effect created by the presence of the centrifugal potential.

The situation is schematically presented in Fig. 1.13

An appropriate model of the actual gravity field is obtained by the so-called *normal potential*.

This is by definition a model, which we can write as

$$U = V_e + V_c = V_e + \frac{1}{2}\omega^2(x^2 + y^2) \tag{1.117}$$

where V_e has to be a potential harmonic outside the reference figure.

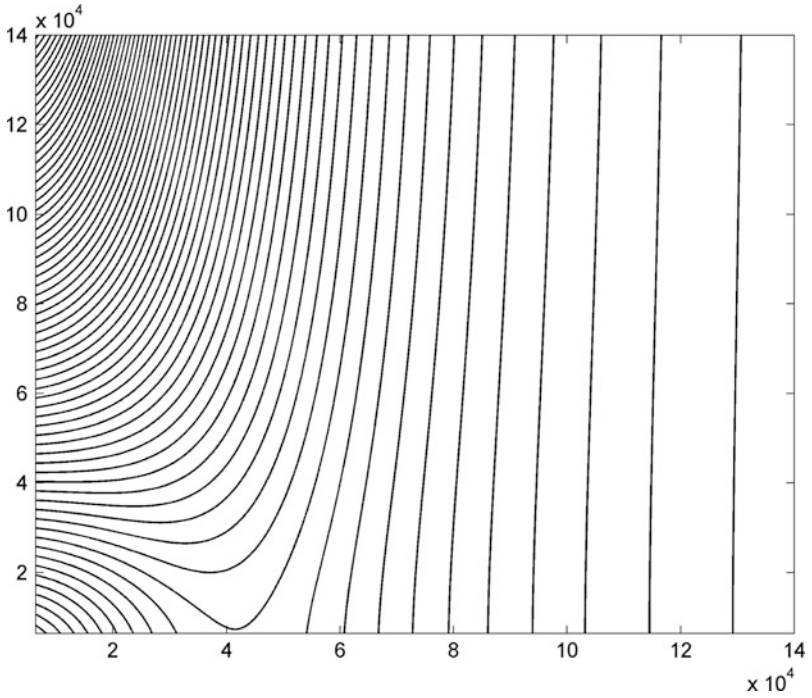


Fig. 1.13 Outlook of equipotential surfaces of W_S, cut on an upper meridian plane. Equidistance is with 2,000 km in Z from 6,400 to 140,000 km. The figure is symmetric around Z and with respect to the equatorial plane

The determination of V_e is done by assuming that one equipotential surface of (1.117), $U(P) = U_0$, is an ellipsoid of revolution, with semi-axes a and b, and that V_e is regular at infinity, namely that $V_e \to 0$ when $r \to \infty$. That the geoid, understood as one of the equipotential surfaces of W which are close to the sea surface, could be well-approximated by an ellipsoid of revolution has been established at the end of the eighteenth century, after the long-standing quarrel initiated by Newton and Cassini, as a result of the famous expeditions organized by the French Academy of Science, to measure arcs of meridians in Ecuador (C.M. de La Condamine) and Lapland (P.L. de Maupertuis and A.C. Clairaut).

From the above discussion we understand that the model we are going to construct in the end will depend only on four parameters: the shape parameters of the ellipsoid (a, b), or alternatively (a, E) or (a, f), the angular velocity ω and the value of U_0. This last parameter, as we shall see, can be substituted by the much more physically meaningful value of the constant GM.

Having to do with the solution of the Laplace equation in the exterior of an ellipsoid, it is only natural to use the ellipsoidal coordinates (1.103) and the corresponding representation of the Laplacian, (1.110).

The boundary condition to be satisfied by U on the ellipsoid $\mathcal{E}$ is basically written as (recall that $E^2 = a^2 - b^2$)

$$U_{\mathcal{E}} = V_e|_{\mathcal{E}} + \frac{1}{2}\omega^2(q^2 + E^2)\sin^2\overline{\vartheta}|_{\mathcal{E}} \tag{1.118}$$
$$= V_e|_{\mathcal{E}} + \frac{1}{2}\omega^2 a^2 \sin^2\overline{\vartheta} = U_0.$$

As a matter of fact (1.118) has to be read in the form

$$V_e|_{\mathcal{E}} = \left(U_0 - \frac{1}{2}\omega^2 a^2 \sin^2\overline{\vartheta}\right), \tag{1.119}$$

and V_e has to satisfy (1.110) for $q > b$.

As proved in Sect. A.4, the solution to this problem is given by the closed formula

$$V_e(q, \overline{\vartheta}) = \left(U_0 - \frac{1}{3}\omega^2 a^2\right)\frac{\arctan\dfrac{E}{q}}{\arctan\dfrac{E}{b}} + \frac{1}{2}\omega^2 a^2 \frac{Q(q)}{Q(b)}\left(\frac{2}{3} - \sin^2\overline{\vartheta}\right), \tag{1.120}$$

where $Q(q)$ is the function (see (1.254))

$$Q(q) = (3q^2 + E^2)\arctan\frac{E}{q} - 3qE. \tag{1.121}$$

Let us see how to express V_e as a function of (a, E, ω^2) and of the constant GM, which we assume to know, since nowadays it can be deduced from satellite tracking observations. This target can be readed by expressing U_0 as function of a, E, ω^2, GM and then substituting in (1.120).

Recalling (1.16) and noting that a mass distribution generating $V(P)$ must have its barycenter at the origin for symmetry reasons, i.e., $\mathbf{b} = 0$, we must have

$$V_e(q, \vartheta) = \frac{GM}{r} + O\left(\frac{1}{r^3}\right), \tag{1.122}$$

when $r \to \infty$.

On the other hand, since

$$q = r\sqrt{1 - \frac{E^2}{r^2}\sin^2\overline{\vartheta}} = r + O\left(\frac{1}{r}\right),$$

we have also

$$\frac{1}{q} = \frac{1}{r} + O\left(\frac{1}{r^3}\right) \text{ or } \frac{1}{r} = \frac{1}{q} + O\left(\frac{1}{q^3}\right). \tag{1.123}$$

Accordingly, (1.122) implies

$$V_e(q, \vartheta) = \frac{GM}{q} + O\left(\frac{1}{q^3}\right). \tag{1.124}$$

But, from (1.121),

$$Q(q) = O\left(\frac{1}{q^3}\right), \tag{1.125}$$

as the reader is invited to verify.

So (1.120), with (1.123)–(1.125), tells us that

$$\frac{GM}{q} + O\left(\frac{1}{q^3}\right) = \left(\frac{U_0 - \frac{1}{3}\omega^2 a^2}{\arctan \frac{E}{b}}\right)\frac{E}{q} + O\left(\frac{1}{q^3}\right),$$

i.e., multiplying by q and taking $q \to \infty$,

$$U_0 = \frac{1}{3}\omega^2 a^2 + \frac{GM}{E} \arctan \frac{E}{b}, \tag{1.126}$$

which is the sought relation (cf. Heiskanen and Moritz 1967). With (1.126) we can rewrite V_e as

$$V_e = \frac{GM}{E} \arctan \frac{E}{q} + \frac{1}{2}\omega^2 a^2 \frac{Q(q)}{Q(b)}\left(\frac{2}{3} - \sin^2 \overline{\vartheta}\right). \tag{1.127}$$

Definition of anomalous potential. Let us first define the anomalous potential T, as

$$\begin{aligned} T(P) &= W(P) - U(P) \\ &= V(P) + V_c(P) - V_e(P) - V_c(P) \\ &= V(P) - V_e(P). \end{aligned} \tag{1.128}$$

We see that T has two fundamental properties. Namely $T(P)$ is harmonic in Ω, i.e.,

$$\Delta T = 0, \quad \text{in } \Omega, \tag{1.129}$$

because both V and V_e are harmonic functions in this domain.

Let us note immediately here that since $V_e(P)$ happens to be harmonic even well inside the ellipsoid $\mathcal{E}$, through most of the earth $T(P)$ satisfies the Poisson equation

$$\Delta T(P) = \Delta V(P) = -4\pi\rho(P). \tag{1.130}$$

As a matter of fact it is possible to define another potential $\overline{V}$ which coincides with V_e outside the ellipsoid, but it is on the same time generated by a mass distribution internal to $\mathcal{E}$ consistent with such external values (Sünkel and Tscherning 1981; Tscherning and Poder 1981). In any case, outside the ellipsoid, (1.130) holds true.

Furthermore, if we choose for M in (1.127) the same value as that of the earth mass, when $r \to \infty$ we find

$$T(P) = \frac{GM}{r} + O\left(\frac{1}{r^2}\right) - \frac{GM}{r} + O\left(\frac{1}{r^3}\right) = O\left(\frac{1}{r^2}\right).$$

If we further choose the reference system (X, Y, Z) by placing its origin at the barycenter of masses, we have, also recalling (1.16) with $\mathbf{b} = 0$,

$$V(P) = \frac{GM}{r} + O\left(\frac{1}{r^3}\right),$$

which used in (1.128) gives the exact asymptotic condition

$$T(P) = O\left(\frac{1}{r^3}\right) \tag{1.131}$$

when $r \to \infty$. This is the second of the two properties mentioned above.

Note that (1.131) holds under the condition that barycenter of the masses, origin of (X, Y, Z), and center of the ellipsoid E are all placed at one and the same point.

Normal gravity vector. By using formula (1.237) and the expressions (1.106) for the vectors $\mathbf{h}_q$, $\mathbf{h}_{\overline{\vartheta}}$ and (1.107), we can compute the vector $\boldsymbol{\gamma} = \nabla U$, i.e., the normal gravity vector, as

$$\begin{aligned} \boldsymbol{\gamma}(q, \overline{\vartheta}) &= \nabla U = \nabla(V_e + V_c) \\ &= \frac{m^2}{p^2}\mathbf{h}_q\left(\frac{\partial V_e}{\partial q} + \frac{\partial V_c}{\partial q}\right) + \frac{1}{p^2}\mathbf{h}_{\overline{\vartheta}}\left(\frac{\partial V_e}{\partial \overline{\vartheta}} + \frac{\partial V_c}{\partial \overline{\vartheta}}\right), \end{aligned} \tag{1.132}$$

where

$$\begin{cases} \dfrac{\partial V_e}{\partial q} = -\dfrac{GM}{E}\dfrac{E}{q^2+E^2} + \dfrac{1}{2}\omega^2 a^2 \dfrac{Q'(q)}{Q(b)}\left(\dfrac{2}{3} - \sin^2\overline{\vartheta}\right) \\ Q'(q) = 6q \arctan\dfrac{E}{q} - \dfrac{(3q^2+E^2)E}{q^2+E^2} - 3E \\ \dfrac{\partial V_c}{\partial q} = \omega^2 q \sin^2\overline{\vartheta} \end{cases} \tag{1.133}$$

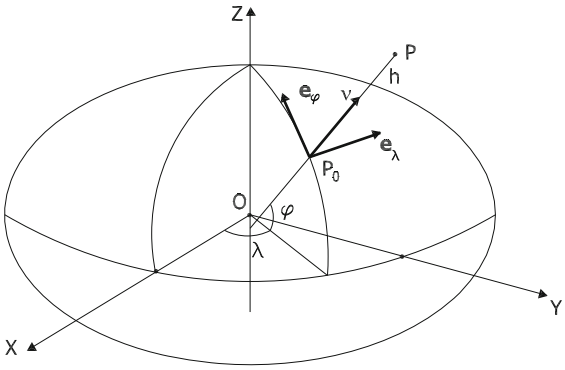


Fig. 1.14 The point P, its geodetic-ellipsoidal coordinates (λ, φ, h) and the triad $(\mathbf{e}_\lambda, \mathbf{e}_\varphi, \boldsymbol{\nu})$

$$\begin{cases} \dfrac{\partial V_e}{\partial \overline{\vartheta}} = -\omega^2 a^2 \dfrac{Q(q)}{Q(b)} \sin\overline{\vartheta}\cos\overline{\vartheta} \\ \dfrac{\partial V_c}{\partial \overline{\vartheta}} = \omega^2 (q^2 + E^2) \sin\overline{\vartheta}\cos\overline{\vartheta}. \end{cases} \tag{1.134}$$

The (1.106), (1.107), (1.132), (1.133) and (1.134) provide the exact expression of the normal gravity at every point in space, the ellipsoidal coordinates of which can be derived from Cartesian coordinates inverting (1.103) (see Remark 4).

By the formula

$$\gamma(q, \overline{\vartheta}) = \left\{ \frac{m^2}{p^2} \left(\frac{\partial V_e}{\partial q} + \frac{\partial V_c}{\partial q} \right)^2 + \frac{1}{p^2} \left(\frac{\partial V_e}{\partial \overline{\vartheta}} + \frac{\partial V_c}{\partial \overline{\vartheta}} \right)^2 \right\}^{1/2}, \tag{1.135}$$

we can compute as well the modulus of the normal gravity vector.

Remark 4. Since we often label points in space by means of geodetic ellipsoidal coordinates (λ, φ, h) it is also interesting to have γ and $\boldsymbol{\gamma}$ as functions of such coordinates, with, in addition, $\boldsymbol{\gamma}$ represented in components with respect to the usual geodetic triad $(\mathbf{e}_\lambda, \mathbf{e}_\varphi, \boldsymbol{\nu})$ pointing from P to east, north and up respectively.

The definition of such quantities is presented in Fig. 1.14 and their analytic relations between geodetic coordinates and (x, y, z) is

$$\begin{vmatrix} x \\ y \\ z \end{vmatrix} = \begin{vmatrix} (\mathcal{N} + h)\cos\varphi\cos\lambda \\ (\mathcal{N} + h)\cos\varphi\sin\lambda \\ \left[(1 - e^2)\mathcal{N} + h\right]\sin\varphi \end{vmatrix}. \tag{1.136}$$

In (1.135) $\mathcal{N}$ is the grand normal, i.e., the curvature radius of the section of the ellipsoid orthogonal to the meridian in P_0 (the orthogonal projection of P on $\mathcal{E}$), and it is given by

$$\mathcal{N} = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi}}. \tag{1.137}$$

Let us remember too that $(\mathbf{e}_\lambda, \mathbf{e}_\varphi, \boldsymbol{\nu})$ are represented, in Cartesian components, by the vectors

$$\mathbf{e}_\lambda = \begin{vmatrix} -\sin\lambda \\ \cos\lambda \\ 0 \end{vmatrix}, \ \mathbf{e}_\varphi = \begin{vmatrix} -\sin\varphi\cos\lambda \\ -\sin\varphi\sin\lambda \\ \cos\varphi \end{vmatrix}, \ \boldsymbol{\nu} = \begin{vmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\lambda \end{vmatrix}. \tag{1.138}$$

So in principle the problem we are talking about is just one of having exact formulas, and computer routines, to perform the direct and inverse transformations

$$\begin{matrix} (\lambda, \beta, q) & \leftrightarrow & (x, y, z) & \leftrightarrow & (\lambda, \varphi, h) \\ \text{(reduced ellipsoidal)} & & \text{(cartesian)} & & \text{(geodetic-ellipsoidal)} \end{matrix}.$$

In this way we can compute $U(h, \varphi), \gamma(h, \varphi)$ as well as

$$\boldsymbol{\gamma} = \mathbf{e}_\varphi[\mathbf{e}_\varphi \cdot \boldsymbol{\gamma}] + \boldsymbol{\nu}[\boldsymbol{\nu} \cdot \boldsymbol{\gamma}]. \tag{1.139}$$

We note that in (1.139) there is no eastward component of $\boldsymbol{\gamma}$, since this vector lies in the meridian plane, i.e., that of $\boldsymbol{\nu}$ and $\mathbf{e}_\varphi$.

The two transformations $(\lambda, \beta, h) \rightarrow (x, y, z)$ and $(\lambda, \varphi, h) \rightarrow (x, y, z)$ are already given by (1.103) (remember that $\beta = \pi/2 - \overline{\vartheta}$) and (1.136) respectively. As for the inverse transformations one can write first

$$tg\lambda = \frac{y}{x}, \tag{1.140}$$

which is valid for both reduced and geodetic ellipsoidal coordinates. Of course, when one is inverting (1.140), the signs of x and y have to be considered in order to place λ in the right quadrant. Then for the reduced ellipsoidal coordinates (β, q) one has the explicit solution

$$\begin{cases} q = \dfrac{1}{\sqrt{2}}\left[r^2 - E^2 + \sqrt{(r^2 - E^2)^2 + 4E^2z^2}\right]^{1/2} \\ tg\beta = \dfrac{\sqrt{q^2 + E^2}}{q}\dfrac{z}{\rho}, \end{cases} \tag{1.141}$$

where

$$r^2 = \rho^2 + z^2, \ \rho^2 = x^2 + y^2. \tag{1.142}$$

As for the geodetic ellipsoidal coordinates one can use the following exact algorithm

$$\begin{cases} \varphi = \arctan \dfrac{z + (e')^2 b \sin^3 \psi}{\rho - e^2 a \cos^3 \psi} \\ h = \dfrac{\rho}{\cos \varphi} - \mathcal{N} \end{cases} \tag{1.143}$$

where $e' = \sqrt{\frac{a^2-b^2}{b^2}}$ is the second eccentricity of the ellipsoid, and

$$\psi = \arctan \frac{a}{b} \frac{z}{\rho} \,. \tag{1.144}$$

More on this subject can be found in the book (Awange and Grafarend, 2005). Although the problem can be exactly solved, many times it is useful to employ approximate formulae, valid in the surrounding of the earth surface, such as the famous Cassinis formula (cf. Heiskanen and Moritz 1967; Moritz 2000),

$$\begin{aligned} \gamma(\varphi, h) &= 978.0327715(1 + 5.30244 \cdot 10^{-3} \sin^2 \varphi \\ &-5.8 \cdot 10^{-6} \sin^2 2\varphi) - (0.30877 - 4.510^{-4} \sin^2 \varphi)h \\ &+72 \cdot 10^{-6} h^2, \end{aligned} \tag{1.145}$$

where the (ellipsoidal) height h has to be given in km and the gravity γ is in Gal. Similarly one can derive an approximate formula for $\gamma_\varphi = \boldsymbol{\gamma} \cdot \mathbf{e}_\varphi$, valid in the topographic layer, with a relative accuracy of the order of 10^{-9}, namely

$$\gamma_\varphi = 5.185960 \cdot \frac{h}{a} \sin 2\varphi,$$

with h (and a) in kilometers and γ_φ in Gal.

1.10 Anomalous Quantities of the Gravity Field and a More Precise Definition of the Geoid

Anomalous potential T. The first and most important anomalous quantity of the gravity field, we have already defined in (1.128); this is the anomalous potential T

$$T(P) = W(P) - U(P). \tag{1.146}$$

As we have already noted, $U(P)$, by adapting its four parameters, provides an excellent approximation of the gravity potential, in the sense that

$$O(T) \sim 10^{-5} O(W) \,; \tag{1.147}$$

therefore T is an ideal unknown field when we shall treat non-linear functionals of it, since the linearization procedures that we will apply will be good with a relative error of the order of 10^{-10}, negligible in the context of the arguments discussed in these notes.

Let us recall that in mathematical terms $Y = 0(\varepsilon)$ means that $0 < A \leq \left|\frac{Y}{\varepsilon}\right| \leq B$, ($A$, B constants); here however we extend the meaning of the symbol to represent the physical order of magnitude of the quantity in parenthesis, or of its maximum absolute value, when it is a function.

We stress again that (1.147) is certainly correct in a neighborhood of the earth surface and therefore a fortiori in the outer space because T is harmonic in Ω and harmonic functions attain their extreme values at the boundary (cf. Part III, Chap. 13, Theorem 4). On the contrary, if we move well inside the masses, W is not anymore harmonic, while U apart from the centrifugal component that cancels with that of W, is in fact still harmonic, with the exception of a small disk (the so-called focal disk) centered at the origin O and lying on the equatorial plane. Therefore the behaviour of W and U start diverging and already at 100 km inside the masses one order of magnitude is lost.

This point is so important that we try to illustrate it by an elementary example.

Example 5. Take a non-rotating spherical planet with an inner sphere, with radius R_0 (= 6,300 km), with a mass content $M_0 \cong 6 \cdot 10^{27}$ gr, and an outer shell (the crust), with a thickness $\delta R_c = 100$ km and a mass density $\rho \sim 2.67$ gr/cm^3, implying a mass $M_c \cong 10^{26}$ gr. Note that $M_0 + M_c$ is roughly equal to the actual mass of the earth.

We take as normal gravity just

$$U = G\frac{(M_0 + M_c)}{r}$$

so that it coincides with W for $r = R = R_0 + \delta R_c$.

However, when we go on the inner surface S_0 we have (cf. Example 2)

$$W|_{S_0} = \frac{GM_0}{R_0} + 2\pi G\rho\left(R^2 - R_0^2\right)$$

$$U|_{S_0} = \frac{GM_0}{R_0} + \frac{GM_c}{R_0} = \frac{GM_0}{R_0} + \frac{4}{3}\pi G\rho\frac{(R^3 - R_0^3)}{R_0}$$

So we can directly compute the relative error $\dfrac{U|_{S_0} - W|_{S_0}}{U|_{S_0}}$, i.e., performing some manipulations and writing $U|_{S_0} = \gamma_0 R_0$ in the denominator and $\frac{4}{3}(R^3 - R_0^3) \sim 4R_0^2(R - R_0)$, we get

$$\left|\frac{U|_{S_0} - W|_{S_0}}{U|_{S_0}}\right| = \frac{2\pi G\rho R_0}{\gamma_0}\left(\frac{R - R_0}{R_0}\right)^2 \sim 2 \cdot 10^{-4},$$

namely an error one order of magnitude larger than the actual anomalous potential.

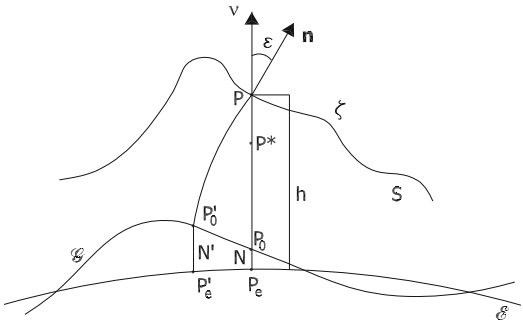


Fig. 1.15 Different heights of P and their relation to $h_P = P_e P, N = P_e P_0,\ H = \overline{PP'}_0$ (along the plumb-line), $h^* = \overline{P_e P}^*, \zeta = \overline{P^* P}$

Geoid undulation N. Next we define a geometric anomaly describing the difference between the geoid $\mathcal{G}$ and the ellipsoid $\mathcal{E}$, namely the geoid undulation N. By definition this is a function of $P_e \in \mathcal{E}$, i.e., of (λ, φ), and it is the height of $\mathcal{G}$ on P_e, measured along the direction of $\boldsymbol{\nu}(P_e)$. Said in another way, if P_0 is the intersection with $\mathcal{G}$ of the normal to the ellipsoid passing through P_e, then (cf. Fig. 1.15)

$$N(P_e) = h(P_0). \tag{1.148}$$

We note immediately that indeed $N(P_e)$ can be either positive or negative, of course depending on the reciprocal definition of $\mathcal{E}$ and $\mathcal{G}$. Here we need to stress that when $W|_{\mathcal{G}} = W_0 = U_0 = U|_{\mathcal{E}}$, and both $\mathcal{G}$ and $\mathcal{E}$ are close (within meters) to the mean surface of the ocean, then N is known experimentally to satisfy the upper bound

$$|N| \leq 120\,\text{m}, \tag{1.149}$$

with a mean square value, approximately given by

$$\{E[N^2]\}^{1/2} = \left\{ \frac{1}{4\pi} \int_\sigma N(\varphi, \lambda)^2 d\sigma \right\}^{1/2} \cong 35\,\text{m}. \tag{1.150}$$

Whence, summarizing, we could say that N is a signal of the order of magnitude of

$$O(N) = 10^{-5} R \tag{1.151}$$

$$R = \text{mean earth radius}.$$

Before we proceed let us establish a relation which is very much used, nowadays that the ellipsoidal height of a point has become available by GPS measurements.

This relation is

$$h_P = H_P + N_{P_e} \tag{1.152}$$

and, though not exact, we claim that it is valid with an error of less than 1 mm for all points of the earth surface.

We start by observing Fig. 1.15, warning the reader that in such a figure the curvature of the plumb-line has been enormously pronounced.

Note also that ε will then be the inclination of $\mathcal{G}$ with respect to $\mathcal{E}$, along the section shown by Fig. 1.15.

When determining orders of magnitude, we can well assume that the plumb-line $\overline{P_0' P}$ has the same inclination ε with respect to $\overline{P_e P}$. A key point is that ε, which we shall study in more detail in the sequel of the section, is experimentally known to be 1 arcmin as a maximum, over the whole surface S,

$$|\varepsilon| \leq 1\text{arcmin} \cong 3 \cdot 10^{-4}\,\text{rad}. \tag{1.153}$$

Accordingly we can claim that even if P is on the top of a mountain 6 km high,

$$\overline{P_e' P_e} \cong H \sin\varepsilon \cong H\varepsilon = 18 \cdot 10^{-4}\,\text{km} = 1.8\,\text{m}. \tag{1.154}$$

Now since $\boldsymbol{\nu}$ is orthogonal to $\mathcal{E}$, the arc $\overline{P_e' P_e}$, less then 2 m long, can certainly be considered as a segment orthogonal to $\overline{P_e P}$. Since if we project the line $(P_e' P_0') \cup (P_0' P)$ orthogonally onto $\boldsymbol{\nu}$ we get exactly h, i.e., calling $P_{\boldsymbol{\nu}}$ the orthogonal projector on $\boldsymbol{\nu}$,

$$\begin{aligned} h &= \left| P_{\boldsymbol{\nu}} \mathbf{r}_{P_0' P} + P_{\boldsymbol{\nu}} \mathbf{r}_{P_e' P_0'} \right| \\ &= H \cos\varepsilon + N'. \end{aligned} \tag{1.155}$$

But (with $H = 6\,\text{km}$!)

$$H \cos\varepsilon \cong H - \frac{1}{2}\varepsilon^2 H\ ; \tag{1.156}$$

and also

$$N' - N \cong \overline{P_e' P_e} tg\varepsilon \cong H\varepsilon^2 \tag{1.157}$$

so that rewriting (1.155) in the form

$$h = H - \frac{1}{2}\varepsilon^2 H + N + \varepsilon^2 H = H + N + \frac{1}{2} H\varepsilon^2 \tag{1.158}$$

we see that (1.152) holds true with an error equal to

$$\frac{1}{2} H\varepsilon^2 \cong 0.3\,\text{mm}. \tag{1.159}$$

Since the case used as an example is really extreme, we consider our statement as proved.

To make exact the definition of geoid undulation, we need to establish on a more solid ground the definition of $\mathcal{G}$ and its relation to $\mathcal{E}$. As a matter of fact there is no unique way to solve such a problem, also because the ellipsoid $\mathcal{E}$, and the attached normal potential U, have to be defined with an approximation purpose, so that any change small enough of their parameters will provide us with another potential as good as the first for our target.

So here we just choose a way to define $\mathcal{G}$ which seems to us linear and clear. Let us start with $\mathcal{E}$. Since the barycenter of the earth can be determined by means of spatial geodetic techniques, we will consider its time-averaged position as known. Similarly, we take as given the direction of the mean rotation axis, with respect to the body of the earth. So we can define $\mathcal{E}$ as an ellipsoid of revolution, centered at the barycenter of masses and with the symmetry polar axis directed as Z.

The geometric flattening of $\mathcal{E}$, $f = (a - b)/a$, or better a kind of its mean value for the actual earth, can be very accurately determined from satellite tracking. So we are left with one geometric parameter only to be fixed and we choose it to be the equatorial radius a. We note here too, that from satellite radar-altimetry we are able today to determine the geometric shape of the ocean surface with an accuracy better than 5 cm, as an average over a footprint of several hundreds meters. Such a surface should be equipotential if there were no currents in the ocean; yet the presence of such (almost) stationary currents, like the Gulf Stream or the Kuroshyo, do impress a stationary deformation to the sea surface with respect to $\mathcal{G}$.

But the magnitude of the separation between the two surfaces, is within a range of a few meters maximally. So it makes sense to say that a is chosen so that $\mathcal{E}$ is close to the ocean surface within meters (what makes a difference of the order of magnitude of $10^{-6}R$).

In the range of meters a can be chosen arbitrarily, i.e., it can be conventionally fixed. The value accepted today is

$$a = 6,378,136.62\,\mathrm{m} \tag{1.160}$$

with an accuracy of $\pm\, 0.10$ m.

Moreover we take as known also the value of ω^2, which is well-observable by astro-geodetic means. Finally we know that GM can also be determined by satellite tracking

$$GM = 398,600,441.5 \cdot 10^6\,\mathrm{m^3\,s^{-2}} \tag{1.161}$$

with an accuracy of the order of $\pm 0.8 \cdot 10^6\,\mathrm{m^3\,s^{-2}}$ (cf. Moritz 2000).

So, once the shape of $\mathcal{E}$ and its placement in space have been secured, a corresponding normal potential U can be computed and the value U_0, attained by U on $\mathcal{E}$, can be computed by (1.126).

By definition the geoid $\mathcal{G}$ is the equipotential surface such that

$$W(P) = W_0 = U_0. \tag{1.162}$$

We note once more explicitly that due to our hypotheses U and W will have the same $\frac{GM}{r}$ term, when $r \to \infty$. Furthermore, since the respective barycenters are placed at the origin O of (X, Y, Z), we may conclude that (1.131) has to hold, namely

$$T(P) = O\left(\frac{1}{r_P^3}\right),\ r_P \to \infty. \tag{1.163}$$

Height anomaly ζ_P. The definition of geoid undulation generalizes to a function defined in space, called the *height anomaly*, $\zeta(P)$.

First we define the so-called normal height $h^*(P)$ as follows: take the line containing the segment $P_e P$ of Fig. 1.15 and find on it the point P^* such that

$$U(P^*) = W(P). \tag{1.164}$$

Then by definition we put

$$h^*(P) = h(P^*). \tag{1.165}$$

Note that (1.164) and (1.165) defines a mapping in space between the points P and P^* according to the relation

$$P \equiv (\lambda, \varphi, h) \ \to \ P^* \equiv (\lambda, \varphi, h^*). \tag{1.166}$$

Through the mapping (1.166) the surface of the earth S is mapped onto another surface, called the *telluroid*, S^* (cf. Heiskanen and Moritz 1967)

$$S^* \equiv \{P^* \equiv (\lambda, \varphi, h_P^*),\ P \in S\}. \tag{1.167}$$

Now, we can define the height anomaly of P as

$$\zeta(P) = h(P) - h^*(P)\ ; \tag{1.168}$$

so, when $P \in S$, ζ_P is basically the separation of the earth surface S with respect to the telluroid S^*.

Let us immediately state that $\zeta(P)$ can be either positive or negative, depending on P. We also see that, according to our Definition (1.162), if P is directly taken on the geoid $\mathcal{G}$, then

$$P \in \mathcal{G} \Rightarrow W(P) = W_0 = U_0 = U(P_e) \to P^* = P_e,$$

i.e., the height anomaly ζ_P becomes the geoid undulation N. Such is the case, with a very good approximation, for every point on the surface of the sea.

Gravity disturbance $\delta\mathbf{g}, \delta g$. The vector gravity disturbance $\delta\mathbf{g}$ is by definition

$$\delta\mathbf{g}(P) = \mathbf{g}(P) - \boldsymbol{\gamma}(P). \tag{1.169}$$

On the other hand the scalar gravity disturbance, or simply gravity disturbance, δg is

$$\delta g(P) = g(P) - \gamma(P). \tag{1.170}$$

Note immediately that, contrary to the convention used almost everywhere in the text, in this case it is

$$\delta g(P) \neq |\delta\mathbf{g}(P)|.$$

In fact we first note that $|\delta\mathbf{g}|$ has the following order of magnitude as a maximum

$$O(|\delta\mathbf{g}|) \cong 10^{-4}\gamma, \tag{1.171}$$

so that we are allowed to linearize expressions in $\dfrac{|\delta\mathbf{g}|}{\gamma}$, neglecting terms of the order of 10^{-8}.

Then we find, from (1.170) and recalling that $\boldsymbol{\nu}(P) = -\dfrac{\boldsymbol{\gamma}(P)}{\gamma(P)}$, with a high degree of approximation,

$$\begin{aligned} g(P) &= |\mathbf{g}(P)| = |\boldsymbol{\gamma}(P) + \delta\mathbf{g}(P)| \\ &= \sqrt{\gamma^2(P) + 2\boldsymbol{\gamma}(P)\cdot\delta\mathbf{g}(P) + |\delta\mathbf{g}(P)|^2} \\ &\cong \gamma(P)\sqrt{1 - 2\frac{\boldsymbol{\nu}(P)\cdot\delta\mathbf{g}(P)}{\gamma(P)}} \\ &\cong \gamma(P)\left\{1 - \frac{\delta g_\nu(P)}{\gamma(P)}\right\} = \gamma(P) - \delta g_\nu(P) \end{aligned} \tag{1.172}$$

where we have called δg_ν the vertical component of $\delta\mathbf{g}$, putting

$$\delta g_\nu = \delta\mathbf{g}\cdot\boldsymbol{\nu}. \tag{1.173}$$

But (1.172) implies

$$\delta g = g(P) - \gamma(P) \cong -\delta g_\nu, \tag{1.174}$$

proving our claim that δg is equal to one component only of $\delta\mathbf{g}$ and not to the whole modulus, $|\delta\mathbf{g}|$.

Free air gravity anomaly, $\Delta\mathbf{g}, \Delta g$. In a way very similar to (1.169) and (1.170) we set up the definition of this new anomaly as

$$\Delta\mathbf{g} = \mathbf{g}(P) - \boldsymbol{\gamma}(P^*) \tag{1.175}$$

$$\Delta g = g(P) - \gamma(P^*). \tag{1.176}$$

First of all note that here g and γ are computed at two different points; in particular, when P is on the earth surface S, P^* is on the telluroid S^*, so that $\Delta\mathbf{g}$ and Δg can be considered as functions of either P or P^*.

Again here it is not true that $\Delta g(P)$ is equal to $|\Delta\mathbf{g}(P)|$.

In this case we can find a relation between $\Delta g, \delta g$ and ζ, in fact from (1.175) we have

$$\Delta g = g(P) - \gamma(P) + \gamma(P) - \gamma(P^*) = \delta g(P) + \gamma(P) - \gamma(P^*). \tag{1.177}$$

But since $\overline{P^*P} = \zeta$, which is a small quantity, we can approximate (1.177) by

$$\begin{aligned} \Delta g &\cong \delta g(P) + \frac{\partial\gamma(P^*)}{\partial h}\zeta(P) \\ &= -\delta g_\nu + \frac{\partial\gamma}{\partial h}\zeta, \end{aligned} \tag{1.178}$$

expression that will become very useful in the sequel.

Deflection of the vertical $\boldsymbol{\varepsilon}, (\eta, \xi)$. We put by definition

$$\boldsymbol{\varepsilon}(P) = \mathbf{n}(P) - \boldsymbol{\nu}(P). \tag{1.179}$$

The first thing to observe is that since both $\mathbf{n}$ and $\boldsymbol{\nu}$ have modulus equal to 1 and $\boldsymbol{\varepsilon}$ is a very small vector,

$$O(|\boldsymbol{\varepsilon}|) \cong 10^{-4}, \tag{1.180}$$

we can safely put

$$\boldsymbol{\varepsilon} \cdot \mathbf{n} \cong \boldsymbol{\varepsilon} \cdot \boldsymbol{\nu} \cong 0 \tag{1.181}$$

and on the same time

$$|\boldsymbol{\varepsilon}| = \varepsilon \cong \text{ angle between } \mathbf{n} \text{ and } \boldsymbol{\nu} \text{ (in rad)}. \tag{1.182}$$

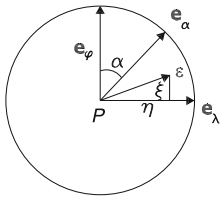


Fig. 1.16 The circumference of directions in the horizontal plane through P, as seen from above

The relation (1.181) tells us that $\boldsymbol{\varepsilon}$ lies in the horizontal plane, so that we can put

$$\boldsymbol{\varepsilon} = \eta \mathbf{e}_\lambda + \xi \mathbf{e}_\varphi \; ; \tag{1.183}$$

the two components η and ξ are the eastward the northward deflections of the vertical.

In particular if we take any vertical plane (i.e., a plane through P containing $\boldsymbol{\nu}_P$) with azimut α with respect to the north (see Fig. 1.16), we find that the projection of $\boldsymbol{\varepsilon}$ on this plane is given by

$$\varepsilon_\alpha = \boldsymbol{\varepsilon} \cdot \mathbf{e}_\alpha = \xi \cos\alpha + \eta \sin\alpha \; . \tag{1.184}$$

In order to fully understand the geometric and the physical significance of $\boldsymbol{\varepsilon}$, we shall find its relation on one side with the "horizontal" coordinates of P, namely with (Λ, Φ) and (λ, φ), on the other side with the gravity disturbance vector $\delta\mathbf{g}$. Remember that, in Cartesian geocentric components,

$$\mathbf{n}_P = \begin{vmatrix} \cos\Phi\cos\Lambda \\ \cos\Phi\sin\Lambda \\ \sin\Phi \end{vmatrix}, \; \boldsymbol{\nu}_P = \begin{vmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{vmatrix} . \tag{1.185}$$

If we put in $\mathbf{n}(\Lambda, \Phi)$

$$\Lambda = \lambda + \delta\Lambda, \; \Phi = \varphi + \delta\Phi \tag{1.186}$$

and we linearize, we find

$$\begin{aligned} \mathbf{n}_P &= \begin{vmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{vmatrix} + \cos\varphi \begin{vmatrix} -\sin\lambda \\ \cos\lambda \\ 0 \end{vmatrix} \delta\Lambda + \begin{vmatrix} -\sin\varphi\cos\lambda \\ -\sin\varphi\sin\lambda \\ \cos\varphi \end{vmatrix} \delta\Phi \\ &= \boldsymbol{\nu}_P + \mathbf{e}_\varphi \delta\Phi + \cos\varphi \mathbf{e}_\lambda \delta\Lambda \end{aligned} \tag{1.187}$$

Comparing (1.183) with (1.187) we see that

$$\eta = \cos\varphi\delta\Lambda = \cos\varphi(\Lambda - \lambda), \; \xi = \delta\Phi = \Phi - \varphi \; . \tag{1.188}$$

Moreover, let us go back to the definition of vector of the vertical; if we use (1.169), (1.170) and (1.173), and perform a linear approximation in $\delta\mathbf{g}$, δg, we get

$$\begin{aligned}\mathbf{n}_P &= -\frac{\mathbf{g}_P}{g_P} = -\frac{\boldsymbol{\gamma}_P + \delta\mathbf{g}}{\gamma - \delta g_\nu} \\ &\cong \frac{\boldsymbol{\nu} - \dfrac{\delta\mathbf{g}}{\gamma}}{1 - \dfrac{\delta g_\nu}{\gamma}} \cong \left(\boldsymbol{\nu} - \frac{\delta\mathbf{g}}{\gamma}\right)\left(1 + \frac{\delta g_\nu}{\gamma}\right) \cong \\ &= \boldsymbol{\nu} - \frac{\delta\mathbf{g}}{\gamma} + \boldsymbol{\nu}\frac{\delta g_\nu}{\gamma}.\end{aligned} \tag{1.189}$$

If we recall (1.173) and we denote by P_ν the orthogonal projection on $\boldsymbol{\nu}$, we can rewrite (1.189) as

$$\boldsymbol{\varepsilon} = \mathbf{n} - \boldsymbol{\nu} = -\frac{1}{\gamma}(I - P_\nu)\delta\mathbf{g}. \tag{1.190}$$

The relation (1.190) tells us, among other things, that $\boldsymbol{\varepsilon}$ is just the horizontal component of $\delta\mathbf{g}$ divided by γ because $(I - P_\nu)$ is just the orthogonal projection on the horizontal plane in P, therefore

$$O(|\boldsymbol{\varepsilon}|) \cong \frac{1}{\gamma}O(|\delta\mathbf{g}|). \tag{1.191}$$

Finally, going back to the definition of $\delta\mathbf{g}$, we see that we can write

$$\delta\mathbf{g} = -\delta g\boldsymbol{\nu} - \gamma\boldsymbol{\varepsilon}. \tag{1.192}$$

Summarizing we could say that there is a general scheme leading to the definition of a geodetic anomaly; namely we must have a physical or geometric (or both) quantity (it can be a scalar, a vector, a tensor etc.) that we express in abstract form as

$$s = F(P; W)\,; \tag{1.193}$$

then we must define some mapping, like (1.166) but not necessarily the same,

$$P \leftrightarrow P^*\,; \tag{1.194}$$

then we define a normal quantity s^*, corresponding to s, as

$$s^* = F(P^*, U) \tag{1.195}$$

and finally the geodetic anomaly of s as

$$Ds = s - s^*. \tag{1.196}$$

In this sense, for instance, the Bouguer gravity anomaly, so much in use in geophysics, is not a geodetic anomaly, since it implies a certain density and a certain distribution of masses and even it cannot be derived from a potential, so we don't include it in this section.

1.11 Summary of Height Systems and Their Relation to the Geodetic Datum

We have seen up to here a number of height systems, i.e., the coordinates used in one way or another to fix the position of a point P in space outside some closed reference surface. It is time first of all to summarize them:

- Geopotential number:

$$C(P) = W_0 - W(P)\,; \tag{1.197}$$

  this is indeed only related to the physical body of the earth and its potential; it requires only that W_0 is fixed and that at least a point P_0 on the geoid be known.
- Dynamic height:

$$H_{\text{dyn}} = \frac{C(P)}{\gamma_0}\,; \tag{1.198}$$

  to be specific γ_0 is a fixed number equal to normal gravity on the ellipsoid at $\varphi = 45°$. This is not conceptually different from $C(P)$, it is only the same coordinate re-scaled in such a way that it is numerically close to a height in the sense of geometry.
- Orthometric height:

$$H_P = \text{length of plumb-line between } P \text{ and the geoid } \mathcal{G}\,; \tag{1.199}$$

  this is also an intrinsic coordinate in the sense that it is only related to physical quantities uniquely derived from the mass distribution. H_P is dimensionally a length and its local variation is close (though not identical) to the leveling increment.
- Geodetic ellipsoidal height:

$$\begin{aligned} h_P = {} & \text{length of the segment } PP_e, \text{ with } P_e \text{ the orthogonal} \\ & \text{projection of } P \text{ on the ellipsoid } \mathcal{E}; \end{aligned} \tag{1.200}$$

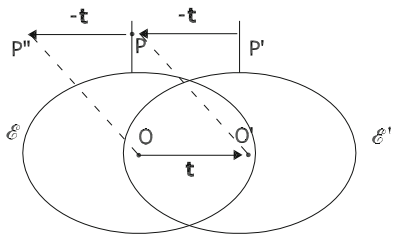


Fig. 1.17 P' has the same coordinates in $\{\mathcal{E}'\}$ as P in $\{\mathcal{E}\}$, P'' has the same coordinates in $\{\mathcal{E}\}$ as P in $(\mathcal{E}')$

the sign is inverted when P is inside $\mathcal{E}$. This is a purely geometric quantity, depending on P and on the choice of the ellipsoid $\mathcal{E}$, also called the choice of the geodetic datum; so if we move $\mathcal{E}$ leaving P fixed with respect to the earth, h_P will change.

- Normal height: if we call P_e the orthogonal projection of P onto the ellipsoid, the normal height h_P^* is defined by

$$U(P^*) = U(\mathbf{r}_{P_e} + h_P^* \boldsymbol{\nu}) \equiv W(P)\,; \tag{1.201}$$

so h_P^* is a mixed quantity which we expect to depend on both the position of P with respect to the earth, and the choice of the geodetic datum $\mathcal{E}$.

Here we want to see how h_P, h_P^* do depend on the choice of $\mathcal{E}$. This is important because both the barycenter and the rotation axis Z are not perfectly known and, even more important, they are changing in time so that we need to understand whether, for instance, every year we have to redo completely the computation of height systems or we can just account for the effects of the variations of $\mathcal{E}$ in some simple way.

That C_P, H_{dyn} and H_P do not vary with $\mathcal{E}$, we have already explained.

So let us see how are things for h. We first of all note that if we move $\mathcal{E}$ with a rototranslation, from the new position of $\mathcal{E}$, we see a point P, fixed in space, as if it had been submitted to a rototranslation opposite to the one imposed to $\mathcal{E}$.

The situation is represented, for a translation only, in Fig. 1.17

The effect of a rototranslation with infinitesimal parameters is known to have, in terms of Cartesian coordinates, the analytical representation

$$d\mathbf{r} = \begin{vmatrix} dx \\ dy \\ dz \end{vmatrix} = \mathbf{t} + \boldsymbol{\varepsilon} \wedge \mathbf{r}, \tag{1.202}$$

where $\mathbf{t}$ is the translation vector and $\boldsymbol{\varepsilon}$ the rotation vector.

Recalling that (cf. (1.136))

$$\begin{aligned}\begin{vmatrix} x \\ y \\ z \end{vmatrix} &= \begin{vmatrix} (\mathcal{N}+h)\cos\varphi\cos\lambda \\ (\mathcal{N}+h)\cos\varphi\sin\lambda \\ \left[(1-e^2)\mathcal{N}+h\right]\sin\varphi \end{vmatrix} \\ &= (\mathcal{N}+h)\begin{vmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{vmatrix} - e^2\mathcal{N}\sin\varphi\begin{vmatrix} 0 \\ 0 \\ 1 \end{vmatrix} \\ &= (\mathcal{N}+h)\boldsymbol{\nu} - e^2\mathcal{N}\sin\varphi\mathbf{e}_z\end{aligned} \tag{1.203}$$

with $e^2 = (a^2 - b^2)/a^2$ and $\mathcal{N}$ the grand normal defined in (1.137), we can differentiate such expression and compare with (1.202).

The differentiation of (1.203) is standard, though lengthy; the relations

$$\frac{\partial}{\partial h}\boldsymbol{\nu} = 0,\ \frac{\partial}{\partial\varphi}\boldsymbol{\nu} = \mathbf{e}_\varphi,\ \frac{\partial}{\partial\lambda}\boldsymbol{\nu} = \cos\varphi\mathbf{e}_\lambda \tag{1.204}$$

can help in this endeavour. The result is

$$d\mathbf{r} = \begin{vmatrix} dx \\ dy \\ dz \end{vmatrix} = dh\boldsymbol{\nu} + (\mathcal{M}+h)d\varphi\mathbf{e}_\varphi + (\mathcal{N}+h)\cos\varphi d\lambda\mathbf{e}_\lambda, \tag{1.205}$$

where $\mathcal{M}$, the radius of curvature of the meridian, is given by

$$\mathcal{M} = \frac{a(1-e^2)}{(1-e^2\sin^2\varphi)^{3/2}}. \tag{1.206}$$

By using (1.202) and (1.205), we see that

$$\mathbf{t} + \boldsymbol{\varepsilon}\wedge\mathbf{r} = dh\boldsymbol{\nu} + (\mathcal{M}+h)d\varphi\mathbf{e}_\varphi + (\mathcal{N}+h)\cos\varphi d\lambda\mathbf{e}_\lambda,$$

so that, taking the scalar product with $\boldsymbol{\nu}$, we find

$$dh = \mathbf{t}\cdot\boldsymbol{\nu} + (\boldsymbol{\varepsilon}\wedge\mathbf{r})\cdot\boldsymbol{\nu} = \mathbf{t}\cdot\boldsymbol{\nu} + (\mathbf{r}\wedge\boldsymbol{\nu})\cdot\boldsymbol{\varepsilon}. \tag{1.207}$$

If we use (1.203) we conclude however that

$$\mathbf{r}\wedge\boldsymbol{\nu} = -e^2\mathcal{N}\sin\varphi\mathbf{e}_z\wedge\boldsymbol{\nu}$$

and, since by direct inspection we see that (cf. Fig. 1.18)

$$-\mathbf{e}_z\wedge\boldsymbol{\nu} = -\cos\varphi\mathbf{e}_\lambda,$$

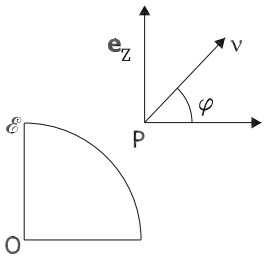


Fig. 1.18 $\mathbf{e}_z$ and $\boldsymbol{\nu}$ in the meridian plane of P

we finally obtain

$$dh = \mathbf{t} \cdot \boldsymbol{\nu} - e^2 \sin\varphi \cos\varphi \mathcal{N} \mathbf{e}_\lambda \cdot \boldsymbol{\varepsilon}. \tag{1.208}$$

If $\mathbf{t}, \boldsymbol{\varepsilon}$ are just errors in the definition of $\mathcal{E}$, they are at most of the order of centimeters or mas (milliarcseconds) and the first term is small but significant, while the second is totally irrelevant. If they represent variations over a time span of years, they can be two orders of magnitude as large and we see that the first term can become very large, and the second, though usually disregarded, enters into the centimetric range.

Whatever it is, the effect of a change of position of $\mathcal{E}$ in space can be accounted for, as for the effects on the ellipsoidal height system, by the simple formula (1.208).

At last let us see that, contrary to intuition, h_P^* does not depend, at least with an approximation to the first order in $d\mathbf{r}$, on changes of position and attitude of $\mathcal{E}$. In fact let us start from the Definition (1.201) and note that if P is fixed with respect to the earth, $W(P)$ does not change so that we must have

$$dU(\mathbf{r}_P^*) = \boldsymbol{\gamma} \cdot d\mathbf{r}_P^* = 0. \tag{1.209}$$

On the other hand

$$\mathbf{r}_P^* = \mathbf{r}_{P_e} + h^* \boldsymbol{\nu}$$

so that

$$d\mathbf{r}_P^* = d\mathbf{r}_{P_e} + dh^* \boldsymbol{\nu} + h^* d\boldsymbol{\nu}. \tag{1.210}$$

But $d\mathbf{r}_{P_e}$ is tangent to the ellipsoid and, since $\boldsymbol{\gamma}$ has a small change of direction along its plumbline, up to the level of the surface of the earth, we can reasonably put

$$\boldsymbol{\gamma}_P \cdot d\mathbf{r}_{P_e} \cong 0. \tag{1.211}$$

Similarly $d\boldsymbol{\nu} \cdot \boldsymbol{\nu} = 0$, because $\boldsymbol{\nu}$ has always modulus 1, and since $\boldsymbol{\gamma}_P$ is almost parallel to $\boldsymbol{\nu}$, we can claim that

$$\boldsymbol{\gamma}_P \cdot d\boldsymbol{\nu} \cong 0. \tag{1.212}$$

Using (1.210)–(1.212) in (1.209) we find, with the approximation above specified,

$$dU = \boldsymbol{\gamma} \cdot d\mathbf{r}_P^* = dh^* \boldsymbol{\gamma} \cdot \boldsymbol{\nu} = -\gamma dh^* = 0, \tag{1.213}$$

i.e., $dh^* = 0$.

Concluding, let us claim that

$$h_P = h_P^* + \zeta_P, \tag{1.214}$$

so that from the above reasoning we see that a variation of geodetic datum $\mathcal{E}$ has on ζ an effect primarily given by the translation $\mathbf{t}$, and more precisely

$$d\zeta_P \cong dh_P \cong -\mathbf{t} \cdot \boldsymbol{\nu} = \mathbf{t} \cdot \frac{\boldsymbol{\gamma}}{\gamma}; \tag{1.215}$$

a rotation of $\mathcal{E}$ has typically an effect two orders of magnitude smaller.

1.12 Exercises

Exercise 1. Prove that

$$\nabla \cdot [F(r)\mathbf{r}] = rF'(r) + 3F(r).$$

Then search for an $F(r)$ such that

$$\nabla \cdot [F(r)\mathbf{r}] = \frac{1}{r}$$

and prove that with the further requirement that $F(r)$ is regular at infinity, it is

$$F(r) = \frac{1}{2r}.$$

Exercise 2. Assume that B is a body with constant density δ_0. By applying the result of Exercise 1 and Gauss' theorem, prove that the Newtonian potential of B is given by

$$T(P) = \frac{1}{2} G\delta_0 \int_S \frac{\mathbf{r}_{PQ}}{r_{PQ}} \cdot \mathbf{n}_Q dS_Q,$$

i.e., the Newton integral is transformed into a surface integral.

Exercise 3. Consider a body of uniform density δ_0 and such that any parallel to one axis, e.g., Z, intersects S only twice, at heights $z_2(\xi, \eta) > z_1(\xi, \eta)$.

Write the Newtonian integral T of B and prove that if $Q_2 \equiv (\xi, \eta, z_2)$, $Q_1 \equiv (\xi, \eta, z_1)$, then

$$\delta g_z \equiv -\frac{\partial T}{\partial z_P} = G\delta_0 \int_{B_0} d\xi d\eta \left[\frac{1}{r_{PQ_2}} - \frac{1}{r_{PQ_1}}\right],$$

where B_0 is projection of B onto the (x, y) plane.

(**Hint:** pass $-\frac{\partial}{\partial z_P}$ under the integral and notice that $-\frac{\partial}{\partial z_P}\frac{1}{r_{PQ}} = \frac{\partial}{\partial z_Q}\frac{1}{r_{PQ}}$).

Exercise 4. Consider a circular cylinder of uniform density δ_0 with base of radius b and height H_0. Assume the lower base is on the (x, y) plane.

Consider a point P on the axis of cylinder at height $H = H_0 + a, (x_P = y_P = 0)$.

By using the result of Exercise 3, prove that

$$\delta g_z = 2\pi G\delta_0[H_0 + \sqrt{b^2 + a^2} - \sqrt{b^2 + H^2}].$$

Noting that, according to our definition in Exercise 3, δg_z is the opposite of the z component of the attraction of the cylinder, prove also geometrically that it is always $\delta g_z > 0$.

(**Hint:** note that, calling ρ the polar coordinate in the (x, y) plane one has

$$\frac{1}{r_{PQ_2}} = \frac{1}{\sqrt{\rho^2 + a^2}}, \quad \frac{1}{r_{PQ_1}} = \frac{1}{\sqrt{\rho^2 + H^2}},$$

and B_0 of Exercise 3 is the circle $\rho \leq b$. Perform the integral in polar coordinates).

Exercise 5. Consider a homogeneous cone of density δ_0, height H_0 and circular basis on the (x, y) plane, with radius b centered at the origin.

The inclination I of this conical mountain is such that $H_0 = b \tan I$.

Take any point P on the axis of the cone (z axis) at height $z = H_0 + a$, $a > 0$ and compute, with the help of Exercise 3, the attraction δg_z of this cone at P.

Prove that

$$\delta g_z = 2\pi G\delta_0 \left[a\cos^2 I \sin I \log \frac{b + a\cos I \sin I + \cos I\sqrt{b^2 + (H_0 + a)^2}}{a\cos I(1 + \sin I)} + \right.$$
$$\left. - \sin^2 I\sqrt{b^2 + (H_0 + a)^2} + H_0 + a\sin^2 I\right].$$

In particular taking $a \to 0$, i.e., going to the point P on the top of the cone, one gets

$$\delta g_z = 2\pi G\delta_0 H_0(1 - \sin I)$$

Exercise 6. This and Exercises 7–9 constitute a guided tour to the computation of the potential T of a homogeneous parallelopiped. Consider a homogeneous parallelopiped D of density δ_0. Place the origin of the axes at the barycenter and assume that

$$D \equiv \{-a \le x \le a,\ -b \le y \le b,\ -c \le z \le c\}.$$

Call (x, y, z) the coordinates of the computation point P and (ξ, η, ζ) the coordinates of the running point in D.

Put

$$r(\xi, \eta, \zeta) = \sqrt{(\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2}$$

and

$$\begin{aligned} S_x &= \{-b \le y \le b,\ -c \le z \le b\} \\ S_y &= \{-a \le x \le a,\ -c \le z \le c\} \\ S_z &= \{-a \le x \le a,\ -b \le y \le b\}, \\ A_\pm &= a \pm x,\ B_\pm = b \pm y,\ C_\pm = c \pm z. \end{aligned}$$

By using the Exercises 1–3 prove that

$$2T(x, y, z) = G\delta_0 \left[\int_{S_x} \left(\frac{A_-}{r(a, \eta, \zeta)} + \frac{A_+}{r(-a, \eta, \zeta)} \right) d\eta d\zeta \right.$$
$$\left. + \int_{S_y} \left(\frac{B_-}{r(\xi, b, \zeta)} + \frac{B_+}{r(\xi, -b, \zeta)} \right) d\xi d\zeta + \int_{S_z} \left(\frac{C_-}{r(\xi, \eta, c)} + \frac{C_+}{r(\xi, \eta, -c)} \right) d\xi d\eta. \right.$$

Exercise 7. Put

$$\begin{aligned} \boldsymbol{\rho} &= (\eta - y)\mathbf{e}_y + (\zeta - z)\mathbf{e}_z,\ \rho = |\boldsymbol{\rho}| \\ \nabla_\rho &= \mathbf{e}_y \frac{\partial}{\partial \eta} + \mathbf{e}_z \frac{\partial}{\partial \zeta} \end{aligned}$$

and show that

$$\frac{1}{r(a, \eta, \zeta)} = \frac{1}{\sqrt{A_-^2 + \rho^2}} = \nabla_\rho \cdot \left[\frac{\sqrt{A_-^2 + \rho^2}}{\rho^2} \boldsymbol{\rho} \right].$$

Apply then the divergence theorem in two dimensions, to the rectangle S_x, proving that

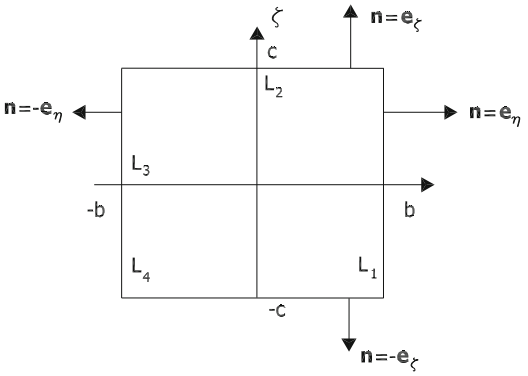


Fig. 1.19 The rectangle S_x in the plane (η, ζ) and its normal field

$$F = \int_{S_x} \frac{1}{r(a,\eta,\zeta)} d\eta d\zeta = B_- \int_{-c}^{c} \frac{\sqrt{A_-^2 + B_-^2 + (\zeta - z)^2}}{B_-^2 + (\zeta - z)^2} d\zeta$$

$$+B_+ \int_{-c}^{c} \frac{\sqrt{A_-^2 + B_+^2 + (\zeta - z)^2}}{B_+^2 + (\zeta - z)^2} d\zeta + C_- \int_{-b}^{b} \frac{\sqrt{A_-^2 + C_-^2 + (\eta - y)^2}}{C_-^2 + (\eta - y)^2} d\eta$$

$$+C_+ \int_{-b}^{b} \frac{\sqrt{A_-^2 + C_+^2 + (\eta - y)^2}}{C_+^2 + (\eta - y)^2} d\eta$$

(**Hint:** remember that in the plane (η, ζ)

$$\int_S \nabla \cdot \mathbf{v} d\eta d\zeta = \int_L \mathbf{v} \cdot \mathbf{n} d\ell$$

where L is the contour of S, covered in counterclockwise sense, $\mathbf{n}$ is the exterior normal to L, that in our case looks like the Fig. 1.19.

Note also that $d\ell$ is the line element which is always positive so that $d\ell = d\zeta$ on L_1, $d\ell = -d\eta$ on L_2, $d\ell = -d\zeta$ on L_3, $d\ell = d\eta$ on L_4).

Exercise 8. Show, by direct differentiation, that the following indefinite integral formula holds

$$\int \frac{\sqrt{A^2 + B^2 + t^2}}{B^2 + t^2} dt = \log(t + \sqrt{A^2 + B^2 + t^2})$$

$$-\frac{A}{B} \arctan \frac{B}{A} \frac{\sqrt{A^2 + B^2 + t^2}}{t}$$

Exercise 9. By combining Exercises 6–8 show that the full computation of the parallelopiped potential can be done through the following formulas

$$
\begin{aligned}
I(H,K,L) &= \int_0^L \frac{\sqrt{H^2+K^2+s^2}}{K^2+s^2}\,ds \\
&= \log\frac{L+\sqrt{H^2+K^2+L^2}}{\sqrt{H^2+K^2}} \\
&\quad -\frac{H}{K}\arctan\frac{K\sqrt{H^2+K^2+L^2}}{HL}+\frac{H}{K}\,\frac{\pi}{2} \\
&\quad \int_{-h}^{h}\frac{\sqrt{H^2+K^2+(s-t)^2}}{K^2+(s-t)^2}\,ds = I(H,K,h-t)-I(H,K,-h-t) \\
&= \log\frac{\sqrt{H^2+K^2+(h-t)^2}+(h-t)}{\sqrt{H^2+K^2+(h+t)^2}-(h+t)} \\
&\quad -\frac{H}{K}\arctan\frac{K\sqrt{H^2+K^2+(h-t)^2}}{H(h-t)}+ \\
&\quad -\frac{H}{K}\arctan\frac{K\sqrt{H^2+K^2+(h+t)^2}}{H(h+t)}
\end{aligned}
$$

$$
\begin{aligned}
&F(A_-,B_-,B_+,C_-,C_+) = B_-[I(A_-,B_-,C_-)-I(A_-,B_-,-C_+)] \\
&+B_+[I(A_-,B_+,C_-)-I(A_-,B_+.-C_+)] \\
&+C_-[I(A_-,C_-,B_-)-I(A_-,C_-,-B_+)] \\
&+C_+[I(A_-,C_+,B_-)-I(A_-,C_+,-B_+)] \\
&2T(x,y,z) = A_-F(A_-,B_-,B_+,C_-,C_+) \\
&+A_+F(A_+,B_-,B_+,C_-,C_+)+B_-F(B_-,A_-,A_+,C_-,C_+) \\
&+B_+F(B_+,A_-,A_+,C_-,C_+)+C_-F(C_-,A_-,A_+,B_-,B_+) \\
&+C_+F(C_+,A_-,A_+,B_-,B_+).
\end{aligned}
$$

Moreover, recognize that, put in this form, the formula requires the computation of 24 logarithms (because each of them appears always twice) and of 48 arctangents. We shall see at the end of Chap. 4 an equivalent formula reducing the computation to 12 logarithms and 24 arctangents.

Exercise 10. This exercise is intended as a preparation for the next one. Prove that

$$
\lim_{z\to 0_+}\frac{1}{4\pi}\frac{z}{\ell_{PQ}^3} \equiv \lim_{z\to 0_+}\frac{1}{4\pi}\frac{z}{[(x-\xi)^2+(y-\eta)^2+z^2]^{3/2}} = \frac{1}{2}\delta(x-\xi)\delta(y-\eta)
$$

(**Hint:** it is enough to prove that

$$
x\neq\xi,\ y\neq\eta \quad \frac{1}{4\pi}\frac{z}{\ell_{PQ}^3}\to 0
$$

and

$$\frac{1}{4\pi}\int_S \frac{z}{\ell_{PQ}^3} d\xi d\eta \equiv \frac{1}{2}, \quad \forall P,\ Z_P > 0)$$

Exercise 11. Apply the third Green's identity (1.61) to prove that, for a boundary S which coincides with the (x, y) plane, one has for a smooth function $u(x, y, z)$, harmonic in the upper half space $(z > 0)$

$$u(x, y, 0) = \frac{1}{2\pi}\int_S \left[-\frac{\partial u}{\partial z}(\xi, \eta, 0)\right] \frac{1}{\ell_{PQ}} d\xi d\eta$$
$$P \equiv (x, y, 0),\ Q \equiv (\xi, \eta, 0)$$

(**Hint:** write (1.61) for a point $P \equiv (x, y, z)$ inside the upper half space and take the limit for $z \to 0_+$ recalling the result of Exercise 10).

Exercise 12. This exercise is intended as a preparation for the next one. Prove that the following integral on the (x, y) plane vanishes

$$\frac{1}{4\pi}\int_S \frac{1}{\ell_{PQ}^3}\left[1 - 3\frac{z^2}{\ell_{PQ}^2}\right] d\xi d\eta = 0$$
$$\forall P \equiv (x, y, z), z > 0\ ;\ Q = (\xi, \eta, 0).$$

(**Hint:** use planar polar coordinates for (ξ, η)).

Exercise 13. Let $u(x, \eta, z)$ be a function harmonic in $\{z > 0\}$ continuous with first and second derivatives in $\{z \geq 0\}$. Prove that one has, on the (x, y) plane,

$$\frac{\partial u(x, y, 0)}{\partial z} = \frac{1}{2\pi}\int \frac{u(\xi, \zeta, 0) - u(x, y, 0)}{\ell_{PQ}^3} d\xi dy$$
$$P \equiv (x, y, 0),\ Q = (\xi, \eta, 0).$$

(**Hint:** write first $\dfrac{\partial u}{\partial z}$ using the third Green's identity for a point P with $z > 0$ and, using the result of Exercise 12, show that

$$\frac{\partial u(x, y, z)}{\partial z} = \frac{1}{4\pi}\int_S \left\{ \frac{[u(\xi, \eta, 0) - u(x, y, 0)]}{\ell_{P.Q}^3}\left[1 - 3\frac{z^2}{\ell_{PQ}^2}\right] + \frac{z}{\ell_{PQ}^3}\frac{\partial u}{\partial z}(\xi, \eta, 0)\right\} d\xi d\eta,\ P \equiv (x, y, z),\ Q = (\xi, \eta, 0).$$

Now take the limit for $z \to 0$, recalling Exercise 10 and observe that, according to the theory of singular integrals (Mikhlin 1957) one can take the limit under the integral of the first term to the right hand side: in fact due to the smoothness of u, one has

$$u(\xi, \eta, 0) - u(x, y, .0) = \frac{\partial u(x, y, 0)}{\partial x}(\xi - x) + \frac{\partial u}{\partial y}(x, y, 0)(y - \eta) + O(\ell_{PQ}^2)).$$

Appendix

A.1

We aim to prove that, as claimed in (1.49) – see also Werner (1974),

$$\Delta_P \frac{1}{r_{PQ}} = -4\pi\delta(P, Q) \tag{1.216}$$

where Dirac's δ is in fact a linear functional acting on a space of continuous functions, or on any subspace of smoother functions, according to the rule

$$\forall f, \quad < \delta_P, f >\equiv \int_{R^3} \delta(P, Q) f(Q) d_3x \equiv f(P). \tag{1.217}$$

Proof. To do that we use the definition of distributional derivative of a locally integrable vector field $\mathbf{v}$ (note that the field $-\frac{\mathbf{r}}{r^3}$ is indeed integrable over any finite ball in R^3 centered to the origin): we say that $\nabla \cdot \mathbf{v} = F$ if and only if for any test function $\varphi(\mathbf{x})$ (i.e., a function continuous with its derivatives of any order and which is identically zero outside a closed set K_φ) it is

$$F(\varphi) = (\nabla \cdot \mathbf{v})\varphi \equiv -\int \nabla\varphi \cdot \mathbf{v} d_3x \tag{1.218}$$

where d_3x is a volume element, and the integral is over the whole space or over K_φ, since outside this set $\nabla\varphi \equiv 0$.

So from (1.49) we compute, in spherical coordinates,

$$-\int \nabla\varphi \cdot \left(-\frac{\mathbf{r}}{r^3}\right) d_3x = \int \left(\frac{\mathbf{r}}{r} \cdot \nabla\varphi\right) \frac{1}{r^2} d_3x \tag{1.219}$$

$$= \int d\sigma \int_0^{+\infty} \frac{\partial\varphi}{\partial r} \frac{1}{r^2} r^2 dr = \int d\sigma[-\varphi(0)] = -4\pi\varphi(0)$$

If we consider the Dirac's δ distribution, namely the distribution defined by

$$\delta(\varphi) = \int \delta(P)\varphi(P)d_3x_P \equiv \varphi(0), \tag{1.220}$$

we see that putting $F = \delta$ and $\mathbf{v} = -\frac{\mathbf{r}}{r^3}$ in (1.218) and comparing (1.219) with (1.220) we can claim that

$$\Delta\frac{1}{r} = \nabla\cdot\left(-\frac{\mathbf{r}}{r^3}\right) = -4\pi\delta\ . \tag{1.221}$$

Note that $\delta(P)$ can be in a sense considered as the *density* of a point mass placed at the origin O.

So if we translate the origin to any point Q, we can write

$$\Delta_P\frac{1}{r_{PQ}} = -4\pi\delta(P,Q), \tag{1.222}$$

where by $\delta(P,Q)$ we mean the *density* implementing the identity

$$\int \delta(P,Q)\varphi(Q)d_3x_Q \equiv \varphi(P).$$

□

A.2

We wish to prove that a single layer potential as (1.53) satisfies the jump relations (1.54), i.e.,

$$\left(\frac{\partial V}{\partial n}\right)_+ - \left(\frac{\partial V}{\partial n}\right)_- = -4\pi G\alpha\ ; \tag{1.223}$$

on this you can see (Miranda 1970; Werner 1974) too. For this purpose we consider, beyond S, the boundary of B, another surface S', and its interior B', as shown in Fig. 1.20.

Let us compute the flux of $\mathbf{g} = \nabla V$ through such arbitrary S', which delimits by intersection an arbitrary subset S_0 of S (cf. Fig. 1.20). We have, with $\mathbf{n}'$ the outer normal of S',

$$\begin{aligned}\int_{S'}\mathbf{g}(P)\cdot\mathbf{n}'_P dS'_P &= G\int_S dS_Q\alpha(Q)\int_{S'}\left(-\frac{\mathbf{r}_{QP}}{\ell^3_{QP}}\cdot\mathbf{n}'_P\right)dS'_P\\ &= -4\pi G\int_S dS_Q\alpha(Q)\int_{B'}\delta(P,Q)dB'_P.\end{aligned} \tag{1.224}$$

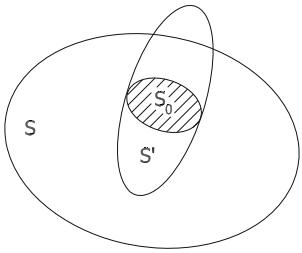


Fig. 1.20 Note that B is the interior of S, B' the interior of S' and $S_0 = S \cap B'$

We note that

$$\int_{S'} -\frac{\mathbf{r}_{QP}}{\ell^3_{QP}} \cdot \mathbf{n}'_P dS'_P = -4\pi \int_{B'} \delta(P,Q) dB'_P = \begin{cases} -4\pi & Q \in B' \\ 0 & Q \notin B', \end{cases} \quad (1.225)$$

i.e., it is, apart from the factor (-4π), the characteristic function of the set B'. Therefore, from (1.225), we find

$$\int_{S'} \mathbf{g}(P) \cdot \mathbf{n}'(P) dS_P = -4\pi G \int_{S_0} \alpha(Q) dS_Q, \quad (1.226)$$

because when Q is outside S_0 the integral (1.225) is zero. Since the identity (1.226) is valid for any S' defining the same S_0, by intersection of B' with S, we can choose S' as in Fig. 1.21. Since the right hand side (RHS) of (1.226) depends on S_0 but not on h, we can also take the limit for $h \to 0$. The integral on the lateral wall of the cylinder then disappears and we have only two integrals left, one on the upper face of S_0, another one on the lower face of S_0. Let us note that, when $h \to 0$, the normal to S_{0+} becomes the outer normal of S_0, so that

$$\int_{S_{0+}} \mathbf{g} \cdot \mathbf{n}' dS' = \int_{S_{0+}} \frac{\partial V}{\partial n'} dS' \to \int_{S_0} \left(\frac{\partial V}{\partial n}\right)_+ dS\,;$$

on S_{0-} however the normal $\mathbf{n}'$ is opposite to $\mathbf{n}$, so that

$$\int_{S_{0-}} \mathbf{g} \cdot \mathbf{n}' dS' = \int_{S_{0-}} \frac{\partial V}{\partial n'} dS' \to \int_{S_0} -\left(\frac{\partial V}{\partial n}\right)_- dS.$$

So, going back to (1.226), we receive

$$\int_{S_0} \left[\left(\frac{\partial V}{\partial n}\right)_+ - \left(\frac{\partial V}{\partial n}\right)_-\right] dS = -4\pi G \int_{S_0} \alpha dS$$

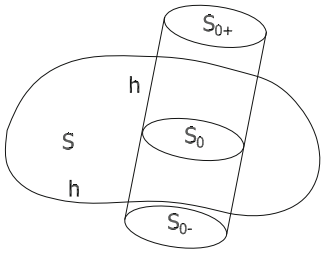


Fig. 1.21 Taking S' in the form of a cylinder orthogonal to S before letting $h \to 0$

and, since S_0 is arbitrary, we must have

$$\left[\left(\frac{\partial V}{\partial n}\right)_+ - \left(\frac{\partial V}{\partial n}\right)_-\right]_S = -4\pi G\alpha. \tag{1.227}$$

The relation (1.227) says that although the potential V is relatively regular across S, its normal derivative has a sharp jump equal to $-4\pi G\alpha$.

A.3

We aim to prove formulas (1.92) and (1.93) (Borisenko and Tarapov 1979).

Let $\mathbf{r}$ be the position vector of the point P to which we attach a system of coordinates $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^T$. We assume that $\{\xi_i\}$ is an orthogonal system, i.e., that vectors tangent to the coordinate lines are orthogonal to one another.

Such three vectors are easy to find as

$$\mathbf{h}_j = \partial_j \mathbf{r}(\boldsymbol{\xi}), \ \left(\partial_j = \frac{\partial}{\partial \xi_j}\right). \tag{1.228}$$

So we know a priori that

$$\mathbf{h}_i \cdot \mathbf{h}_j = h_i^2 \delta_{ij}, \ (h_i = |\mathbf{h}_i|). \tag{1.229}$$

Note immediately that the following fundamental relation holds

$$\partial_i \mathbf{h}_j = \partial_i \partial_j \mathbf{r}(\boldsymbol{\xi}) = \partial_j \partial_i \mathbf{r}(\boldsymbol{\xi}) = \partial_j \mathbf{h}_i. \tag{1.230}$$

Moreover the $\{\mathbf{h}_j\}$ are related to the metric, expressed in $\{\xi_i\}$ coordinates, through the two relations

$$d\mathbf{r} = \Sigma_i \mathbf{h}_j \, d\xi_j \tag{1.231}$$

$$|d\mathbf{r}|^2 = \Sigma_{i,j} \mathbf{h}_i \cdot \mathbf{h}_j \, d\xi_i \, d\xi_j = \Sigma_i h_i^2 d\xi_i^2. \tag{1.232}$$

Now take any smooth $F(\mathbf{r})$; by definition of ∇ we must have

$$dF = \nabla F \cdot d\mathbf{r} = \Sigma \nabla F \cdot \mathbf{h}_j d\xi_j \equiv \Sigma \partial_j F d\xi_j \tag{1.233}$$

so that

$$\partial_j F = \mathbf{h}_j \cdot \nabla F. \tag{1.234}$$

On the other hand, since the basis $\{\mathbf{h}_j\}$ is orthogonal, for any vector $\mathbf{v}$ we have

$$\Sigma_j \frac{\mathbf{h}_j}{h_j^2}(\mathbf{h}_j \cdot \mathbf{v}) \equiv \mathbf{v}. \tag{1.235}$$

By applying (1.235) to (1.234) one gets

$$\nabla F = \left(\Sigma_j \frac{\mathbf{h}_j}{h_j^2} \partial_j \right) F \; ; \tag{1.236}$$

since F in (1.236) is arbitrary, one can claim that

$$\nabla = \Sigma_j \frac{\mathbf{h}_j}{h_j^2} \partial_j . \tag{1.237}$$

Now we can pass to compute, with the help of (1.229) and (1.230),

$$\begin{aligned} \Delta &= \Sigma_{i,j} \frac{\mathbf{h}_i}{h_i^2} \partial_i \cdot \left[\frac{\mathbf{h}_j}{h_j^2} \partial_j \right] \\ &= \Sigma_{i,j} \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{h_i^2 h_j^2} \partial_{ij} + \Sigma_{i,j} \frac{\mathbf{h}_i \cdot \partial_i \mathbf{h}_j}{h_i^2 h_j^2} \partial_j + \\ &\quad - \Sigma_{i,j} \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{h_i^2} \frac{\partial_i (h_j^2)}{h_j^4} \partial_j \\ &= \Sigma_j \frac{1}{h_j^2} \partial_j^2 + \Sigma_{ij} \frac{\mathbf{h}_i \cdot \partial_j \mathbf{h}_i}{h_i^2 h_j^2} \partial_j + \\ &\quad - 2\Sigma_j \frac{(\partial_j h_j)}{h_j^3} \partial_j \end{aligned} \tag{1.238}$$

Let us consider together the second and third term in (1.238); for instance put $j = 1$, then the coefficient of ∂_1 is

$$
\begin{aligned}
j = 1 \quad & \Sigma_i \frac{\mathbf{h}_i \cdot \partial_1 \mathbf{h}_i}{h_i^2 h_1^2} - 2\frac{\partial_1 h_1}{h_1^3} \\
& = \Sigma_i \frac{\partial_1 h_i}{h_i h_1^2} - 2\frac{\partial_1 h_1}{h_1^3} \\
& = -\frac{\partial_1 h_1}{h_1^3} + \frac{\partial_1 h_2}{h_2 h_1^2} + \frac{\partial_1 h_3}{h_3 h_1^2} \\
& = \frac{1}{h_1 h_2 h_3} \partial_1 \left(\frac{h_2 h_3}{h_1} \right).
\end{aligned}
\tag{1.239}
$$

where we have used the obvious relation

$$
\mathbf{h}_i \cdot \partial_j \mathbf{h}_i = \frac{1}{2} \partial_j h_i^2 = h_i \partial_j h_i .
$$

By cycling the indexes and setting $H = h_1 h_2 h_3$, we see that

$$
\Sigma_i \frac{\mathbf{h}_i \cdot \partial_j \mathbf{h}_i}{h_i^2 h_j^2} - 2\frac{\partial_j h_j}{h_j^3} = \frac{1}{H} \partial_j \frac{H}{h_j^2}
\tag{1.240}
$$

so that, if we go back to (1.238), we can write

$$
\begin{aligned}
\Delta &= \frac{1}{H} \Sigma_j \frac{H}{h_j^2} \partial_j^2 + \frac{1}{H} \Sigma_j \left[\partial_j \frac{H}{h_j^2} \right] \partial_j \\
&= \frac{1}{H} \Sigma_j \partial_j \left[\frac{H}{h_j^2} \partial_j \right].
\end{aligned}
\tag{1.241}
$$

A.4

We want to prove formula (1.120), expressing the harmonic part of the normal potential. We refer to Sect. 1.9 for the notation. For a different approach to the determination of the normal potential, consult Heiskanen and Moritz (1967), Chap. 2.

First of all note that both the boundary surface $\mathcal{E}$ as well as the boundary condition (1.119), are cylindrically symmetric, so we expect that the sought solution $V_e = V_e(q, \overline{\vartheta})$ be independent of λ too. To determine a potential from Laplace equation and its values on the boundary, as in (1.119), is the Dirichlet problem. That such a problem has a unique solution depending with continuity from boundary data, is discussed at length in Chap. 13 of Part III.

Based on this consideration we can try to find our solution by a suitable guess and if we are able to prove that it works, then this is the sought one.

Given the shape of the Laplace operator and of the condition (1.119), we guess that a solution should have the form

$$V_e = A(q) - B(q)\sin^2\overline{\vartheta}. \tag{1.242}$$

We can immediately state that, if (1.242) is correct, owing to (1.119), A and B should satisfy for $q = b$ the relations

$$A(b) = V_0,\ B(b) = \frac{1}{2}\omega^2 a^2. \tag{1.243}$$

In addition, if we want V_e to be a regular potential, we must have

$$A(q) \to 0,\ B(q) \to 0 \text{ when } q \to \infty \tag{1.244}$$

Note that, since from (1.103) we have

$$r^2 = q^2 + E^2\sin^2\overline{\vartheta},$$

$q \to \infty$ is one and the same thing as $r \to \infty$.

Substituting the trial solution (1.242) into (1.110) and separating the two terms, one independent of $\overline{\vartheta}$, the other proportional to $\sin^2\overline{\vartheta}$, we get the differential system

$$\begin{cases} (q^2 + E^2)A'' + 2qA' - 4B = 0 \\ (q^2 + E^2)B'' + 2qB' - 6B = 0, \end{cases} \tag{1.245}$$

to be integrated with the boundary conditions (1.243) and (1.244).

To integrate (1.245) is a standard exercise, that we do for the sake of completeness.

We start with the second equation and first we look for a particular integral in the form

$$\overline{B} = pq^2 + c \qquad (p, c \text{ constants}). \tag{1.246}$$

We immediately find

$$\overline{B} = 3q^2 + E^2. \tag{1.247}$$

Then we put into (1.245)

$$B = \overline{B} \cdot v \tag{1.248}$$

and we see that the new unknown v has to satisfy the new differential equation

$$(q^2 + E^2)(3q^2 + E^2)v'' + [12q(q^2 + F^2) + 2q(3q^2 + E^2)]v' = 0.$$

We write this in the separated form

$$\frac{v''}{v'} = -\left[\frac{2q}{q^2+E^2} + \frac{12q}{3q^2+E^2}\right] \tag{1.249}$$

and integrate, obtaining

$$v' = \frac{C}{(q^2+E^2)(3q^2+E^2)^2} \equiv D\left[\frac{1}{q^2+E^2} - 3\frac{3q^2-E^2}{(3q^2+E^2)^2}\right]. \tag{1.250}$$

The reader can verify the second identity and discover that the new constant D is related to C by $C = 4E^4D$.

Before performing the last integration step we notice that

$$\int \frac{dq}{q^2+E^2} = \frac{1}{E}\arctan\frac{q}{E} \equiv \frac{1}{E}\left[\frac{\pi}{2} - \arctan\frac{E}{q}\right]$$

and that

$$-3\int \frac{3q^2-E^2}{(3q^2+E^2)^2}dq = \frac{3q}{3q^2+E^2}.$$

So the integral of (1.250) is

$$\begin{aligned} v &= -\frac{D}{E}\left[\arctan\frac{E}{q} - \frac{3qE}{3q^2+E^2}\right] + L + \frac{\pi}{2E} = \\ &\equiv G\left[\arctan\frac{E}{q} - \frac{3qE}{3q^2+E^2}\right] + H, \end{aligned} \tag{1.251}$$

with an obvious meaning of the constants.

Returning to B we get

$$B = G[(3q^2+E^2)\arctan\frac{E}{q} - 3qE] + H(3q^2+E^2). \tag{1.252}$$

Since we must have $B(q) \to 0$ when $q \to \infty$, we see that it has to be $H = 0$.

In fact, with the help of the Taylor formula

$$\arctan x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 + \ldots,$$

one verifies that the term in square parenthesis tends to zero. So, in order to satisfy the boundary relation

$$B(b) = \frac{1}{2}\omega^2 a^2 \tag{1.253}$$

it is enough to put

$$Q(q) = (3q^2 + E^2)\arctan\frac{E}{q} - 3qE \tag{1.254}$$

and

$$B(q) = \frac{1}{2}\omega^2 a^2 \frac{Q(q)}{Q(b)}. \tag{1.255}$$

The form of A is more immediate to find, since, by combining the two equations (1.245), one sees that

$$F = A - \frac{2}{3}B$$

has to satisfy

$$(q^2 + E^2)F'' + 2qF' = 0. \tag{1.256}$$

Considering that $F(q)$ has to tend to 0 for $q \to \infty$, because of (1.244) and (1.256) integrates in

$$F(q) = C\arctan\frac{E}{q},$$

i.e.,

$$A(q) = C\arctan\frac{E}{q} + \frac{2}{3}B(q). \tag{1.257}$$

By using the first of (1.243), C is determined and we get the final expression

$$V_e(q, \overline{\vartheta}) = \left(U_0 - \frac{1}{3}\omega^2 a^2\right)\frac{\arctan\dfrac{E}{q}}{\arctan\dfrac{E}{b}} + \frac{1}{2}\omega^2 a^2 \frac{Q(q)}{Q(b)}\left(\frac{2}{3} - \sin^2\overline{\vartheta}\right), \tag{1.258}$$

where $Q(q)$ is explicitly given by (1.254).

Chapter 2
Observables of Physical Geodesy and Their Analytical Representation

2.1 Outline of the Chapter

As we have shown in Chap. 1, the gravity potential W can be split into a known normal potential U plus the anomalous potential T; thus knowing T means knowing W.
Through the whole book we try to show how to relate T to quantities that can be observed on the earth surface or even in space, by using satellite technology. The first step in this direction is to study how to represent every geodetic observable quantity as a function of T.
Since T is our unknown, we have first to define what is the functional space to which it belongs; in this book we will use Hilbert spaces only, because they are much simpler than more general spaces and can be essentially treated as an infinite dimensional analogue of Euclidean spaces, with a very similar geometry.
A self-contained introduction to Hilbert spaces can be found in Part III, Chap. 12, although the reader that does not want to go deeper into mathematical technicalties, can in any way follow the text, only accepting here and there some statements without proof.

A numerical variable, function of T that is ranging in some Hilbert space H, is a functional of T. This functional can be linear or non-linear. Most of our actual observables in geodesy are non-linear functionals of $W = U + T$, and since T is much smaller than U, it is not surprising that we expect to be able to linearize the observation equations. The concept is made more precise through the definition of the Frechet and Gateaux differentials in Sect. 2.2.

Basically we can summarize the situation for the earth by saying that, as far as we want to determine a geoid with 1 cm accuracy, the linearized theory presented in the book is applicable.

Then we consider in Sect. 2.3 the most common observables related to T, we find their analytical form and we linearize them explicitly. In doing so it is convenient to consider combinations of variables, often geometric and physical quantities, mixed in a way which is very typical for geodesy.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_2,

By performing this job we find all the relevant functional relations expressing all the anomalous quantities described in Chap. 1, in terms of the anomalous potential T. In particular we find the relation between T and height anomalies, including the geoid undulations.

In the development of the section we encounter the case that the orthometric height is considered as an observable.

This is not a true geodetic observable in strict Molodensky's sense, because the definition of orthometric height implies the knowledge of the mass density between the earth surface and the geoid. Nevertheless it is not difficult to show that it is enough to know the mass density with a quite realistic approximation, or we can even fix it at a conventional mean value of $2.67\,\mathrm{g\,cm^{-3}}$, to be able to derive from true observables, namely the levelling increments, orthometric heights accurate to the centimeter level.

This is shown in the last part of Sect. 2.3 and the related estimates are fully developed and analyzed through Sect. 2.4. The key result of this section is formula (2.70) which can be interpreted (see Remark 3) according to the practical formula (2.75), making use of the so-called *Bouguer anomaly*.

Now that all the main relations between T and the observables have been established in a linearized form, in the effort of approximating T, we can use any further knowledge of factors that affect this potential and its functionals to reduce the unknown part of T, so to say we try to eat T morsel by morsel. This will be done in the following chapters, in terms of different wavelength components, but here in Sect. 2.5 the principle is established as the *remove–restore* concept.

Basically it implies that known gravitational effects can be subtracted from the free air gravity anomaly Δg and, once a solution has been found with the reduced data, we add back to it the piece of potential T due to the same known effects.

As such this principle is just another expression of the fact that now all the relations between the relevant quantities are linear. When we manipulate the relations found in Sect. 2.3 and apply them not to the full anomalous potential T, but to the residual unknown part of it corresponding to a maximum of a few meters of height anomaly, we are allowed to use one further approximation in our formulas, which is often useful, particularly in analytical studies. This consists in substituting a simple spherical potential instead of the normal potential U when this is present directly or through its functionals (e.g. through normal gravity γ).

This concept of spherical approximations is analyzed in Sect. 2.5. The procedure introduces a relative error somewhere between 10^{-2} and 10^{-3} and it is therefore justified only if T is reduced to a small component.

However it has to be stressed that nowadays this simplification has no particular reason to be applied when we work out numbers, since the exact expressions are as easy to be computed electronically; so its use has to be confined to simple qualitative and analytical considerations.

2.2 Observables and Observation Equations: Linearization

An observation is by definition an operation that, applied to a certain physical system, provides us with a real number. The operation, which is performed under conditions controlled as much as possible, is intended to provide the magnitude of a certain quantity q, in the sense that the number obtained, the observation q_0, is supposed to be

$$q_0 = q + \nu \tag{2.1}$$

with ν the observation error.

The number ν, which is obviously never known, however displays some peculiar behaviour: it is unstable, in the sense that if we repeat the observation measurement *under the same conditions*, we find another value q_0, i.e. by definition another value of ν, but its instability has a statistical character, in the sense that most of the times it does not change too much its absolute value.

Under these condition ν is modelled as a random variable, typically with

$$E\{\nu\} = 0,\ \sigma_\nu^2 = E(\nu^2) < +\infty. \tag{2.2}$$

We have already used the symbol E in a different context, as linear eccentricity. Here E is used for expectation over a probability distribution. The context should make clear the meaning of the symbol each time.

Here however we are interested in the quantity q we wanted to measure; in the case of physical geodesy this is generally a function of the position of the point (or points) involved in the measurement, through its coordinates $\mathbf{r}_P$, of the gravity field, e.g. through its potential W, and of a number of ancillary parameters, that we collect in a vector $\mathbf{x}$ of unknowns, e.g. parameters describing the transmission of e.m. waves through the air or parameters relative to the state of the measuring instruments etc.

So we can say that

$$q = F[\mathbf{r}_P, \mathbf{x}; W], \tag{2.3}$$

if the measurement is "pointwise", i.e. it refers to a specific point only; otherwise (2.3) contains more points $\{\mathbf{r}_{Pi}\}$.

What is a function of $\mathbf{r}_P$ and of $\mathbf{x}$ is common knowledge. We concentrate then on the meaning of being a function of W. This is a similar concept, with the difference that now W has to be chosen in some space having an infinite number of dimensions, because a general set of functions cannot be described with the help of a finite number of degrees of freedom.

Exactly in the same way as when we write $q = F(\mathbf{r}_P)$ we implicitly mean that $\mathbf{r}_P$ is ranging over R^3, or some subset of it, when we write $q = F[W]$, we have to specify too what is the set of elements on which W has to range.

In our case we specify this by assuming that

$$W = U + T \tag{2.4}$$

with T in some subset of the space of all functions harmonic in Ω, $\mathcal{H}(\Omega)$. Naturally although functions in such a space are very smooth in Ω (continuous with all their derivatives) they can display a very bad and rough behaviour at the boundary, so we have to select a subspace H of $\mathcal{H}(\Omega)$ in such a way that T is smooth enough as to guarantee that every functional F representing a physically feasible measurement be bounded. In other words if we put $T \in H$ into $F(U+T)$, for every F we need, we must be sure that we get a finite number, because otherwise we are trying to observe a quantity which is not measurable.

To give a precise functional formulation of this statement is not easy. However let us agree that we want *at least* $\mathbf{g} = \nabla W$ to be not too bad on the boundary S, i.e. that mean values of $\mathbf{g}$ on small patches of S be bounded. This can be most easily translated into the other requirement

$$\int_S |\mathbf{g}(Q)|^2 dS_Q < +\infty; \tag{2.5}$$

since

$$\mathbf{g} = \boldsymbol{\gamma} + \nabla T \tag{2.6}$$

and since $\boldsymbol{\gamma}$ is certainly a regular vector on S, we can convert (2.5) into

$$\int_S |\nabla T|^2 dS_Q < +\infty. \tag{2.7}$$

So a reasonable space H in which T has to be chosen could be defined through the requirement that in H a norm is defined according to

$$\|T\| = \left\{ \int_S |\nabla T|^2 dS \right\}^{1/2}. \tag{2.8}$$

For technical reasons, to be found in Part III, Chap. 13, instead of (2.8) an equivalent formulation is given by putting the not too restrictive constraint that S be star-shaped, i.e. it could be described by an equation of the form

$$r = R(\vartheta, \lambda). \tag{2.9}$$

In this case (2.8) is modified according to

$$\|T\|_1 = \left\{ \int_\sigma |\nabla T(R, \vartheta, \lambda)|^2 R^3(\vartheta, \lambda) d\sigma \right\}^{1/2} \tag{2.10}$$

with σ the unit sphere and $d\sigma$ its area element (cf. Part III, Sect. 14.2).

Remark 1. That (2.10) satisfies the definition of a norm (cf. Part III, Definition 8) can be verified as an exercise by the reader, with the help of the theory explained in Part III, Sect. 12.2. Only one point is more delicate, namely to prove that

$$\|T\|_1 = 0 \Rightarrow T = 0. \tag{2.11}$$

As a matter of fact if $\|T\|_1 = 0$ we have indeed $|\nabla T| = 0$ (almost everywhere) on S; therefore it has to be $T = C$ (constant) on S. On the other hand $|\nabla T| = 0$ implies that $\dfrac{\partial T}{\partial n} = 0$ on S too. So if we apply the Dirichlet identity (1.58) to the exterior space Ω, where T is harmonic, we find

$$\int_{\Omega} |\nabla T|^2 d\Omega = -\int_S T \frac{\partial T}{\partial n} dS = -C \int_S \frac{\partial T}{\partial n} dS = 0$$

so that it must be $T = C$ in the whole of Ω.

On the other hand T has to be regular at infinity, so that it has to be identically zero through Ω.

We note also that the norm (2.8), or the equivalent norm (2.10), can be related to the definition of a scalar product, i.e.

$$\|T\|_1^2 = \langle T, T \rangle_1$$

where

$$\langle T, V \rangle_1 = \int_\sigma (\nabla T \cdot \nabla V) R^3(\vartheta, \lambda) d\sigma. \tag{2.12}$$

Now that we have defined a norm and a Hilbert space structure in H we can also define what is the meaning of "linearizing" the functional F.

Let us remember (see Part III, Definition 11) that a continuous linear functional on H is a mapping, defined on whole H, $L : H \to \mathbf{R}$, such that

$$\forall \lambda, \mu \in \mathbf{R},\ \forall u, v \in H, \quad L(\lambda u + \mu v) = \lambda L(u) + \mu L(v).$$

Remember also that $y = o(\varepsilon)$ means that

$$\lim_{\varepsilon \to 0} \frac{y}{\varepsilon} = 0.$$

Then we say that $F(u), u \in H$, is differentiable at a "point" $\overline{u} \in H$ if there is a continuous linear functional $L(\,)$ such that, $\forall h \in H$,

$$F(\overline{u} + h) - F(\overline{u}) - L(h) = o(\|h\|); \tag{2.13}$$

in this case we say that $L(h)$ is the Frechet differential of F at $\overline{u}$ and we write (2.13) in the form

$$dF(\overline{u}, h) = L(h). \tag{2.14}$$

Note that since H is a Hilbert space, as a consequence of the famous Riesz theorem (cf. Part III, Theorem 2), every linear bounded functional on H can be represented in the form of a scalar product, namely $\exists \xi \in H$, such that

$$L(h) = \langle \xi, h \rangle_H . \tag{2.15}$$

The element ξ is called the Frechet derivative of F at $\overline{u}$ and usually denoted as $F'(\overline{u})$ or $F_u(\overline{u})$. So (2.14) can be written as

$$dF(\overline{u}, h) = \left\langle F'(\overline{u}), h \right\rangle_H . \tag{2.16}$$

In what follows we shall consider physical quantities that do depend on T through functionals that are everywhere differentiable in H.

In finite dimensional spaces, the definition (2.13) of the differential is pretty much the same, with the only difference that we have an Euclidean modulus of the increment, $|h|$, instead of the norm, $\|h\|$. However when we have to "compute" the derivative of a function $F(\mathbf{x})$ we use the more comfortable concept of gradient, $\nabla F(\overline{\mathbf{x}})$, such that

$$dF(\overline{\mathbf{x}}, \mathbf{h}) = \nabla F(\overline{\mathbf{x}}) \cdot \mathbf{h}. \tag{2.17}$$

In practice the gradient is computed by taking partial derivatives along all axes. In a similar way we define the gradient, or Gateaux derivative, of $F(u)$ at $\overline{u}$ by computing the limit

$$\begin{aligned} &\lim_{t \to 0} \frac{1}{t} \{F(\overline{u} + th) - F(\overline{u})\} \\ &= \frac{d}{dt} F(\overline{u} + th)|_{t=0} = L(h) = \langle \nabla F(\overline{u}), h \rangle_H . \end{aligned} \tag{2.18}$$

One can easily prove that if F is Frechet differentiable, then the Gateaux derivative $\nabla F(\overline{u})$ exists and

$$F'(\overline{u}) \equiv \nabla F(\overline{u}) \tag{2.19}$$

so that (2.18) becomes a comfortable tool to compute $F'(\overline{u})$.

The converse of the above statement is known to be false already in R^3, but we shall not be concerned with this problem, since we shall assume that all our $F(u)$ are regular enough for (2.19) to be true.

The *linearization* of a functional $F(u)$ around a "point" $\overline{u}$ in H is then just the use of the approximate expression

$$F(\overline{u}+h) \cong F(\overline{u}) + \left\langle F'(\mathbf{u}), h \right\rangle_H ; \tag{2.20}$$

thanks to (2.13) the error in (2.20) is $o(\|h\|)$. As a matter of fact, if we assume that F is two times Frechet differentiable we can even prove that the error is at least $O(\|h\|^2)$, i.e. in (2.20) we are neglecting only quadratic terms in h. This is particularly useful in physical geodesy, since to "linearize" our observation equations we shall systematically put

$$F(W) = F(U+T) \cong F(U) + dF(U,T). \tag{2.21}$$

In doing so, remembering that $O(T) \cong 10^{-5} O(U)$, $O(|\nabla T|) \cong 3 \cdot 10^{-5} O(|\boldsymbol{\gamma}|)$ etc, we obtain a relative precision in (2.21) better than 10^{-4}; for instance, in terms of the geoid, which is at most $\sim$100 m, we get relations accurate to better than 1 cm, which is within the target of this book.

The peculiar character of the observation equations of physical geodesy is that many times the observables are functions of T as well as of the observation point (or points) $\mathbf{r}_P$ which has totally or partially unknown geometric coordinates, as in (2.3). Therefore the general form of linearized geodetic equations is obtained as follows: let

$$\begin{cases} \mathbf{r}_P = \widetilde{\mathbf{r}}_P + \delta \mathbf{r}_P \\ \mathbf{x} = \widetilde{\mathbf{x}} + \boldsymbol{\xi}, \end{cases} \tag{2.22}$$

so that $\widetilde{\mathbf{r}}_P$ and $\widetilde{\mathbf{x}}$ are approximate values for $\mathbf{r}_P$ and $\mathbf{x}$, then we shall write systematically

$$\begin{aligned} q_0 = q + \nu \cong F(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U) \\ + F_{\mathbf{r}}(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U) \cdot \delta \mathbf{r}_P + F_{\mathbf{x}}(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U) \cdot \boldsymbol{\xi} + L(T) + \nu \end{aligned} \tag{2.23}$$

with $L(T)$ computed from (2.18), i.e.

$$L(T) = \lim_{t \to 0} \frac{1}{t} \{ F(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U + tT) - F(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U) \}. \tag{2.24}$$

We close the remark by noting that if we put

$$\widetilde{q} = F(\widetilde{\mathbf{r}}_P, \widetilde{\mathbf{x}}, U),$$

the known term of (2.24), namely $q_0 - \widetilde{q}$, is exactly the "geodetic anomaly" of q, Dq.

In the next section we shall see several examples of linearization for the combination of different observable quantities q and with different choices of the mapping $\mathbf{r}_P \leftrightarrow \widetilde{\mathbf{r}}_P$.

This general viewpoint to the linearization of observation equations, mixing geometric and physical quantities, was developed in geodesy in the 1970s and it has been designated as the Integrated Geodesy approach (see Moritz 1980; Krarup 2006, Chap. 18).

2.3 The Linearized Observation Equations of Physical Geodesy

In this section we shall consider suitable combinations of elementary observables.

Such elementary variables are:

(a) Geometric variables like $(\lambda, \varphi, h)_P$ or $(x, y, z)_P$ in a geocentric system. Quantites like these can nowadays be obtained point by point by GPS with an accuracy of a few centimeters worldwide, or areawise by radar interferometry. On oceans radar-altimeters give the coordinates of P again with a few centimeters of accuracy; on land one can obtain from satelliteborne radar missions the mean value of h over squares of side between 10 and 100 m, with a variable accuracy, say between 1 and 5 m.

In addition many times we can claim we know (λ, φ) of the point P, derived from classical geodetic techniques and photogrammetry; when photogrammetry and, more recently, laser scanning from an aerial platform are served by GPS and inertial systems, we are again able to derive (λ, φ, h) with an accuracy in the range of $\sim$5 cm;

(b) Physical variables like (Λ, Φ, g, W) which are typically obtained by astrogeodetic observations, (Λ, Φ), or by gravimetry, g, or by combining levelling with gravimetry, W. The astrogeodetic coordinates (Λ, Φ) can be obtained with an error of the order of 0,1 arcsec (corresponding to a shift of $\sim$3 m on the earth surface); g can be obtained with a very high accuracy, down to the 1 μGal level, though, as already explained, we hardly need that the measurement error be below the 0.1 mGal level for geodetic purposes; W can be obtained with an accuracy of some 0.1 $\mathrm{m^2\,s^{-2}}$ (or 10^3 Gal cm $= 10^{-2}$ g.p.u.).

As a matter of fact what is really observable is not directly W but a W difference between two points. It is for this reason that rigorously we should say that we observe $W(P) - \overline{W}$, where $\overline{W}$ is some reference unknown potential value related to the particular height system in which we operate.

Nevertheless we shall for the moment consider $W(P)$ as if it were directly observable and will introduce the proper changes into the next section.

More variables, like the gravity gradients, specially $\dfrac{\partial g}{\partial H}$, are observable by means of gradiometers, however we shall not dwell on that in this section as such

observations, on the earth surface S, constitute a pretty small data set available for the reconstruction of the gravity field;

(c) Finally, another quantity needs to be considered, which is both of geometric and physical character, namely the orthometric height H_P. This is really not directly observed, and in principle by going back to the original measurements it should always be possible to convert the observation of H into an observation of W; nevertheless there is a lot of information on H for which the original observations cannot be retrieved, so we deem it useful to include H as an elementary observable. Strictly connected with H is a real native elementary measurement, namely the leveling increment. This will be treated at the end of the section.

Remark 2. Note that many times instead of observing directly a quantity $q(P)$, at P, we rather observe increments, i.e. $q(P) - q(Q)$, between two point P and Q. Yet for the matter of linearization it will be clear how to make the generalization from the observation equation of $q(P)$ to that of the increment $q(P) - q(Q)$.

As explained at the end of Sect. 2.2, in order to perform a correct linearization we need simultaneously to establish a map $\mathbf{r}_P \leftrightarrow \widetilde{\mathbf{r}}_P$, involving three subsidiary relations; it is for that reason that we shall work out observation equations for quadruples of geodetic observables or more:

1. (λ, φ, h, W); this is the simplest case because we know the coordinates of P and then we can put straightforwardly

$$\widetilde{\mathbf{r}}_P = \mathbf{r}_P; \tag{2.25}$$

so we have

$$W(P) = U(P) + T(P). \tag{2.26}$$

In this case $L(T) \equiv T(P)$ i.e. the functional L is just the evaluation of T at P (see Part III, Definition 21),

2. (λ, φ, h, g): here again we know the coordinates of P and we can use the identity mapping (2.25). Yet the observation of $g(P)$ is not any more a linear functional because

$$g(P) = |\nabla W(P)| = |\boldsymbol{\gamma}(P) + \nabla T(P)|; \tag{2.27}$$

since $|\nabla T| < 10^{-4}|\boldsymbol{\gamma}|$ we can linearize (2.27).
Recalling (2.18) we find

$$\begin{aligned} \frac{d}{dt}\,|\boldsymbol{\gamma}(P) + t\nabla T(P)||_{t=0} &= \left.\frac{[\boldsymbol{\gamma}(P) + t\nabla T(P)] \cdot \nabla T(P)}{|\gamma(P) + t\nabla T(P)|}\right|_{t=0} \\ &= \frac{\boldsymbol{\gamma}(P)}{\gamma(P)} \cdot \nabla T(P). \end{aligned} \tag{2.28}$$

So, using the definition of gravity disturbance (1.170), the observation equation of $g(P)$ in this case is

$$\delta g(P) = g(P) - \gamma(P) = \frac{\boldsymbol{\gamma}(P)}{\gamma(P)} \cdot \nabla T(P); \tag{2.29}$$

if the point P is on the earth surface, or nearby, (2.29) can be safely approximated by

$$\delta g(P) \cong -\boldsymbol{\nu} \cdot \nabla T(P) = -\frac{\partial T}{\partial h}(P). \tag{2.30}$$

Note that the linear functional $L(\,)$ in (2.30) is a combination of the operator $\dfrac{\partial}{\partial h}$ with the functional of evaluation at P,

3. (λ, φ, W, g); this is probably the most important combination of observables at least in the classical sense discussed by Molodensky, Yurkina, Eremeiev (cf. Molodensky et al. 1962) and developed by many authors in physical geodesy (see for instance Heiskanen and Moritz 1967; Moritz 1980; Krarup 2006).
We note that in this case the knowledge of the coordinates of P is incomplete; however recalling the definition of normal height h_P^* and of height anomaly ζ (cf. (1.165), (1.168)) we have

$$\mathbf{r}_P \equiv (\lambda, \varphi, h) \Leftrightarrow \widetilde{\mathbf{r}}_P = (\lambda, \varphi, h^*) \tag{2.31}$$

where

$$W(P) = W(\lambda, \varphi, h) = U(P^*) = U(\lambda, \varphi, h^*).$$

Since

$$h = h^* + \zeta \tag{2.32}$$

we can write

$$\begin{aligned} 0 &= W(\lambda, \varphi, h^* + \zeta) - U(\lambda, \varphi, h^*) = \\ &\cong W(\lambda, \varphi, h^*) - U(\lambda, \varphi, h^*) + \frac{\partial W}{\partial h}(\lambda, \varphi, h^*)\zeta. \end{aligned} \tag{2.33}$$

But the last term can be written

$$\frac{\partial W}{\partial h}\zeta = \frac{\partial U}{\partial h}\zeta + \frac{\partial T}{\partial h}\zeta \cong -\gamma(\lambda, \varphi, h^*)\zeta + \frac{\partial T}{\partial h}\zeta; \tag{2.34}$$

since $0\left(\frac{\partial T}{\partial h}\zeta\right) \cong \cdot 10^{-4}\gamma\zeta$ (see (1.171)) we can neglect the second term in the right hand side of (2.34) and write

$$\frac{\partial W}{\partial h}\zeta \cong -\gamma\zeta,$$

which, used in (2.33), gives us

$$\gamma(\lambda, \varphi, h^*)\zeta(P^*) = T(\lambda, \varphi, h^*) \tag{2.35}$$

or

$$\zeta(P^*) = \frac{T(P^*)}{\gamma(P^*)}; \tag{2.36}$$

(2.34), or (2.35), is known as Bruns's relation (Heiskanen and Moritz, 1967). Note that (2.36) holds true with a relative error of better than 10^{-4} and since $O(|\zeta|) = 100\,\text{m}$, this means an error smaller than 1 cm in ζ.
Moreover when the linearization point $\widetilde{P}$ is directly on the ellipsoid, as it happens when P is on the sea surface, we have indeed $\zeta(\widetilde{P}) = N(\widetilde{P})$.
We have now to couple (2.36) with the observation equation of $g(P)$, which in this case writes

$$\begin{aligned} g(\lambda, \varphi, h) = g(\lambda, \varphi, h^* + \zeta) &\cong g(\lambda, \varphi, h^*) + \frac{\partial g}{\partial h}(\lambda, \varphi, h^*)\zeta \\ &\cong \gamma(\lambda, \varphi, h^*) + \delta g(\lambda, \varphi, h^*) + \frac{\partial \gamma(\lambda, \varphi, h^*)}{\partial h}\zeta \end{aligned}$$

Re-arranging and recalling (2.29) and (2.36) we find

$$\begin{aligned} \Delta g(\lambda, \varphi, h^*) &= g(\lambda, \varphi, h) - \gamma(\lambda, \varphi, h^*) \\ &= \frac{\boldsymbol{\gamma}}{\overline{\gamma}} \cdot \nabla T + \frac{\gamma'}{\gamma} T, \end{aligned} \tag{2.37}$$

where we have denoted $f' \equiv \frac{\partial f}{\partial h}$ for the sake of brevity; finally with approximation (2.30) one gets

$$\Delta g(\widetilde{P}) = -T' + \frac{\gamma'}{\gamma} T \tag{2.38}$$

which is also known as the *fundamental equation of physical geodesy*.

We note, for future reference, that (2.38) can be cast into the nice form

$$\frac{\Delta g}{\gamma} = -\frac{\partial}{\partial h}\left(\frac{T}{\gamma}\right), \tag{2.39}$$

Telluroid. The function $\zeta(\lambda, \varphi)$, as we know, is called the *height anomaly*. By mapping $\zeta(\lambda, \varphi)$ above the ellipsoid we get a surface not too far from the geoid, but not coinciding with it, sometimes called the co-geoid. We strongly underline that knowing ζ on the earth surface is one and the same thing as knowing the anomalous potential on it because of Bruns's relation (2.36) and, subsequently, T everywhere outside the masses, as a solution of the Dirichlet problem.

Furthermore through the mapping (2.31), i.e. by moving the point P of a quantity $-\zeta$ along the ellipsoidal normal, we generate another surface, an image of the earth surface, which is called the **telluroid**,

4. (Λ, Φ, g, W): we mention this combination not to further elaborate it analytically but only for historical reasons, because this has been the first problem considered by Molodensky et al. 1962. On the other hand the knowledge of (Λ, Φ) is available for such a little number points on the earth surface that we don't need to dwell on it. By the way, when an observation of (Λ, Φ) by a Zenith camera is done nowadays, it is very easy also to get the position of P by GPS, a case which is treated in the next point,

5. $(\lambda, \varphi, h, \Lambda, \Phi)$: here we could consider Φ and Λ separately or together, which is equivalent to saying that we observe $\mathbf{n}$ at a point P of known coordinates. But then we can directly compute the vector of the deflection of the vertical

$$\boldsymbol{\varepsilon} = \mathbf{n}(P) - \boldsymbol{\nu}(P)$$

and recalling (1.190), with $\delta\mathbf{g} = \nabla W(P) - \nabla U(P) = \nabla T(P)$, we find the observation equation

$$\boldsymbol{\varepsilon}(P) = -\frac{1}{\gamma}(I - P_\nu)\nabla T. \tag{2.40}$$

We note that $(I - P_\nu)\nabla T$ is just the horizontal gradient of T, i.e. the component of ∇T orthogonal to $\boldsymbol{\nu}$.
By decomposing (2.40) into the northwise and eastwise directions, recalling (1.187) and (1.188), we find

$$\begin{vmatrix} \Phi - \varphi \\ \cos\varphi(\Lambda - \lambda) \end{vmatrix} = \begin{vmatrix} \xi \\ \eta \end{vmatrix} = -\frac{1}{\gamma}\begin{vmatrix} \mathbf{e}_\varphi \cdot \nabla T \\ \mathbf{e}_\lambda \cdot \nabla T \end{vmatrix}, \tag{2.41}$$

which is the sought observation equation,

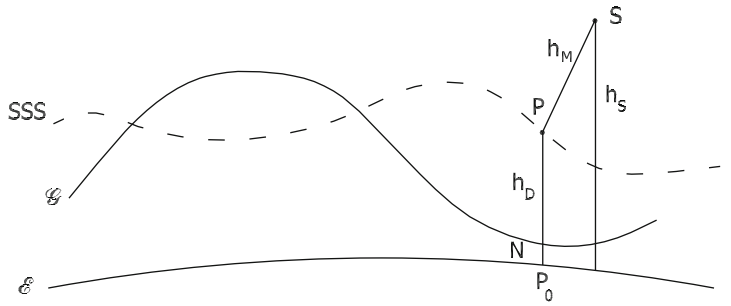


Fig. 2.1 $\mathcal{E}$ ellipsoid, $\mathcal{G}$ geoid, SSS Stationary Sea Surface, h_S height of the satellite S, h_M height measured by radar altimeter, h_D dynamic height of the sea due to geographic currents, N geoid undulation

6. (λ, φ, h, H): there are two different contexts in which such a combination matters: on oceans or on land.

   (a) *On ocean*: in this case (λ, φ, h) is known from satellite radar-altimetry; note that, accordingly, h, averaged in time from many different tracks of the satellite over the same area, has the meaning of mean sea level, which we assume to be stationary in time.
   Furthermore the orthometric height of P, i.e. the height of the Stationary Sea Surface (SSS) over the geoid ($\mathcal{G}$), is due to the presence of stationary oceanic currents and water density variations, and it can be modelled and predicted.
   The situation is illustrated in Fig. 2.1.
   As we can see the following relations hold

$$h_P = h_S - h_M,$$

   where h_S is known from satellite tracking and h_M is the radar altimeter measurement,

$$h_P = h_D + N$$

   where $h_D = H_P$ is the dynamic height predicted by oceanographic models. Therefore, summarizing, we get the observation equation (cf. (2.36))

$$\frac{T(P_0)}{\gamma(P_0)} = N = h_S - h_M - h_D = h_P - H_P \tag{2.42}$$

   which is indeed linear and is referred to the point P_0 on the ellipsoid, while the actual surface SSS is at a distance of the order of N from $\mathcal{E}$, because h_D has a maximum magnitude below 3 m.
   So we can say that we have (2.42) at the actual surface of the earth.

(b) *On land*: in this case we assume for instance that at the same point P we have GPS observations, providing (λ, φ, h), and a benchmark of a levelling line, where the orthometric height H_P has been computed and we don't have anymore the original information that would allow the direct computation of W_P.
Indeed we can always say that from such observations we can compute

$$N = \frac{T(P_0)}{\gamma(P_0)} = h_P - H_P \tag{2.43}$$

but we don't like to use directly (2.43) as an observation equation because now the point P_0, which is on the ellipsoid, can be kilometers within the masses and far away from P. This heavily contradicts the basic Molodensky principle that all physical geodesy could be done with quantities referred to the surface only. Obviously, in one way or another the masses between earth surface and geoid will enter into the observation equation, because they are fundamental for the definition of orthometric height and they have certainly been used when orthometric height have been computed (see the point 8 in this section). So we shall do it indirectly showing how, with some supplementary information, we can derive from our data the value of $W(P)$, controlling that only a coarse information on the masses is needed, because our formulas are little sensitive to it. Of course in this case we shall be happy to arrive at a result approximated to a few centimeters in terms of the height anomaly.
To this aim we first write

$$\begin{aligned} W(P) = U(h^*) = U(H + h^* - H) = \\ \cong U(H) - \gamma(H)(h^* - H); \end{aligned} \tag{2.44}$$

which is nothing but the definition of normal height h^*, suitably linearized around H, a known quantity. Note that in (2.44) the dependence of functions from (λ, φ) has been skipped, because it plays no role. Note also that both $U(H), \gamma(H)$ are known as they are computed using H in analytical expressions. Furthermore, since

$$h = h^* + \zeta = H + N,$$

we have

$$h^* - H = N - \zeta. \tag{2.45}$$

In the next section it will be shown that, with some approximations, one has

$$N_{P_0} - \zeta_P \cong \frac{\Delta g_P}{\gamma_0} H_P - \frac{2\pi G\rho}{\gamma_0} H_P^2 \tag{2.46}$$

where Δg_P is the free air anomaly at P, γ_0, is a constant standard value, e.g. $\gamma_0 = 981\,\text{Gal}$, which is the mean normal gravity value. The standard density ρ of the crust is $\rho = 2.67\,\text{g}\,\text{cm}^{-3}$ and a variation of density of $\pm 10\%$ can be considered as very large. Since for $H = 10^3\,\text{m}$

$$O\left(\frac{2\pi G\rho}{\gamma_0}H^2\right) \sim 0.1\,\text{m} \tag{2.47}$$

we see that a 10% variation of ρ on a thickness of 1,000 m gives to $N - \zeta$ a variation of 1 cm, which is within our accuracy target. Similarly in the first addendum of (2.46) we find a term depending on the free air anomaly Δg_P, which is not supposed to be known by measurements. So it has to be derived from a free air anomaly map with an error that can easily amount to $\delta\Delta g_P = 10\,\text{mGal}$. Nevertheless such an error, when we put $H = 1{,}000\,\text{m}$, has an effect on $N - \zeta$ of the order of

$$O\left(\frac{\delta\Delta g}{\gamma_0}H\right) \cong 10^{-5}10^3\,\text{m} = 1\,\text{cm},$$

which is again within our target. Concluding we could say that by combining (2.44) and (2.46) we can transform our data into a value of $W(P)$, with an error up to a few centimeters in height anomaly.

7. (λ, φ, g, H): this is a very traditional but rather mixed set of observables. Similarly to the discussion in point 6, because of the presence of H, we need to make some further approximation in order to arrive at an observation equation at the surface level. To this aim we start from the observation equation of the free air anomaly (cf. (2.37))

$$\Delta g = -T' + \frac{\gamma'}{\gamma}T \tag{2.48}$$

which however we cannot use directly because to compute Δg we need h_P^*, i.e. $W(P)$, as by definition

$$\Delta g = g(P) - \gamma(h_P^*). \tag{2.49}$$

Nevertheless, we can write (2.49) as

$$\begin{aligned}\Delta g &= g(P) - \gamma(H_P) - \gamma'(H_P)(h_P^* - H_P) \\ &= g(P) - \gamma(H_P) - \gamma'(H_P)(N - \zeta).\end{aligned} \tag{2.50}$$

If we define a different gravity anomaly, namely

$$\Delta\widetilde{g} = g(P) - \gamma(H_P) \tag{2.51}$$

we rewrite (2.50) as

$$\Delta g = \Delta\widetilde{g} - \gamma'(H_P)(N - \zeta), \tag{2.52}$$

where $\Delta\widetilde{g}$ is known from our basic information.
Now consider that $O(\gamma') \sim 0.3\,\text{mGal}\,\text{m}^{-1}$ and $(N - \zeta)$ is known to be less than 2 m, so that

$$O(\gamma'(N - \zeta)) \leq 1\,\text{mGal}.$$

It follows that if in (2.46) we use $\Delta\widetilde{g}$ instead of Δg, i.e. we write

$$N - \zeta \cong \frac{\Delta\widetilde{g}}{\gamma_0} H_P - \frac{2\pi G\rho}{\gamma_0} H_P^2 \tag{2.53}$$

where all terms are known or computable, we find $N - \zeta$ with error smaller then $10^{-6} H_P$, i.e. less then 1 mm/km. Since this is irrelevant, we can use (2.53) to compute $N - \zeta$ and (2.52) to compute Δg. Finally, with Δg on the surface, we can use (2.48) as an observation equation;

8. $(\lambda, \varphi, \Delta g, \delta L)$: in this case (λ, φ) need to be known with a rough approximation for which cartographic coordinates could be enough. Similarly Δg is assumed to be known from a gravity map, say with an error up to 10 mGal. On the contrary δL, the levelling increment, is the true precise measurement. Note that the individual observation is a step δL observed along a levelling line, winding from an initial point A to a final point B; the levelling increments are then added along the levelling line from A to B.
Since a typical horizontal length of a single step is 100 m while a typical length of the line joining A to B is some kilometers, we shall collect our measurements in the form

$$\Delta_{AB} L = \int_{\widehat{AB}} \delta L. \tag{2.54}$$

Let us first examine closely the individual term δL: it is (cf. (1.86))

$$\begin{aligned} \delta L &= \mathbf{n}_P \cdot d\mathbf{r}_P \\ &= (\mathbf{n}_P - \boldsymbol{\nu}_P) \cdot d\mathbf{r}_P + \boldsymbol{\nu}_P \cdot d\mathbf{r}_P \\ &= \boldsymbol{\varepsilon} \cdot d\mathbf{r}_P + dh, \end{aligned} \tag{2.55}$$

where (cf. (2.40))

$$\boldsymbol{\varepsilon} = -\frac{1}{\gamma} \nabla_h T. \tag{2.56}$$

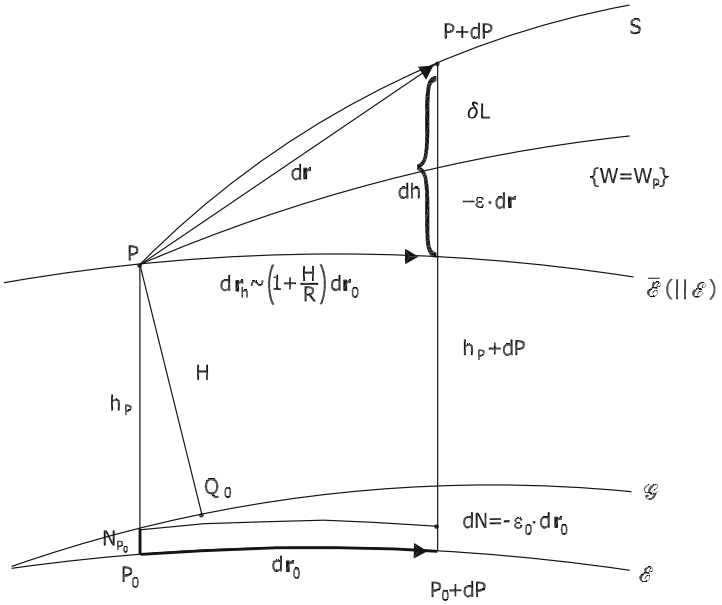


Fig. 2.2 The geometrical setting of the spirit levelling: S earth surface, $W = W_P$ equipotential through P, $\overline{\mathcal{E}}$ parallel to $\mathcal{E}$ through P, $\mathcal{G}$ geoid, $\mathcal{E}$ ellipsoid, $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_0$ deflections of the vertical, respectively in P (on the surface) and in P_0 (on the ellipsoid)

In (2.56) we have denoted with ∇_h the horizontal gradient, $(I - P_v)\nabla$.
The situation is represented in Fig. 2.2, where the relation (2.55) could be derived by a simple geometrical reasoning.

By integrating (2.55) along the line $\overset{\frown}{AB}$ on the surface S we can write

$$\Delta_{AB} L = h_B - h_A + \int_{\overset{\frown}{AB}} \boldsymbol{\varepsilon} \cdot d\mathbf{r}_P \tag{2.57}$$

We note that (2.57) is the *observation equation* of $\Delta_{AB} L$ although it has never been used, in this form, in geodetic literature. The reason is that in (2.57) the linear functional $F(\,)$ of T, i.e.

$$F(T) = \int_{\overset{\frown}{AB}} -\frac{1}{\gamma} \nabla_h T \cdot d\mathbf{r}_P ,$$

is not pointwise but it does depend on the line $\overset{\frown}{AB}$. Although later on we shall learn how to deal with that, we account here for an approximation procedure that transforms (2.57) into an observation equation for the increment of the orthometric height.

The result is summarized into the formula

$$\Delta_{AB}L = (N_B - \zeta_B) - (N_A - \zeta_A) - \int_{\widehat{AB}} \frac{g - \gamma_0}{\gamma_0}\delta L + H_B - H_A, \quad (2.58)$$

with γ_0 some mean constant gravity value, for instance $\gamma_0 = 981$ Gal. The proof has to be found in Sect. A.1.

In geodetic literature the term

$$OC = (N_A - \zeta_A) - (N_B - \zeta_B) + \int_{\widehat{AB}} \frac{g - \gamma_0}{\gamma_0}\delta L \quad (2.59)$$

is called *orthometric correction* (Heiskanen and Moritz 1967).

We note that the third term in (2.59) can be computed by a rough knowledge of $\Delta g \sim \dfrac{g - \gamma_0}{\gamma_0}$, as we can have from a free air anomaly map. As for the other terms, going back to (2.46) we can write

$$(N_A - \zeta_A) - (N_B - \zeta_B) \cong \frac{\Delta g_A + \Delta g_B}{2\gamma_0}(H_A - H_B) \quad (2.60)$$

$$+\frac{\Delta g_A - \Delta g_B}{\gamma_0}\left(\frac{H_A + H_B}{2}\right) - \frac{2\pi G\rho}{\gamma_0}(H_A - H_B)2\left(\frac{H_A + H_B}{2}\right).$$

In such an expression we are allowed to substitute $-\Delta_{AB}L \cong (H_A - H_B)$ and to use a very roughly approximated value for $\dfrac{H_A + H_B}{2}$ (e.g. with an error of 10 or 20 m), to get a result better than our usual range of accuracy, in fact accurate to a few millimeters. Therefore OC in (2.59) can be considered as a "correction", known up to millimeters, and (2.58) can be finally written as

$$H_B - H_A = \Delta_{AB}L + OC \quad (2.61)$$

as an observation equation for the orthometric height only.

Summarizing this long presentation we could say that, where the modern positioning techniques allow us to know the coordinates of the observation point, we can easily write observation equations of δg and $\boldsymbol{\varepsilon}$; when gravity and levelling data have been correctly used (i.e. geopotential numbers and normal heights computed) we have observation equations for the free air gravity anomaly Δg. When orthometric heights are used as data we are forced to use suitable approximate formulas for $N - \zeta$, that will be justified in the next section, to reconduct our observation equations to the previous form and in any way referring now to surface quantities.

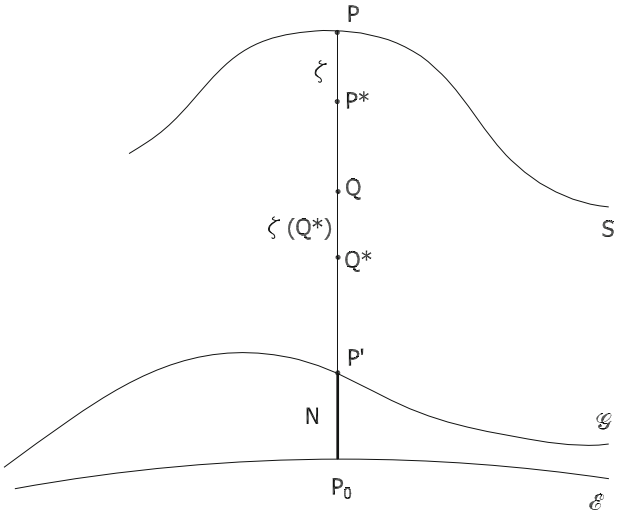


Fig. 2.3 Integration path for (2.62)

2.4 On the Relation Between Height Anomalies and Geoid Undulations

As we have seen in the previous section, there are measurements and relations which oblige the geodesist to "enter" into the first layers of the masses. In particular it is interesting to know $N_{P_0} - \zeta_P$, where P is on S and P_0 its orthogonal projection on $\mathcal{E}$, for all the reasons explained in Sect. 2.3, points 7 and 8. We follow here the approach presented in Sansò and Vaníček (2006) which is a refinement of the more classical reasoning of Heiskanen and Moritz (1967).

In addition to that we underline that when we have solved the main problem of physical geodesy, namely we have built a mathematical model of ζ_P, we can "test" this model whenever we have a point P at which we know both H and h (and then $N = h - H$ too).

This point is important and sometimes misunderstood in geodetic literature, therefore we shall come back to that later on.

The first step to solve our problem is devoid of further approximations, rather than those implied by the (rigorous) linearization procedures. In fact we know that (cf. (2.36), (2.39))

$$-\frac{\partial}{\partial h}\left(\frac{T}{\gamma}\right) = -\frac{\partial}{\partial h}\zeta = \frac{\Delta g}{\gamma}; \tag{2.62}$$

a careful inspection of the way in which (2.62) has been derived shows that it holds at any point Q^* such that $h(Q^*) = h^*_Q$, where Q runs along the ellipsoidal normal from P' up to P, while Q^* runs along the same normal from P_0 to P^* (see Fig. 2.3).

Note that it is mandatory to specify what is the independent variable in (2.62) because $\Delta g = g(Q) - \gamma(Q^*), \zeta = h_Q - h_{Q^*}$ are as a matter of fact functions of two points and they can be correctly attributed to one or the other.

So if we integrate (2.62) in h^* from P_0 to P^* we get (recall that $\zeta_{P_0} = N_{P_0}$)

$$-(\zeta_{P^*} - N_{P_0}) = N_{P_0} - \zeta_{P^*} = \int_0^{h_P^*} \frac{\Delta g}{\gamma} dh^*. \tag{2.63}$$

This relation shows that we have to continue Δg into the masses in order to derive $N - \zeta$. For this purpose we have to use the relation (1.81), already derived in Chap. 1, with the warning that in that context we have denoted by ℓ a curvilinear coordinate along the plumbline, which now we know to be the orthometric height H. So we can write

$$\frac{\partial g}{\partial H} = -2Cg + 4\pi G\rho - 2\omega^2; \tag{2.64}$$

where C is the mean curvature of the equipotential surface.

Let us note that basically (2.64) is nothing but Poisson's equation for W, written in local curvilinear coordinates, adapted to the geometry of the gravity field. So in $\frac{\partial g}{\partial H}$ we recognize the second derivative in vertical direction of W, and it is not difficult to see that $2Cg$ represents the "horizontal" Laplacian of W, i.e. the Laplace-Beltrami, Δ_t, operator for the equipotential surface, applied to W.

As such we shall never be able to know exactly C (as well as ρ) in (2.64) without having solved before the problem of determining W. Yet we will show that one can play the game of *sensitivity of the result and suitable approximations* for C and ρ, so that we are able to derive an equation for $\frac{\partial}{\partial h^*}\Delta g$, controlling the error at the centimetric level, which is our target.

This painful work is performed in Sect. A.2, where we arrive at the equation

$$\frac{\partial}{\partial h^*}\Delta g = -2C_0\Delta g + 4\pi G\rho, \tag{2.65}$$

where C_0 is the mean curvature of the ellipsoid at P_0 and ρ is also fixed to a constant value, e.g. $\rho = 2.67\,\mathrm{g\,cm^{-3}}$.

With that in mind we can integrate (2.65) to get

$$\Delta g(h^*) = \Delta g_P e^{2C_0(h_P^* - h^*)} + \frac{4\pi G\rho}{2C_0}\left[1 - e^{2C_0(h_P^* - h^*)}\right]. \tag{2.66}$$

Note that the solution (2.66) satisfies the initial condition

$$\Delta g(h_P^*) = \Delta g_P,$$

a quantity that we assume to be given on the surface S.

Since for every P on the earth surface $O(C_0[h_P^* - h^*]) \sim 10^{-3}$, we can safely linearize the exponentials in (2.66), obtaining

$$\Delta g(h^*) = \Delta g_P[1 + 2C_0(h_P^* - h^*)] - 4\pi G\rho(h_P^* - h^*). \tag{2.67}$$

The formula (2.67) provides the continuation of Δg_P into the masses, down to the ellipsoid. The most relevant error in (2.67) depends on the imperfect knowledge of ρ, and it can amount to several milligals.

Finally, we can use (2.67) into (2.63); it is not difficult to verify that

$$\frac{\Delta g(h^*)}{\gamma(h^*)} \sim \frac{\Delta g(h^*)}{\gamma_0} \tag{2.68}$$

with γ_0 some constant value at the ellispoid. In fact, neglecting the dependence of $\gamma(h^*)$ from h^* gives rise to errors absolutely irrelevant with the present criteria. So (2.63) is easily integrated to

$$N_{P_0} - \zeta_P = \frac{\Delta g_P}{\gamma_0}[h_P^* + C_0 h_P^{*2}] - \frac{2\pi G}{\gamma_0}\rho h_P^{*2}. \tag{2.69}$$

In such equation we evaluate, with h up to 6,000 m,

$$O\left(\frac{\Delta g}{\gamma}C_0 h^2\right) \cong 10^{-4} \cdot 10^{-3} h \leq 0.6\,\text{mm}$$

which is below the millimeter level even for high mountains. Therefore we can reduce (2.42) to

$$N_{P_0} - \zeta_P = \frac{\Delta g_P}{\gamma_0} h_P^* - \frac{2\pi G\rho}{\gamma_0} h_P^{*2} \tag{2.70}$$

If we substitute $h_P^* = H_P + (N_P - \zeta)$ into (2.70) we see that all terms containing $N - \zeta$ are negligible and we have then proved that

$$N_{P_0} - \zeta_P = \frac{\Delta g_P}{\gamma_0} H_P - \frac{2\pi G\rho}{\gamma_0} H_P^2, \tag{2.71}$$

which is the sought equation.

Remark 3. In geophysics it is costumary (cf. also Heiskanen and Moritz (1967) and Torge (2001)) to define the Bouguer anomaly as

$$\Delta g_B = g_P - (2\pi G\rho)H - \left(\frac{\partial\gamma}{\partial h}\right)_0 H - \gamma_0; \tag{2.72}$$

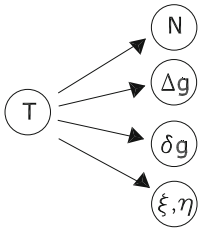


Fig. 2.4 The anomalous potential and some derived fields

if we put approximately

$$\gamma(h^*) \sim \gamma_0 + \frac{\partial \gamma}{\partial h} H$$

we see that (2.72) can be written as

$$\Delta g_B \cong \Delta g_P - (2\pi G\rho) H. \tag{2.73}$$

Comparing with (2.71) we see that one can write

$$N_{P_0} - \zeta_P \cong \frac{\Delta g_B}{\gamma_0} H; \tag{2.74}$$

since $\gamma_0 \cong 10^3$ Gal, if one gives Δg_B in Gal and H in km one gets $N - \zeta$ in meters, or

$$(N - \zeta)(\mathrm{m}) = \Delta g_B(\mathrm{Gal}) \cdot H(\mathrm{km}), \tag{2.75}$$

which is a formula often encountered in literature (cf. Heiskanen and Moritz 1967, Chap. 8, Sect. 13).

2.5 The Remove–Restore Concept

To summarize what we have done till now we could say that after the definitions contained in Chap. 1, we have learned how to relate one to the other the geodetic quantities and, basically, all of them to the gravity potential W.

Since we can always put $W = U + T$, with U an excellent mathematical model that using a few parameters (a, e, ω, GM), can catch the behaviour of W, with a relative accuracy between 10^{-4} and 10^{-5}, we are now permitted to work with "linearized" relations where the anomalous field T is the new unknown object and appears in all the equations only in linear form.

So, from T we can compute with linear operators other fields, as for example it is schematically represented in Fig. 2.4.

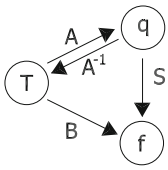


Fig. 2.5 The commutative diagram with T, the observables q and the unknown f

In many instances we have a situation, represented in Fig. 2.5, where from T we can derive two fields q and f, one of which is, in some sense, observable and the other is what we would like to derive

Example 1. Classical is the following example, which is also central for these lecture notes: assume that $q = \Delta g$ restricted to the telluroid is the "observable" field, given by

$$q = \Delta g = -\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma} T = AT; \tag{2.76}$$

take as target field that of height anomalies, $f = \zeta$, describing the separation from the telluroid to the actual earth surface, then (cf. (2.36))

$$\zeta = \frac{1}{\gamma} T = B \cdot T. \tag{2.77}$$

The problem is to find the solver, i.e. the operator S which we can formally write

$$S = BA^{-1};\ f = Sq. \tag{2.78}$$

When the telluroid is approximated by a sphere (see Chap. 3) the operator S takes the name of *Stokes's operator*. Naturally (2.78) is meaningful only if the inverse of A, A^{-1}, exists and is well behaving; this problem will occupy us in the next chapter and is more thoroughly discussed in Part III, Chap. 15.

Example 2. In classical geodetic surveying one uses a total station that, given two points P, Q, provides the observation of the distance D_{PQ} and of the angles (ϑ, α) defined through the relations

$$\cos\vartheta = \mathbf{n}_P \cdot \mathbf{e}_{PQ} \tag{2.79}$$

($\mathbf{e}_{PQ}$ unit vector in the direction PQ)

and

$$tg\alpha = \frac{\mathbf{e}_{PQ} \cdot \mathbf{n}_P \wedge \mathbf{e}_0}{\mathbf{e}_{PQ} \cdot \mathbf{e}_0} \tag{2.80}$$

($\mathbf{e}_0 =$ unit vector in the horizontal plane through P defined by the instrument);

(ϑ, α) are also called respectively zenithal and azimuthal angles (cf. Fig. 2.6)

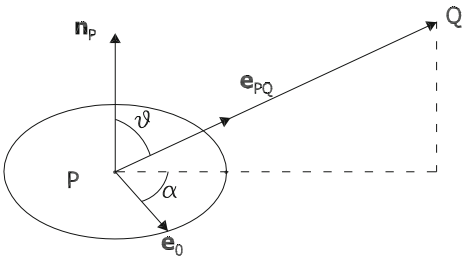


Fig. 2.6 Schematic view of the total station observables

In order to use (2.79) and (2.80) in a simple geometric way one should know $\mathbf{n}_P$ as function of the coordinates of P. So we need $\boldsymbol{\varepsilon}_P$ and we might be willing to estimate it from known free air gravity anomalies in the area. In this case we have again

$$q = \Delta g = AT$$

as in (2.76) and (cf. (2.41))

$$f = \boldsymbol{\varepsilon} = \mathbf{B}T = -\mathbf{e}_\varphi \left(\frac{1}{\gamma}\mathbf{e}_\varphi \cdot \nabla T\right) - \mathbf{e}_\lambda \left(\frac{1}{\gamma}\mathbf{e}_\lambda \cdot \nabla T\right). \tag{2.81}$$

When we use the sphere as a coarse approximation of the telluroid, the solver $S = \mathbf{B}A^{-1}$ in this case takes the name of Vening-Meinesz operator (Heiskanen and Moritz 1967)

Since we have been so successful in including a lot of information on W in a model controlled by few parameters, the possibility has been considered of continuing this job by building other mathematical models which, with a finite number of suitable parameters, would allow us to better approximate T. In other words we construct a model T_M, so that we can put

$$T = T_M + T_r, \tag{2.82}$$

where the subscript r stands for "residual" anomalous potential, not to be confused with a radial derivative.

We anticipate that there are actually two types of models that contribute to T_M; one is global, T_{GM}, and will be treated in detail in the next chapter, the other one, T_{tM}, is much more local, and is used to better account for the short range effects of masses distributed into the topographic layer, i.e. between the actual surface S and some reference surface. This will be better discussed in Chap. 4. For the moment we just say that we can build a model T_M such that

$$O\left(\frac{T_r}{\gamma}\right) \sim 1\,\mathrm{m}; \tag{2.83}$$

this is almost two orders of magnitude smaller than $O(\zeta) = O\left(\frac{T}{\gamma}\right) \cong 100\,\mathrm{m}$.

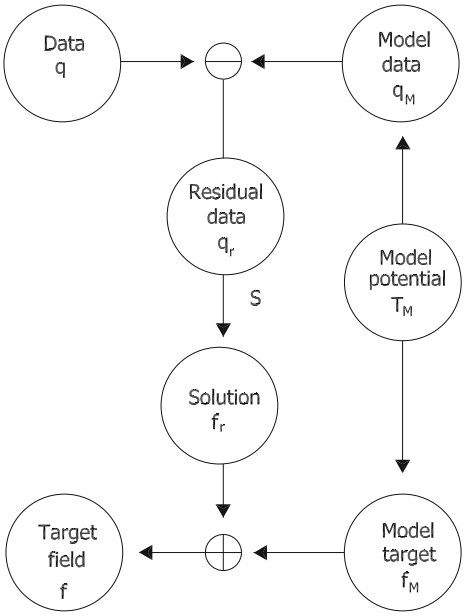


Fig. 2.7 The remove–restore chain

The use of T_r, instead of the whole T, as our unknown allows some significant simplifications of all the expressions where T_r enters. This subject, also described as an application of the *spherical approximation method*, will be discussed in the next section.

Here we want to stress only that the processing chain in this case can be represented as in Fig. 2.7.

As we see the idea is that we subtract from data, i.e. we remove, the knowledge of q that we are able to evaluate from T_M, namely q_M; then we process the residual data q_r to get f_r and we add back to it, i.e. we restore, the knowledge of f, which again we can derive from T_M, to get the final solution.

So the "remove–restore" principle is nothing but claiming that the problem of computing f from q is linear, in our approximation range; as such this principle is indisputable. The advantage of this approach lies in that in the central step, namely the computation of f_r from q_r, we may use a number of rough approximations due to the fact that we know a priori that q_r is small and so model errors in the computation of $S = BA^{-1}$ have a much smaller impact on the final solution.

2.6 The Spherical Approximation Procedure

This is a procedure that rationally exploits the discussion of the previous section. In particular, let $A, B, \ldots$ be any linear operator or functional up to here considered; for instance, consider all the linearized observational functionals of Sect. 2.3.

Many times these contain quantities related to the gravity field U, exactly because they have been derived by linearizing non-linear functionals with respect to $W = U + T$, taking U as the linearization *point*.

The *spherical approximation* consists in systematically using the approximation

$$U \cong \frac{GM}{r} \tag{2.84}$$

in coefficients that do multiply T, i.e. if $a(U)$ is any function of U and $L(T)$ any linear functional of T, independent of U, we shall put

$$a(U)L(T) \cong a\left(\frac{GM}{r}\right) L(T). \tag{2.85}$$

It is important to stress that since (2.84) implies a relative error of the order $e^2 \cong 0.7 \cdot 10^{-2}$ in U and then in $a(U)$ as well, we expect a similar error in $a(U)L(T)$ too.

If the approximation is used for the whole T for which we know that $O\left(\frac{T}{\gamma}\right) \sim 10^2$ m, we might end up with an approximation of the order of 70 cm, in terms of geoid, which is absolutely too coarse for the target established in this book. Nevertheless, if we repeat the procedure when only T_r is used, so that $O\left(\frac{T_r}{\gamma}\right) \sim 1$ m for instance, we expect an error in geoid of the order of 1 cm, which is in the range we can accept.

Remark 4. It has to be stressed that *spherical approximation* does not mean we are approximating the earth surface or the telluroid with a sphere, but at most the ellipsoid $\mathcal{E}$ with a sphere. In addition a procedure like this should never be applied before linearization, because then we would find errors much larger than our target, as it is illustrated in Example 4.

Example 3. We use Bruns's equation (2.36) to illustrate the idea. This equation is

$$\zeta = \frac{T}{\gamma} \tag{2.86}$$

and if we use the expression of the ellipsoidal γ_e (see (1.145)) simplified to

$$\gamma_e \cong \gamma_a(1 + 5 \cdot 10^{-3} \sin^2 \varphi - 0.3 \cdot 10^{-3} h)$$
$$(\gamma \text{ in Gal}, h \text{ in km})$$

we find

$$\zeta_e \cong \frac{T}{\gamma_a}(1 - 5 \cdot 10^{-3} \sin^2 \varphi + 0.3 \cdot 10^{-3} h). \tag{2.87}$$

The spherical approximation in this case would be

$$\begin{cases} U_s \sim \dfrac{GM}{r},\ \gamma_s = \dfrac{GM}{r^2} \\ r \cong R + h, \zeta_s = \dfrac{T}{\gamma_s} \end{cases} \tag{2.88}$$

with R the mean radius of the ellipsoid.

We note that γ_a is such that

$$\gamma_a = \gamma(\varphi = 0,\ h = 0) \cong \frac{GM}{R^2} \tag{2.89}$$

so that if we take $h = 0$ in (2.87) and $r = a$ in (2.88), we find

$$\begin{aligned} \zeta_e - \zeta_s &= \frac{T}{\gamma_a}(1 - 5 \cdot 10^{-3} \sin^2 \varphi) - \frac{T}{\gamma_a} \\ &= -\frac{T}{\gamma_a} 5 \cdot 10^{-3} \sin^2 \varphi. \end{aligned}$$

This shows that the order of magnitude of the error is

$$O(\zeta_e - \zeta_s) = 5 \cdot 10^{-3} O(\zeta_s) \cong 0.5\,\text{m},$$

which, as anticipated, is by far too large. If on the contrary we apply (2.88) only to the residual T_r and we assume that $O\left(\frac{T_r}{\gamma_a}\right) \cong 1$ m, we find

$$O\left((\zeta_r)_e - (\zeta_r)_s\right) \cong 0{,}5 \text{ cm}$$

which is within our target.

Example 4. Take the approximation used in Example 3 for γ_e and note that, close to the ellipsoid, one can write

$$U_e \cong W_0 - \gamma_a(h + 5 \cdot 10^{-3} \sin^2 \varphi \cdot h - 0.15 \cdot 10^{-3} h^2), \tag{2.90}$$

with

$$W_0 = \frac{GM}{R}.$$

Since

$$U_s \cong \frac{GM}{R + h} \cong W_0 - \gamma_a h + \frac{\gamma_a}{R} h^2, \tag{2.91}$$

comparing with (2.90) one finds

$$U_e(\varphi, h) \cong U_s(\varphi, h) - 5\gamma_a \cdot 10^{-3} \sin^2 \varphi \cdot h.$$

Therefore if we apply the spherical approximation directly to W, and not to coefficients multiplying linear expressions in T, we see that we significantly modify the definition of T, in fact

$$W - U_s \cong -5\gamma_a 10^{-3} \sin^2 \varphi \cdot h + T,$$

which means that $\frac{T}{\gamma}$ is modified by a term of the order of 5 m for 1 km of altitude.

The two examples above show that in any case if one wants to use the spherical approximation, this has to be done correctly only *after* linearization and *after* the reduction of T to a residual component T_r. Now, in order to be more precise, let us specify that the use of a spherical approximation implies

$$U \sim \frac{GM}{r} \tag{2.92}$$

$$\boldsymbol{\gamma} \sim -\frac{GM}{r^2}\mathbf{e}_r \tag{2.93}$$

$$\gamma \sim \frac{GM}{r^2} \tag{2.94}$$

$$\boldsymbol{\nu} \sim \mathbf{e}_r \tag{2.95}$$

$$\gamma' \sim -2\frac{GM}{r^3} \tag{2.96}$$

$$r \sim R + h \tag{2.97}$$

With the use of such formulas we find for the main observables considered in Sect. 2.3

$$\zeta \cong \frac{T}{\frac{GM}{r^2}} \tag{2.98}$$

$$\delta g = -\frac{\partial T}{\partial r} \tag{2.99}$$

$$\Delta g = -\frac{\partial T}{\partial r} - \frac{2}{r}T \tag{2.100}$$

$$\boldsymbol{\varepsilon} = -\frac{1}{\gamma}\left(\mathbf{e}_\lambda \frac{1}{r\cos\varphi}\frac{\partial T}{\partial \lambda} + \mathbf{e}_\varphi \frac{1}{r}\frac{\partial T}{\partial \varphi}\right); \tag{2.101}$$

note has to be taken that (λ, φ) in (2.101) are the spherical longitude and latitude and $(\mathbf{e}_\lambda, \mathbf{e}_\varphi)$ are spherical unit vectors too.

Remark 5. It has to be said that with the present computing capacity, the need of simplifying formulas to facilitate the numerical work has not reason to be any more considered. Therefore the use of spherical approximation should generally be restricted to analytical applications or to simulations for noise propagation studies.

2.7 A Review of Observation Equations with Unknown Reference Potential

As promised in Sect. 2.3 we need now to review our observation equations in which $W(P)$ was introduced as an observable and adopt a more realistic model accounting for the fact that what we can observe in reality is only a potential difference. For instance we observe $W(P) - \overline{W}$, where $\overline{W}$ is the potential at some reference point $\overline{P}$. When $\overline{P}$ is a tide gauge we expect such a point to be lying in the vicinity of the geoid in a range of a few meters. Yet the value $\overline{W}$ will be different from W_0 because $\overline{P}$ is not exactly on the geoid $\mathcal{G}$.

Let us put

$$\widetilde{W}(P) = W_0 + W(P) - \overline{W}; \tag{2.102}$$

since W_0 is known and $W(P) - \overline{W}$ observed, we can take $\widetilde{W}(P)$ itself as an observable and see what happens to observation equations if $\widetilde{W}(P)$ instead of $W(P)$ is considered as known.

To proceed, we note first that (2.102) can be written as well as

$$\widetilde{W}(P) = W(P) + \delta W_0 \tag{2.103}$$

with

$$\delta W_0 = W_0 - \overline{W}; \tag{2.104}$$

δW_0 is an unknown parameter that will enter into our observation equations, into the vector $\mathbf{x}$ according to our general scheme of formula (2.3).

Returning to Sect. 2.3 we find for (2.26) the new formulation

$$\widetilde{W}(P) - U(P) = T(P) + \delta W_0, \tag{2.105}$$

where on the LHS we have known and on the RHS we have unknown quantities. Equation 2.30, for gravity disturbance, is unchanged, because it refers to a point P of known coordinates.

The case of point 3 culminating in (2.36) and (2.38), needs to be carefully reviewed.

In fact since now only $\widetilde{W}(P)$ is available, we cannot compute the normal height h_P^*. At most we can put, as a new definition of the linearization height,

$$\begin{aligned} U(\widetilde{h}) &= \widetilde{W}(P) = W(P) + \delta W_0 \\ &= U(h^*) + \delta W_0. \end{aligned} \tag{2.106}$$

Note that here too, as in Sect. 2.3, we highlight only the dependence of functions on height variables, neglecting the horizontal coordinates of points. From (2.106) we see that

$$\delta W_0 = U(\widetilde{h}) - U(h^*) = \gamma(\widetilde{h})\delta h \tag{2.107}$$

where we have put

$$\delta h = h^* - \widetilde{h}. \tag{2.108}$$

Accordingly we can compute from the observed $g(P)$ a different kind of anomaly, i.e.

$$D\widetilde{g}(\widetilde{h}) = g(h) - \gamma(\widetilde{h}) \tag{2.109}$$

which we elaborate in the following from

$$\begin{aligned} D\widetilde{g} &= g(h) - \gamma(h^*) + \gamma(h^*) - \gamma(\widetilde{h}) \\ &\cong \Delta g + \gamma'(\widetilde{h})\delta h. \end{aligned} \tag{2.110}$$

Taking (2.107) into account, (2.110) becomes

$$D\widetilde{g} = \Delta g + \frac{\gamma'}{\gamma}\delta W_0. \tag{2.111}$$

Finally, recalling (2.38), we find the modified observation equation

$$D\widetilde{g} = -\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma}T + \frac{\gamma'}{\gamma}\delta W_0. \tag{2.112}$$

The case of point 4, in section 2.3, has not been worked out, so we will not consider it here. The equations of 5 do not change because again here we assume P to have known coordinates. The cases of points 6 and 7 are in fact modified because in the present situation the orthometric height cannot be considered as observable. In fact if we take the point $\overline{P}$ as reference (origin) for a new system of orthometric heights we will have for any point P (see Fig. 2.8)

$$H_P = \overline{H}_P + \delta H_P \tag{2.113}$$

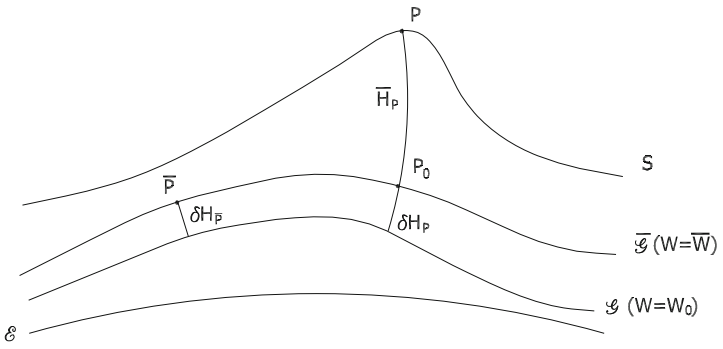


Fig. 2.8 The geometry of the change of height datum form $\mathcal{G}$ to $\overline{\mathcal{G}}$

where only $\overline{H}_P$ is available from measurements. We need now to relate δH_P to $\delta H_{\overline{P}}$ and this last to δW_0, which is our basic unknown parameter.

This is easy to do by writing the linearized relation

$$\delta W_0 = W_0 - \overline{W} \cong g_{P_0} \delta H_P, \tag{2.114}$$

holding true for every P on the surfaces, including the actual P. Since γ_{P_0} differs form g_{P_0} at most by a factor $10^{-4}\gamma$, (2.114) can be further approximated as

$$\delta W_0 = \gamma_{P_0} \delta H_P, \tag{2.115}$$

or, going back to (2.113)

$$H_P = \overline{H}_P + \frac{\delta W_0}{\gamma_{P_0}}. \tag{2.116}$$

This was the sought relation that can be substituted into any observation equation where use of H_P is made. Note that (2.115) implies

$$\delta H_P = \frac{\gamma_{\overline{P}}}{\gamma_{P_0}} \delta \overline{H}. \tag{2.117}$$

Since at the level of the sea γ_P varies at most by a factor $5 \cdot 10^{-3}\gamma$ from pole to equator, we see that with a displacement $\delta \overline{H} = 2\,\mathrm{m}$ of the reference surface we have a variability of δH_P at most of 1 cm. In other words, for many applications the change of reference surface of the heights can be accounted for by the addition of a constant to observed orthometric heights. Finally we don't discuss here the leveling equation because in that context there is only a very weak dependence on Δg, a quantity that changes with δW_0.

The more realistic situation where equations like (2.112) with many different unknown constants δW_{01} due to different origins of different height systems is analyzed in detail in Part II, Chap. 11.

2.8 Exercises

Exercise 1. In the spirit of the proof of (2.38), consider the correspondence of $\mathbf{r}_P \equiv (\lambda, \varphi, h)$ with any other approximate point $\mathbf{r}_{\widetilde{P}} = (\lambda, \varphi, \widetilde{h})$; then put $\zeta = h - \widetilde{h}$ and prove that instead of (2.36) and (2.38) the two generalized relations hold

$$\zeta = \frac{1}{\gamma(\widetilde{P})}\{T(\widetilde{P}) - [W(P) - U(\widetilde{P})]\},$$

$$-T' + \frac{\gamma'(\widetilde{P})}{\gamma(\widetilde{P})}T = g(P) - \gamma(\widetilde{P}) + \frac{\gamma'(\widetilde{P})}{\gamma(\widetilde{P})}[W(P) - U(\widetilde{P})].$$

Observe that if $\widetilde{h} = H$ is chosen, then ζ derived by the above formula is directly the geoid undulation N, in view of (1.152).

(**Hint:** note that, to the first order,

$$W(P) = U(\widetilde{h} + \zeta) + T(P) \cong U(\widetilde{P}) - \gamma(\widetilde{P})\zeta + T(\widetilde{P})$$

$$g(P) = \gamma(\widetilde{h}) + \gamma'(\widetilde{h})\zeta - \frac{\partial T}{\partial h}(\widetilde{P})$$

and continue as in Sect. 2.3, point 3).

Exercise 2. Consider the case of point 4 in Sect. 2.3, and derive the corresponding linearized observation equation which, applied at the boundary, gives rise to the so-called *vector Molodensky problem*. To do that consider the mapping

$$\lambda_{\widetilde{P}} = \Lambda_P, \ \varphi_{\widetilde{P}} = \Phi_P, \ U(\widetilde{P}) = W(P).$$

Put

$$\boldsymbol{\xi} = \mathbf{r}_P - \mathbf{r}_{\tilde{P}}, \ \Delta\mathbf{g} = \mathbf{g}(P) - \boldsymbol{\gamma}(\widetilde{P}),$$

$$M(\widetilde{P}) = [M_{ik}(\widetilde{P})] = \left[\frac{\partial \gamma_i}{\partial x_k}(\widetilde{P})\right];$$

M is the matrix of second derivatives of the normal potential, also called *Marussi tensor*.

Prove that the sought equations are

$$\boldsymbol{\xi} = M^{-1}[\Delta\mathbf{g} - \nabla T]$$

$$-\boldsymbol{\gamma} \cdot M^{-1}\nabla T + T = -\boldsymbol{\gamma} \cdot M^{-1}\Delta\mathbf{g}$$

(**Hint:** write

$$\mathbf{g}(P) = \boldsymbol{\gamma}(P) + \nabla T(P) \cong \gamma(\widetilde{P}) + M\boldsymbol{\xi} + \nabla T(\widetilde{P})$$
$$W(P) = U(P) + T(P) \cong U(\widetilde{P}) + \boldsymbol{\gamma}(\widetilde{P}) \cdot \boldsymbol{\xi} + T(\widetilde{P}),$$

derive $\boldsymbol{\xi}$ from the first and substitute into the second, observing that $W(P) = U(\widetilde{P})$).

Exercise 3. Find a direct, though more loosely approximated relation between H and h^*, for a point P where (λ, φ, g) are also known, considering that

$$U(h^*) \cong U_0 - \gamma(P_0)h^*$$
$$W(H) \cong W_0 - \left[g_P - \frac{\partial g}{\partial H}H\right]H$$
$$U_0 = W_0$$
$$-\frac{\partial g}{\partial H} \cong -\frac{\partial \gamma}{\partial h}(P_0) + 4\pi G\rho \cong 0.1966\,\mathrm{Gal\,km^{-1}}.$$

Appendix

A.1

We want to find a manageable expression for the sum of leveling increments along a line, proving (2.58).

To this aim we go back to (2.55) and substitute

$$dh = dH + dN = dH - \boldsymbol{\varepsilon}_0 \cdot d\mathbf{r}_0$$

in it. We receive (see Fig. 2.2 for the notation)

$$\delta L = (\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0) \cdot d\mathbf{r} + \boldsymbol{\varepsilon}_0 \cdot (d\mathbf{r} - d\mathbf{r}_0) + dH$$
$$= [(\mathbf{n} - \mathbf{n}_0) - (\boldsymbol{\nu} - \boldsymbol{\nu}_0)] \cdot d\mathbf{r}_h + \frac{H}{R}\boldsymbol{\varepsilon}_0 \cdot d\mathbf{r}_0 + dH,$$

because, with a good approximation, $d\mathbf{r} - d\mathbf{r}_0 = \dfrac{H}{R}d\mathbf{r}_0 + dh\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}_0$ is orthogonal to $\boldsymbol{\nu}$. Since $\int_{\widehat{AB}} \boldsymbol{\varepsilon}_0 \cdot d\mathbf{r}_0$ is the variation of N, which is at most a few meters, even for points A, B far away dozens of kilometers, and $\dfrac{H}{R} < 10^{-3}$, we can drop the term $\dfrac{H}{R}\boldsymbol{\varepsilon}_0 \cdot d\mathbf{r}_0$; in other words we can take $d\mathbf{r}_h \sim d\mathbf{r}_0$ in this computation. Now, recalling (1.75), we can write

$$\mathbf{n} - \mathbf{n}_0 = \int_{P_0}^{P} \nabla_h \log g dH$$

and similarly

$$\boldsymbol{\nu} - \boldsymbol{\nu}_0 = 0 = \int_{P_0}^{P} \nabla_h \log \gamma_0 dh;$$

where γ_0 is constant, along the vertical line, so that the latter identity reduces to $0 = 0$, because $\boldsymbol{\nu}$ is indeed constant along the normal to the ellipsoid. So we have

$$\begin{aligned}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0 &= \int_{P_0}^{P} \nabla_h \log \frac{g}{\gamma_0} dh = \int_{P_0}^{P} \nabla_h \log \left(1 + \frac{g - \gamma_0}{\gamma_0}\right) dh \cong \\ &= \int_{P_0}^{P} \nabla_h \left(\frac{g - \gamma_0}{\gamma_0}\right) dh = \nabla_h \int_{P_0}^{P} \left(\frac{g - \gamma_0}{\gamma_0}\right) dh + \\ &\quad - \left(\frac{g(P) - \gamma_0}{\gamma_0}\right) \nabla_h h_P .\end{aligned}$$

The last step is justified by the well-known differentiation rule

$$D_x \int_0^{g(x)} f(x,t)dt = f[x, g(x)] \cdot g'(x) + \int_0^{g(x)} \frac{\partial}{\partial x} f(x,t)dt.$$

Summarizing and going back to (2.118), we find

$$\delta L = \nabla_h \left[\int_{P_0}^{P} \left(\frac{g - \gamma_0}{\gamma_0}\right) dh\right] \cdot d\,\mathbf{r}_h - \frac{g(P) - \gamma_0}{\gamma_0} dh + dH. \tag{2.118}$$

As it is shown in Sect. 2.4,

$$\int_{P_0}^{P} \frac{g - \gamma_0}{\gamma} dh \cong N_{P_0} - \zeta_P, \tag{2.119}$$

for which an explicit formula, as function of H, is given by (2.71). Moreover in (2.118) we can substitute δL back for dh to the effect that one can write

$$\delta L = d(N - \zeta) - \frac{g(P) - \gamma_0}{\gamma_0} \delta L + dH,$$

which finally integrated along the line $\overset{\frown}{AB}$ yields

$$\Delta_{AB} L = (N_B - \zeta_B) - (N_A - \zeta_A) - \int_{\overset{\frown}{AB}} \frac{g - \gamma_0}{\gamma_0} \delta L + H_B - H_A, \tag{2.120}$$

namely the formula we wanted to prove.

A.2

We want to prove formula (2.65) for the vertical gradient of Δg as function of the normal height h^*. We adopt symbols and notation of Sect. 2.4. To this aim we note first of all that in (2.63) we need Δg, so that we have to convert (2.64) into an equation for the vertical continuation of Δg.

To this aim we write the analogous of (2.64) for the normal field, i.e.

$$\frac{\partial \gamma}{\partial h} = -2C_0\gamma - 2\omega^2; \tag{2.121}$$

note that (2.121) can be written for any point along the ellipsoidal normal, for instance at Q^* instead of Q, but we are not allowed to substitute $\frac{\partial}{\partial h^*}$ for $\frac{\partial}{\partial h}$ in (2.121) because h^* *is not* a linear function of h. So we must transform $\frac{\partial}{\partial H}$ in (2.64) into $\frac{\partial}{\partial h}$, then we subtract (2.121) computed at Q^* from (2.64) and finally we transform $\frac{\partial}{\partial h}$ into $\frac{\partial}{\partial h^*}$.

As for $\dfrac{\partial g}{\partial H}$ we can write

$$\frac{\partial g}{\partial H} = \mathbf{n}\cdot\nabla g = (\mathbf{n}-\boldsymbol{\nu})\cdot\nabla g + \boldsymbol{\nu}\cdot\nabla g \cong \boldsymbol{\varepsilon}\cdot\nabla\gamma + \boldsymbol{\nu}\cdot\nabla g. \tag{2.122}$$

In (2.122) we evaluate the order of magnitude

$$\begin{aligned} O(\boldsymbol{\varepsilon}\cdot\nabla\gamma) &= O(\boldsymbol{\varepsilon}\cdot\nabla_t\gamma) = O\left(|\boldsymbol{\varepsilon}|\frac{1}{R}\frac{\partial\gamma}{\partial\varphi}\right) \\ &= O\left(|\boldsymbol{\varepsilon}|\frac{5\cdot10^{-3}\gamma}{R}\right) = 5\cdot10^{-7}\frac{\gamma}{R} \end{aligned} \tag{2.123}$$

where we have used (1.145) and (1.181).

Therefore this term contributes to g, and then to Δg, at height h with an error $\delta\Delta g$ of the order of magnitude of $5\cdot10^{-7}\frac{\gamma}{R}h$, or, equivalently, of $5\cdot10^{-7}\frac{\gamma}{R}H$.

As a consequence of (2.63), to evaluate the error induced by neglecting $\boldsymbol{\varepsilon}\cdot\nabla g$ in computing $N-\zeta$ one has to assess the order of magnitude of $\delta\Delta g$ integrated in H, i.e., observing that in the topographic layer one has $O\left(\frac{H}{R}\right)\sim 10^{-3}$,

$$O(\delta[N-\zeta]) = O\left(\boldsymbol{\varepsilon}\cdot\nabla\gamma\frac{H^2}{\gamma}\right)\sim 5\cdot10^{-7}\frac{H^2}{R}\sim 5\cdot10^{-10}H; \tag{2.124}$$

this shows that the term in question doesn't matter in our computation. So we can write

$$\frac{\partial g(Q)}{\partial H}\cong\boldsymbol{\nu}\cdot\nabla g(Q) = \frac{\partial}{\partial h}g(Q)$$

in (2.64) and work on the right hand side with an obvious approximation to arrive at the equation

$$\frac{\partial g(Q)}{\partial h} = -2[C(Q) - C_0(Q^*)]\gamma - 2C_0(Q^*)g(Q) + 4\pi G\rho - 2\omega^2. \quad (2.125)$$

If we can prove that in (2.125) the term

$$[C(Q) - C_0(Q^*)]\gamma \cong [C(Q) - C_0(Q)]\gamma + C_0'(Q)\gamma\zeta \quad (2.126)$$

is negligible, we are left with the equation

$$\frac{\partial}{\partial h}g(Q) = -2C_0(h^*)g(Q) + 4\pi G\rho - 2\omega^2 \quad (2.127)$$

We evaluate (2.126) in two steps. First we use the following estimate, derived from several numerical experiments,

$$O([C(Q) - C_0(Q)]) \cong \frac{10^{-3}}{R}\ ; \quad (2.128)$$

as always, $O(\,)$ means the order of magnitude of the maximum value, as the standard deviation of $C(Q) - C(Q_0)$ is easily one order of magnitude smaller. Then we evaluate the impact of this term on $N - \zeta$ by considering the corresponding error $[C(Q) - C_0(Q)]\gamma$ integrated in H, once to give its impact on g, and then a second time, divided by γ, to give the impact on $N - \zeta$ (see (2.63)). The result is

$$O(\delta[N - \zeta]) = O\left([C(Q) - C_0(Q)]\gamma \cdot \frac{H^2}{\gamma}\right) \cong O\left(\frac{10^{-3}H}{R} \cdot H\right) = 10^{-6}H,$$

which is negligible because it gives at maximum an error of 1 mm/km of altitude. As for the second addendum in (2.126) we use the rough approximation

$$|C_0'| \cong \frac{1}{R^2},$$

yielding

$$O(\delta[N - \zeta]) = O\left(C_0'\gamma\zeta \cdot \frac{H^2}{\gamma}\right) = O\left(\frac{H^2}{R^2}\zeta\right) = 10^{-6}\zeta;$$

this is totally negligible since it is below the millimeter for any height up to 6,000 m.

So we know that (2.127) is correct and we can subtract (2.121) from it, to get

$$\frac{\partial}{\partial h}g(Q) - \frac{\partial}{\partial h}\gamma(Q^*) = -2C_0(Q^*)[g(Q) - \gamma(Q^*)] + 4\pi G\rho,$$

namely

$$\frac{\partial}{\partial h}\Delta g = -2C_0(h^*)\Delta g + 4\pi G\rho. \tag{2.129}$$

Now from

$$h = h^* + \zeta$$

we see that (cf. (2.62))

$$\frac{\partial}{\partial h} = (1 - \zeta')\frac{\partial}{\partial h^*} = \left(1 + \frac{\Delta g}{\gamma}\right)\frac{\partial}{\partial h^*}.$$

So, omitting all second order terms that are easily verified to be negligible, we write (2.129) in the form

$$\frac{\partial}{\partial h^*}\Delta g = -2C_0(h^*)\Delta g + 4\pi G\rho. \tag{2.130}$$

Finally, we want to show that in (2.130) we can consider C_0 and ρ as constants.

We reason again in terms of orders of magnitude of maximum errors. So if we use the rough estimate

$$O(|C_0(0) - C_0(h^*)|) = \frac{1}{R} - \frac{1}{R + h^*} \cong \frac{h^*}{R^2},$$

we see that one has for the error $\delta(N - \zeta)$, after the usual double integration on H,

$$O(\delta[N - \zeta]) = O\left(\frac{h^*}{R^2}\gamma\frac{H^2}{\gamma}\right) = 10^{-6}h^*,$$

namely 1 mm/km of altitude in worst case.

In parallel one can consider that in the crust ρ can vary around its mean value, $\overline{\rho} = 2.67\,\mathrm{g\,cm^{-3}}$, by no more than 10%, but

$$0,1 \cdot 4\pi G\overline{\rho} \cong 0.02\,\mathrm{mGal\,m^{-1}}$$

so that the corresponding error on $\delta[N - \zeta]$ is of the order of

$$\begin{aligned} O(\delta[N - \zeta]) &= O\left(0.1 \cdot 4\pi G\overline{\rho}\frac{H^2}{\gamma}\right) \\ &= 2 \cdot 10^{-8}H^2 (H \text{ in meters}) \end{aligned} \tag{2.131}$$

Therefore, with $H = 10^3$ m, our maximum error becomes 2 cm, which is certainly not too small. Yet the following has to be considered: first of all sometimes we have geological maps that could help us to use a value of ρ good up to 1%, giving in (2.131) an error smaller by one order of magnitude; a variation of 0.267 g cm^{-3} in the surface density has to be considered very large. Finally, this is certainly the most uncertain information we can have in physical geodesy so that, when we really need $N - \zeta$, we have to live with errors of this magnitude.

So now (2.130) can be written as

$$\frac{\partial}{\partial h^*}\Delta g = -2C_0\Delta g + 4\pi G\rho \tag{2.132}$$

with C_0 and ρ considered as constants.

Chapter 3
Harmonic Calculus and Global Gravity Models

3.1 Outline of the Chapter

The chapter is devoted to the construction and manipulation of so-called *global models of the anomalous potential*.

These are basically truncated series of spherical or ellipsoidal harmonics. These functions are so important in physical geodesy that they need to be carefully introduced and their mathematical properties have to be known by everyone dealing with gravity field representations.

As always, we start from Newton's formula relating mass density and gravitational potential. If we use a similar representation for the normal potential, we may conclude that the anomalous potential can be represented too in the form of a Newtonian integral. Now the development of Newton's kernel, i.e. the inverse of the distance between two points, in a series of polynomials called Legendre polynomials, is a very classical issue presented in Sect. 3.2.

Legendre polynomials are then studied in Sect. 3.3.

In particular their integral properties (a reproducing property by convolution on the unit sphere as well as the L^2 orthogonality on the unit interval $[-1,1]$) and differential properties are established. In this way we obtain a first representation of the potential as a series of harmonic functions, each decreasing at infinity as an inverse power of r. The series is clearly converging outside any sphere encompassing all the masses.

In Sect. 3.4 the so-called *surface spherical harmonics* $\{Y_{nm}\}$ are introduced. The precise construction of these functions is delayed to Part III, where the full theory is derived from the study of spaces of harmonic polynomials. One basic result proved in Part III is the so-called *summation theorem* reported in (3.54).

This provides a fundamental relation between spherical harmonics of degree n and order m and the corresponding Legendre polynomials of degree n.

If one then defines the solid spherical harmonics $\{S_{nm}\}$ as the surface spherical harmonics of degree n divided by r^{n+1}, one immediately sees that our T can be

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_3,

expressed a series of these solid spherical harmonics, converging on any sphere lying outside the masses.

The sequence $\{Y_{nm}\}$ is then studied in the space of functions which are square integrable (L^2) on the unit sphere; it turns out that this is an orthonormal complete sequence implying that any L^2 function can be developed into a series of $\{Y_{nm}\}$. This fact, together with the statement that $S_{nm} = r^{-(n+1)}Y_{nm}$ are harmonic functions, which coincide with Y_{nm} on the unit sphere, allow the solution of classical geodetic problems for the sphere giving rise to the use of Poisson, Hotine and Stokes kernels.

Such problems, though not realistic, mimic for the case of a spherical boundary other problems that can be formulated as *boundary value problems* where the unknown T has to be harmonic outside a given surface S, and it has to satisfy some differential relation on the surface itself.

However a second theorem, namely Theorem 3, stating that given any reasonable surface S the traces of $\{S_{nm}\}$ on S form a complete system in $L^2(S)$, is even more important for the practice of building approximate solutions to geodetic boundary value problems (BVP). These in fact bring us much closer to a realistic situation than the previous examples with a spherical boundary.

So till now we have learnt how to solve exactly a BVP for the Laplace equation in the exterior of a spherical domain, typically we have Stokes's formula, but we have a realistic problem with a non-spherical surface and boundary values (e.g. gravity anomalies) on it.

If we could find a function harmonic in a domain larger than the exterior of S, in fact harmonic down to some internal sphere (also called *Bjerhammar sphere*), we could still use Stokes's representation for this function and impose on it that the boundary values of the gravity anomalies be attained on S.

This is not possible in general; the values of a harmonic function and of all its derivatives inside the domain of harmonicity are extremely smooth and so only very particular functions on S can have a harmonic continuation down to an internal sphere.

Nevertheless since real data are only pointwise and finite in number, we can always interpolate them by a function harmonic down to any fixed internal sphere. This point of view, which is also strictly related to Theorem 3, is established in Sect. 3.5 in the form of a new Theorem 4 known in geodesy as *Krarup's theorem*.

In Sect. 3.6 the spherical set up of the previous two sections, is generalized to domains with ellipsoidal boundary. It is proved that by the use of suitable ellipsoidal coordinates, we can build a new system of functions, called *ellipsoidal harmonics*, that are orthonormal in the space of functions square integrable on the ellipsoid and even complete in such a space.

Numerical instability problems related to ellipsoidal harmonics are discussed and effective, computable approximate formulas are given.

In Sect. 3.7 we formally establish the problem of the determination of T from Δg in the form of a BVP, namely the Molodensky problem, discussing as well other BVP's that might become even more important in future.

Numerical methods, typical of functional analysis, like least squares or Galerkin method, are discussed later in Part III, Sect. 14.5 of the book. In that chapter the relation of these methods to more practical geodetic solutions, is also highlighted.

Finally in Sect. 3.8 we discuss two indexes that, though very coarse, are quite essential in expressing the quality of the solution, accounting for two distinct effects.

The first is the presence of noise in the observations used to estimate the global model. The noise in fact propagates from the measurements to the solution and determines what is called the *commission error*. This is basically the average of the L^2 norm of the error function constructed propagating the noise from measurements to the harmonic coefficients of the global model.

The second effect on the other hand is the error that we commit because, instead of estimating the full anomalous potential, we aim only at its projection on a finite dimensional subspace, generated by linear combinations of solid spherical harmonics up to a maximum degree. The norm of the reminder is the omission error. This has an easy relation to the coefficients left out from the truncated series, when this is convergent.

In fact, this is the sum of the squares of all coefficients of degree higher than N. But of course this is just an unknown quantity which we will never know a priori. However by looking at the so-called degree variances (i.e. the sum over all orders of the squares of the coefficients of a certain degree) one can guess some law for its decay that can allow the computation of the omission error.

One law of this kind, of historical nature, is Kaula's law; other laws, much more realistic, are shown in the text. The above mentioned models can be used as different cases to make predictions and this has the scope to give a feeling of the range of variability of this error, which after all depends from a pure guess based on empirical data.

3.2 The Newton Integral Representation of the Anomalous Potential

We have defined in Sect. 1.10 the anomalous potential of the gravity field as

$$T(P) = W(P) - V(P). \tag{3.1}$$

This definition eliminates the centrifugal potential and leaves us with

$$T(P) = V(P) - V_e(P) \tag{3.2}$$

where $V(P)$ is the actual gravitational potential of the earth, i.e. the Newtonian integral (1.14), while $V_e(P)$ is the ellipsoidal gravitational potential, given explicitly by formula (1.127).

Table 3.1 A simplified version of the PREM model. For the first two layers we give average values; for the others we give values across discontinuities; for a rough approximation one can imagine a linear dependence on depth within the layers

Depth (km)	Earth layer	Density (g cm^{-3})
0	Topographic layer	2.67
33	Crust	2.8
400	Upper mantel	3.3
		3.5
670	Transition zone	3.7
		4.1
2,900	Lower mantel	4.4
		5.6
5,100	Outer core	10.0
		12.3
6,400	Inner core	12.9
		13.2

Indeed if we knew exactly the mass density, $\rho(Q)$, we would have a little need of physical geodesy, in fact physical geodesy is precisely the science of how to deal with the gravity field without knowing ρ.

However we are aware that a certain ρ exists and we have to some extent a knowledge of this function by means of various geophysical observations and models; primarily geodynamic models relating seismic observations to the density distribution.

In fact we have already shown through Examples 1 and 2, that many (in fact infinite) internal mass distributions generate the same outer potential and this proves that the density cannot be derived from the knowledge of the outer gravity field only.

However guessing the internal mass distribution is an old scientific problem which can be traced back to Clairaut and his *Theorie de la figure de la terre, tirée des principes de l'hydrostatique* (1743). On this item, its geodynamical and geodetic relevance see also Moritz (1990) and Sabadini and Vermeersen (2004). Here we give, just for information, the model of an inner density distribution derived from a famous preliminary earth model (PREM) by Dziewonsky and Anderson (1981) (Table 3.1).

Naturally a model like this, where the density is only a function of depth, i.e. of the radius, can generate only an exterior field of the type $\frac{GM}{r}$, as shown in Sect. 1.3. In particular it does not even account for the ellipsoidal shape of the earth nor for the topography. Yet one can prove that there is a density ρ_e which is layered, i.e. it is constant in layers between concentric ellipsoids, and generates an exterior potential equal to V_e (Marussi 1985; Sünkel and Tscherning 1981; Moritz 1990). Here we are not interested in a precise definition of ρ_e, but rather in knowing that it exists and that it can be interpreted as a kind of average of the actual density ρ in each layer. As a result of this reasoning we see that we can put

$$T(P) = V(P) - V_{e\rho}(P)$$
$$= G\int_B \frac{\rho(Q)}{\ell_{PQ}} dB_Q - G\int_{B_e} \frac{\rho_e(Q)}{\ell_{PQ}} dB_Q, \tag{3.3}$$

where B_e is the volume occupied by the ellipsoid with surface $\mathcal{E}$. If we define a density anomaly as

$$\delta\rho(Q) = \begin{cases} \rho(Q) & \text{in } B\backslash B_e \\ -\rho_e(Q) & \text{in } B_e\backslash B \\ \rho(Q) - \rho_e(Q) & \text{in } B_e \cap B, \end{cases} \tag{3.4}$$

we see that (3.3) can be written as a unique Newtonian integral

$$T(P) = G\int_B \frac{\delta\rho(Q)}{\ell_{PQ}} dB_Q. \tag{3.5}$$

It has to be clear that (3.4) and (3.5) do hold when P is outside B, and when P is in the topographic layer, $B\backslash B_e$. As a matter of fact, (3.5) rather ignores the case $P \in B_e\backslash B$, because this set is so small (in fact so thin) and mostly related to the oceanic area that it is not so relevant for the present discussion. However when P enters into the ellipsoid the potential

$$V_{e\rho} = G\int_{B_e} \frac{\rho_e(Q)}{\ell_{PQ}} dB_Q \tag{3.6}$$

becomes different from V_e, i.e.

$$V_{e\rho}(P) \neq V_e(P), \ P \in B_e\ ; \tag{3.7}$$

in fact $V_e(P)$ is still harmonic in B_e, apart from a small area on the equatorial plane, while $V_{e\rho}$ is obviously not harmonic. Therefore, when defining $T(P)$ inside the ellipsoid, one has to be careful and state explicitly whether one uses the original definition (3.1) or rather one wants to use (3.3). As geodesists we don't suffer of this ambiguity because we don't need to go inside $\mathcal{E}$ more than a few hundred meters, and therefore we shall use irrespectively (3.1) and (3.5).

Now we want to pick up an argument that we have considered in Sect. 1.3 and push it further; namely we want to study the behaviour of $T(P)$ when r_P is large enough.

To this aim let us consider Fig. 3.1; we call a Brillouin sphere any sphere that encloses completely the masses and we denote by $\overline{R}$ the minimum among the radius of the Brillouin spheres.

Then take any point P with $r_P > \overline{R}$, so that for sure $P \in \Omega$. We can write, $\forall Q \in B$,

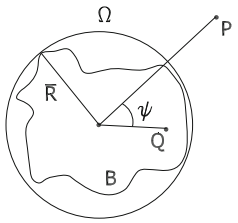


Fig. 3.1 The minimal Brillonin sphere and the points $P, r_P > \overline{R}$ and $Q, r_Q \leq \overline{R}$

$$\frac{1}{\ell_{PQ}} = \frac{1}{\sqrt{r_P^2 + r_Q^2 - 2r_P r_Q \cos\psi}} = \frac{1}{r_P}\frac{1}{\sqrt{1 + s^2 - 2st}} \tag{3.8}$$

where we have put

$$s = \frac{r_Q}{r_P},\ t = \cos\psi. \tag{3.9}$$

Since

$$\forall Q \in B \quad s = \frac{r_Q}{r_P} \leq \frac{\overline{R}}{r_P} < 1, \tag{3.10}$$

the above function is regular and even analytic in s because

$$1 + s^2 - 2st > 0,\ \forall t\ (|t| \leq 1)$$

when (3.10) is satisfied, and we can develop it into a power series in s, which is uniformly convergent for $Q \in B$,

$$\frac{1}{\ell_{PQ}} = \frac{1}{r_P}\sum_{n=0}^{+\infty} s^n P_n(t) = \sum_{n=0}^{+\infty} \frac{r_Q^n}{r_P^{n+1}} P_n(\cos\psi). \tag{3.11}$$

The functions $P_n(t)$ turn out to be polynomials in t and are called *Legendre polynomials*; they will be studied in detail in the next section.

If we substitute in (3.5) we get

$$T(P) = \sum_{n=0}^{+\infty} \frac{G}{r_P^{n+1}} \int_B r_Q^n P_n(\cos\psi)\delta\rho(Q)dB_Q. \tag{3.12}$$

Since, using a system of spherical coordinates (r, ϑ, λ), we have

$$\cos\psi = \mathbf{e}_P \cdot \mathbf{e}_Q = \frac{\mathbf{r}_P}{r_P} \cdot \frac{\mathbf{r}_Q}{r_Q} \tag{3.13}$$
$$= \sin\vartheta_P \sin\vartheta_Q \cos(\lambda_P - \lambda_Q) + \cos\vartheta_P \cos\vartheta_Q$$

we see that (3.12) gives us a representation of the anomalous potential of the form

$$T(P) = T(r, \vartheta, \lambda) = \sum_{n=0}^{+\infty} \frac{G T_n(\vartheta_P, \lambda_P)}{r_P^{n+1}} \tag{3.14}$$

where

$$T_n(\vartheta_P, \lambda_P) = \int_B r_Q^n P_n(\cos\psi)\delta\rho(Q)dB_Q \tag{3.15}$$

and $\cos\psi$ is taken from (3.13).

The series (3.14) converges uniformly with respect to (ϑ_P, λ_P) outside any sphere with radius larger than $\overline{R}$.

3.3 Legendre Functions

In this section we want to study the functions $P_n(t)$ and draw some conclusions from the representation (3.14) (see also Heiskanen and Moritz (1967), Chap. 1, Krarup (2006), Chap. 13). We start form the definition of $P_n(t)$ as coefficients of the Taylor series of the function

$$G(s,t) = \frac{1}{\sqrt{1 + s^2 - 2st}}, \tag{3.16}$$

also called the *generating function* of Legendre polynomials.

So we have

$$G(s,t) = \sum_{n=o}^{+\infty} s^n P_n(t) \tag{3.17}$$

the series being convergent in the interval

$$0 \leq s < 1. \tag{3.18}$$

Note that in the end we want to substitute $t = \cos\psi$, so we can restrict ourselves to study $P_n(t)$ in the interval

$$-1 \leq t \leq 1 \tag{3.19}$$

corresponding to

$$0 \leq \psi \leq \pi. \tag{3.20}$$

Since (3.17) is a Taylor series, we can compute $P_n(t)$ from

$$P_n(t) = \frac{1}{n!} D_s^n G(s,t)|_{s=0}. \tag{3.21}$$

In this way for instance we can get

$$P_0 \equiv 1,\ P_1(t) \equiv t,\ P_2(t) = \frac{1}{2}(3t^2 - 1),\ P_3(t) = \frac{1}{2}(5t^3 - 3t), \tag{3.22}$$

suggesting that $P_n(t)$ are polynomials of degree n, with the same parity as n, i.e. even for n even and odd for n odd. We shall soon see that this is the case, however we will need a more handy tool than formula (3.21). In fact consider that $G(s,t)$ satisfies identically the relation

$$(1 + s^2 - 2st) D_s G(s,t) = (t - s) G(s,t). \tag{3.23}$$

If we insert the series (3.17) into (3.109) and equate the coefficients of the same powers in s, we find the remarkable recursive relation

$$(n+1) P_{n+1}(t) = (2n+1) t P_n(t) - n P_{n-1}(t)\ ; \tag{3.24}$$

since we already know that $P_0 \equiv 1$, $P_1 \equiv t$, (3.24) allows the direct computation of $P_n(t)$ for any t.

Furthermore, not only (3.24) provides us with a rule for a very fast computation of P_n up to n equal to several thousands, but also gives us the possibility of better understanding the nature of $P_n(t)$.

First of all we now see that if P_{n-1}, P_n are polynomials of degree $n-1$ and n respectively, then P_{n+1} is a polynomial of degree $n+1$; furthermore, if P_{n-1} has a certain parity and P_n the opposite parity, then P_{n+1} has the same parity as P_{n-1}.

Since this is true for $n = 0$ and $n = 1$, we see that the conclusion holds $\forall n$.

Moreover, by taking $t = \pm 1$, (i.e. $\psi = 0$ or π) in (3.16) and (3.17), we find

$$\sum_{n=0}^{+\infty} s^n P_n(\pm 1) = \frac{1}{(1 \pm s)} = \sum_{n=0}^{+\infty} (\mp s)^n\ ; \tag{3.25}$$

(3.25) has to be an identity in s, so we have proved that

$$P_n(1) = 1,\ P_n(-1) = (-1)^n. \tag{3.26}$$

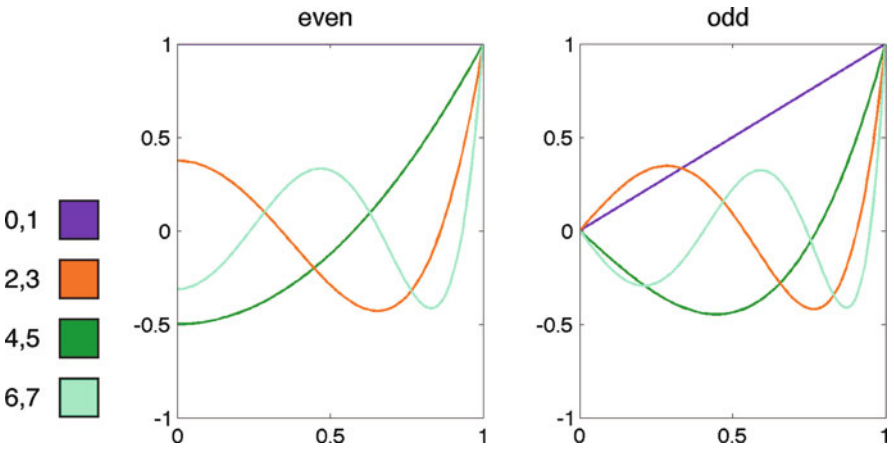


Fig. 3.2 Plot of the Legendre polynomials up to degree 7

Another important property of $P_n(t)$ we mention, namely that

$$|P_n(t)| \leq 1, \ \forall t \in [-1, 1]. \tag{3.27}$$

In fact, reversing the above reasoning we see that if for some $\bar{t}$ one has $P_n(\bar{t}) = \pm 1$ then, $\forall s < 1$

$$\frac{1}{\sqrt{s^2 - 2s\bar{t} + 1}} \equiv \sum_{n=0}^{+\infty} (\pm 1)^n s^n \equiv \frac{1}{1 \mp s}$$

implying that $\bar{t} = \pm 1$. Since $P_n(0) = 0$ when n is odd and, using (3.24), $|P_n(0)| < 1$ when n is even and since $P_n(t)$ cannot cross the barrier ± 1, as explained above, the relation (3.27) has to hold

The interested reader can find more proofs in Szegö (1948). A quick look at the plot of the first Legendre polynomials will help us in viewing their properties. In particular, note the oscillating behaviour of P_n, far from $t = \pm 1$, and for larger values of n (Fig. 3.2).

We turn now to study the differential features of the functions $P_n(t)$, first of all establishing that they are solutions of the Legendre differential equation.

We start from (3.12) and we recall that, whatever is $\delta\rho$ in B, $T(P)$ is certainly harmonic for $r > \overline{R}$. By recalling (1.99) and noting that

$$\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r} \right) \frac{1}{r^{n+1}} = n(n+1) \frac{1}{r^{n+3}},$$

by applying

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\Delta_\sigma$$

to (3.12), (see (1.101), (1.102)), we find

$$\begin{aligned}\Delta T(P) &= \Sigma \frac{G}{r_P^{n+3}} \int_B [n(n+1)P_n(\cos\psi) + \Delta_\sigma P_n(\cos\psi)] r_Q^n \delta\rho(Q) dB \\ &= 0 \end{aligned} \tag{3.28}$$

Since (3.28) has to be true $\forall r_P > \overline{R}$, we find that all the integrals in there have to be zero. Since

$$\int_B [n(n+1)P_n(\cos\psi) + \Delta_\sigma P_n(\cos\psi)] r_Q^n \delta\rho(Q) dB_Q = 0 \tag{3.29}$$

has to hold whatever is $\delta\rho$, we may conclude that

$$\forall n,\ \Delta_\sigma P_n(\cos\psi) + n(n+1)P_n(\cos\psi) \equiv 0. \tag{3.30}$$

On the other hand $\cos\psi$ is given by (3.13) and Q in (3.30) is an arbitrary point of B. So if we choose the unit vector $\mathbf{e}_Q$ along the Z axis we have

$$\cos\psi = \cos\vartheta_P,$$

and, using (1.102) and (3.30), becomes

$$\left(\frac{\partial^2}{\partial\vartheta^2} + \mathrm{ctg}\vartheta\frac{\partial}{\partial\vartheta}\right) P_n(\cos\vartheta) + n(n+1)P_n(\cos\vartheta) \equiv 0. \tag{3.31}$$

If we put

$$t = \cos\vartheta$$

in (3.31) and note that

$$\mathrm{ctg}\vartheta\frac{\partial}{\partial\vartheta} = \mathrm{ctg}\vartheta(-\sin\vartheta)\frac{d}{dt} = -t\frac{d}{dt}, \quad \frac{\partial^2}{\partial\vartheta^2} = -t\frac{d}{dt} + (1-t^2)\frac{d^2}{dt^2}$$

we receive

$$(1-t^2)\frac{d^2}{dt^2}P_n(t) - 2t\frac{d}{dt}P_n(t) + n(n+1)P_n(t) = 0, \tag{3.32}$$

which is well-known in literature as the *Legendre equation*. So we can say that $P_n(t)$ is the solution of (3.32), satisfying the boundary conditions

$$P_n(-1) = (-1)^n, P_n(1) = 1. \tag{3.33}$$

We note also that (3.32) can be written in the more concise and mathematically convenient form

$$\frac{d}{dt}(1-t^2)\frac{d}{dt}P_n(t) + n(n+1)P_n(t) = 0. \tag{3.34}$$

It is not difficult to show (see also Part III, Sect. 13.6) that if a polynomial of degree n is a solution of (3.32), then its coefficients are fixed up to a multiplicative constant.

This constant can always be chosen in such a way that the second of (3.33) is satisfied. Then $P_n(t)$ turns out to have the same parity as n, so that the first of (3.33) is automatically satisfied.

So polynomial solutions of (3.32), with conditions (3.33), are fixed and unique. In Part III, Exercise 9, it is proved that

$$P_n(t) = \frac{1}{2^n n!} D_t^n (t^2-1)^n \tag{3.35}$$

is a solution of (3.32); this is quite clearly a polynomial of degree n. In fact it is proved there that $\{P_n(t)\}$ do coincide with our Legendre polynomials, which are defined in a different way. So (3.35), known in literature as *Rodrigues formula*, becomes an alternative expression for $P_n(t)$.

Another recursive relation, particularly useful to compute first derivatives of $P_n(t)$, is derived from the identity

$$G(s,t) + 2sD_sG(s,t) = \frac{1-s^2}{s} D_t G(s,t)\ ; \tag{3.36}$$

in fact substituting (3.17) end equating the coefficients of the same powers of s we obtain

$$P'_{n+1}(t) = P'_{n-1}(t) + (2n+1)P_n(t). \tag{3.37}$$

A combination of (3.37), multiplied by n, with (3.24), differentiated, provides another useful relation, i.e.

$$P'_{n+1} = tP'_n + (n+1)P_n. \tag{3.38}$$

Let us stress that (3.24) together with (3.38) and (3.32) provides us with a powerful tool to compute sequentially $P_n(t)$, $P'_n(t)$, $P''_n(t)$ for all n up to any fixed high degree N.

We can turn now to study the integral properties of $P_n(t)$, which will be of fundamental importance in the sequel.

Such properties can be summarized in the formula

$$(\ell + n + 1)\frac{1}{4\pi}\int P_\ell(\cos\psi_{P_0Q})P_n(\cos\psi_{PQ})d\sigma_Q = P_n(\cos\psi_{P_0P})\delta_{\ell n}. \quad (3.39)$$

The proof can be found in Sect. A.1.
Formula (3.39) allows to draw three conclusions:

(a) Any two Legendre functions $P_n(\cos\psi_{P_0Q})$, $P_\ell(\cos\psi_{PQ})$ with different degrees $(\ell \neq n)$, are orthogonal in $L^2(S_1)$ whatever are the directions of $\mathbf{e}_{P_0}$ and $\mathbf{e}_P$,

$$\frac{1}{4\pi}\int P_\ell(\cos\psi_{P_0Q})P_n(\cos\psi_{PQ})d\sigma_Q = 0,\ \ell \neq n \quad (3.40)$$

(b) Letting $\mathbf{e}_{P_0} = \mathbf{e}_P$ in (3.39), i.e. $\cos\psi_{P_0P} = 1$, and $\ell = n$, we find

$$\frac{1}{4\pi}\int P_n^2(\cos\psi_{P_0Q})d\sigma_Q = \frac{1}{2n+1} \quad (3.41)$$

(c) When $\mathbf{e}_{P_0} \neq \mathbf{e}_P$, $\ell = n$, we find the *reproducing* formula

$$P_n(\cos\psi_{P_0P}) = \frac{2n+1}{4\pi}\int P_n(\cos\psi_{P_0Q})P_n(\cos\psi_{PQ})d\sigma_Q. \quad (3.42)$$

Formula (3.42) is essential for the analysis in Part III.

Remark 1. Take $\mathbf{e}_{P_0} = \mathbf{e}_P = \mathbf{e}_z$ in (3.39); then, noting that $d\sigma = \sin\vartheta d\vartheta d\lambda = -dtd\lambda$, we find

$$\frac{1}{4\pi}\int_0^{2\pi} d\lambda \int_{-1}^{1} P_\ell(t)P_n(t)dt = \frac{1}{2}\int_{-1}^{1} P_\ell(t)P_n(t)dt = \frac{\delta_{\ell n}}{2n+1}. \quad (3.43)$$

This equation shows that the sequence of polynomials $\{P_n(t)\}$ is orthogonal in $L^2[(-1,1])$ and furthermore

$$\|P_n(t)\|^2_{L^2([-1,1])} = \int_{-1}^{1} P_n^2(t)dt = \frac{2}{2n+1}. \quad (3.44)$$

Even more, although we won't make so much use of Legendre polynomials in one dimension, we have to note that $\{P_n(t)\}$ is a complete sequence in $L^2([-1,1])$. In fact, note that one has $1 = P_0, t = P_1, t^2 = \frac{1}{3}(2P_2 + P_0), t^3 = \frac{1}{5}(2P_3(t) + 3P_1)$ and so forth; then t^k can be expressed for every k as a linear combination of $\{P_n(t)\}$ and the same will be true for any polynomial in t. On the other hand a famous theorem by Weiestrass (cf. Riesz and Nagy 1965; Yosida 1978) claims that any continuous function $f_c(t)$ can be uniformly approximated on any bounded interval by

a suitable polynomial $Q_N(t)$. Since $\{\int_{-1}^{1}[f_c(t)-Q_N(t)]^2dt\}^{1/2} \leq \sqrt{2}\varepsilon$ if $|f_c(t)-Q_N(t)|<\varepsilon$, $f_c(t)$ is arbitrarily well-approximated in $L^2([-1,1])$ by $Q_N(t)$.

On the other hand any $f(t) \in L^2([-1,1])$ can be approximated as well as we like by a suitable continuous function $f_c(t)$, so that we have

$$\begin{aligned}\|f(t)-Q_N(t)\| &\leq \|f(t)-f_c(t)\| + \|f_c(t)-Q_N(t)\| \\ &\leq \varepsilon + \sqrt{2}\varepsilon = (1+\sqrt{2})\varepsilon,\end{aligned} \tag{3.45}$$

i.e. the space of polynomials in t is everywhere dense in $L^2([-1,1])$, very much like the space of rational numbers is dense in that of real numbers. Since $Q_N(t)$ can be expressed as a linear combination of $P_0, P_1 \ldots P_N$, we see that $\{P_n(t)\}$ is a complete orthogonal basis in $L^2([-1,1])$ and the following representation

$$f(t) = \sum_{n=0}^{+\infty} \frac{(2n+1)}{2} P_n(t) \left(\int_{-1}^{1} f(t') P_n(t') dt' \right) \tag{3.46}$$

holds for any square integrable $f(t)$ (see Part III, Definition 19 and Proposition 10 or Riesz and Nagy (1965), Yosida (1978)).

Remark 2. Let us remark that from $P_0 = 1$, $P_1 = \cos\psi_{PQ} = \mathbf{e}_P \cdot \mathbf{e}_Q$, we can write from (3.12)

$$\begin{aligned}(r_P > \overline{R}), \qquad T(P) = G \Bigg\{ &\frac{1}{r_P} \int_B \delta\rho(Q) dB_Q \\ &+ \frac{\mathbf{e}_P}{{r_P}^2} \cdot \int_B \delta\rho(Q) r_Q \mathbf{e}_Q dB_Q + O\left(\frac{1}{r_P^3}\right) \Bigg\}.\end{aligned} \tag{3.47}$$

Therefore if the normal field is made in such a way that $\int_{B_e} \rho_e(Q) dB_Q \equiv M$, i.e. the mass generating U is the same as that generating W, and if in addition we are using a geocentric system, such that

$$\int_B \delta\rho(Q)\mathbf{r}_Q dB_Q = \int_B \rho(Q)\mathbf{r}_Q dB_0 - \int_{B_e} \rho_e(Q)\mathbf{r}_Q dB_Q = 0, \tag{3.48}$$

then we have indeed

$$T(P) = O\left(\frac{1}{r_P^3}\right) \tag{3.49}$$

as we have anticipated in (1.131).

3.4 Spherical Harmonics

Consider the following family of functions of (ϑ, λ), i.e. defined on the unit sphere S_1, depending on two indexes (n, m):

$$Y_{nm}(\vartheta, \lambda) = \overline{P}_{nm}(\vartheta) f_m(\lambda), \tag{3.50}$$

$$n = 0, 1, 2 \ldots, \; m = -n \ldots, 0, \ldots n$$

$$f_m(\lambda) = \begin{cases} \cos m\lambda & m = 0, 1 \ldots n \\ \sin |m|\lambda & m = -n, \ldots -1 \end{cases} \tag{3.51}$$

$$\overline{P}_{nm}(\vartheta) = \sqrt{(2 - \delta_{mo})(2n + 1)\frac{(n - m)!}{(n + m)!}} P_{nm}(\vartheta) \tag{3.52}$$

$$P_{nm}(\vartheta) = (1 - t^2)^{m/2} D_t^m P_n(t) \tag{3.53}$$

$$t = \cos \vartheta$$

By definition these are called *surface spherical harmonics* of degree n and order m; $P_{nm}(\vartheta)$ are called *associated Legendre functions of the first kind*, $\overline{P}_{nm}(\vartheta)$ *normalized associated Legendre functions*.

This sequence and its relation to functions harmonic in space is studied in depth in Part III, Chap. 13; in this section we limit ourselves to recall some results highlighting the possibility of constructing, by means of linear combinations, useful approximate models of the anomalous potential, usually called *global models*.

We start by stating a famous theorem, the proof of which can be found in Part III, Sect. 13.2, Theorem 2.

Theorem 1 (Summation theorem). *The following identity holds*

$$P_n(\cos \psi_{PQ}) = \frac{1}{2n + 1} \sum_{m=-n}^{n} Y_{nm}(\vartheta_P, \lambda_P) Y_{nm}(\vartheta_Q, \lambda_Q). \tag{3.54}$$

To understand the relevance of this theorem to our matters, let us substitute (3.54) into (3.14) and (3.15) and rearrange; we obtain

$$T(P) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \widetilde{T}_{nm} \frac{Y_{nm}(\vartheta_P, \lambda_P)}{r_P^{n+1}} \tag{3.55}$$

$$\widetilde{T}_{nm} = \frac{G}{(2n + 1)} \int_B r_Q^n Y_{nm}(\vartheta_Q, \lambda_Q) \delta\rho(Q) dB_Q; \tag{3.56}$$

the series (3.55), as we know, is convergent for $r_P > \overline{R}$. From (3.56) we see that the numerical coefficients $\widetilde{T}_{nm}$ are different in dimension for every degree n; to avoid

this ugly characteristic it is customary to modify (3.55) in such a way as to express T by means of non-dimensional coefficients.

Namely we put

$$T(P) = \frac{GM}{R}\sum_{n=0}^{+\infty}\sum_{m=-n}^{n} T_{nm}\left(\frac{R}{r_P}\right)^{n+1} Y_{nm}(\vartheta_P, \lambda_P) \tag{3.57}$$

$$T_{nm} = \frac{1}{2n+1}\frac{1}{M}\int_B \left(\frac{r_Q}{R}\right)^n Y_{nm}(\vartheta_Q, \lambda_Q)\delta\rho(Q)dB_Q, \tag{3.58}$$

where R can be any radius related to the earth; common is the choice

$$R = 6{,}371\,\text{km}, \tag{3.59}$$

namely the mean radius of the earth ellipsoid. Such a value is indeed very close, but not strictly equal to the Brillouin radius.

If we go back to (1.16) we see that our typical choice of the normal potential U and of the relative position of the earth ellipsoid to the masses, implies

$$T_{00} = 0, \quad T_{1,m} = 0 \quad (m = -1, 0, 1) \tag{3.60}$$

so that the series (3.57) in fact starts from the degree $n = 2$ and (3.49) is always satisfied. We also note that with this definition we can count on the estimate $O(T_{nm}) \sim 10^{-5}$ or smaller. We shall see later on how to make this estimate tighter.

We notice now that since $T(P)$ has to be a harmonic function whatever are the numerical coefficients $\{T_{nm}\}$ in (3.57), we must also have

$$r > R, \qquad \Delta\left[\left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta, \lambda)\right] = 0. \tag{3.61}$$

If we use (1.100) and (1.102), i.e. the spherical representation of the Laplacian, and we take into account the relations

$$\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right)\left(\frac{1}{r^{n+1}}\right) = n(n+1)\frac{1}{r^{n+3}} \tag{3.62}$$

$$\frac{\partial^2}{\partial\lambda^2} Y_{nm}(\vartheta, \lambda) = -m^2 Y_{nm}(\vartheta, \lambda), \tag{3.63}$$

we find that the associated Legendre functions have to satisfy the Legendre equation of order m (cf. Part III, Remark 2)

$$(1-t^2)P''_{nm}(t) - 2tP'_{nm}(t) + \left[n(n+1) - \frac{m^2}{1-t^2}\right]P_{nm}(t) = 0 \tag{3.64}$$

We note that if we put $m = 0$ in (3.64) we retrieve the simple Legendre equation (3.32); this is consistent with the fact that if we put $m = 0$ into (3.52) and (3.53) we find

$$\overline{P}_{n0}(t) = \sqrt{2n+1}P_n(t). \tag{3.65}$$

We note explicitly as well that in this way $\{\overline{P}_{n0}(t)\}$ are not L^2 normalized on the interval $[-1, 1]$ (compare (3.44)), but are indeed L^2 normalized on the unit sphere.

Another remark which is an immediate consequence of (3.62) and of the formula

$$\Delta = \left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\Delta_\sigma,$$

is that the important relation holds

$$\Delta_\sigma Y_{nm}(\vartheta, \lambda) = -n(n+1)Y_{nm}(\vartheta, \lambda), \tag{3.66}$$

i.e. $\{Y_{nm}(\vartheta, \lambda)\}$ are eigenfunction of the Laplace-Beltrami operator Δ_σ. More precisely, considering (3.63) too, we can claim that $Y_{nm}(\vartheta, \lambda)$ is an eigenfunction of Δ_σ, with eigenvalue $-n(n+1)$, and an eigenfunction of $\frac{\partial^2}{\partial\lambda^2}$, with eigenvalue $-m^2$. This fact will be exploited later on.

The functions

$$S_{nm}(r, \vartheta, \lambda) = \left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta, \lambda)$$

are usually called exterior solid spherical harmonics. The adjective exterior refers to the fact that they are harmonic outside the origin up to infinity, as opposed to the functions $r^n Y_{nm}(\vartheta, \lambda)$, used in most of Part III, Chap. 13, which are harmonic in the whole space, but not bounded at infinity.

Very much like the Legendre polynomials, also the functions $\overline{P}_{nm}(t)$ can be sequentially computed by means of recursive relations.

There are two principal types of such relations, one on the degree n, the other one on the order m; these are

$$\overline{P}_{n+1,m}(t) = A_{nm}t\overline{P}_{nm}(t) - B_{nm}\overline{P}_{n-1,m}(t) \tag{3.67}$$

with

$$A_{nm} = \left[\frac{(2n+1)(2n+3)}{(n+1-m)(n+1+m)}\right]^{(1/2)}$$

$$B_{nm} = \left[\frac{(2n+3)(n+m)(n-m)}{(2n-1)(n+1-m)(n+1+m)}\right]^{(1/2)},$$

and

$$\overline{P}_{n,m+1}(t) = \frac{2t}{\sqrt{1-t^2}} m C_{nm} \overline{P}_{nm}(t) - C_{nm} D_{nm} \overline{P}_{n,m-1}(t). \tag{3.68}$$

with

$$C_{nm} = \left[\frac{1}{(n-m)(n-m+1)} \right]^{(1/2)}$$
$$D_{nm} = [(n+m)(n-m+1)]^{(1/2)} \sqrt{1+\delta_{m1}}.$$

The relations (3.67) are triggered by

$$\overline{P}_{m-1,m}(t) \equiv 0, \ \overline{P}_{mm}(t) = \sqrt{\frac{2(2m+1)}{(2m)!}} (1-t^2)^{(m/2)}. \tag{3.69}$$

while (3.68) are triggered by

$$\overline{P}_{n0}(t) = \sqrt{2n+1} P_n(t), \ \overline{P}_{n1} = \sqrt{\frac{2(2n+1)}{n(n+1)}} (1-t^2)^{(1/2)} P'_n(t), \tag{3.70}$$

where $P_n(t)$, $P'_n(t)$ are computed according to (3.24) and (3.38).

In Part III, Proposition 7 and the following, there are proofs of such relations as well as a discussion on their numerical implementation. At present with degrees up to some thousands and all orders, the best is to compute $\overline{P}_{mm}$ suitably rescaled and use them in (3.67) and (3.69), dividing at the end the result by the scale factor. Such scale factor can be very large, however being computed separately in exponential form, does not destroy significant digits in the process of the numerical computation.

By differentiating (3.67) one gets a recursive relation, useful for the computation of the derivatives P'_{nm}; in fact

$$\overline{P}'_{n+1,m}(t) = A_{nm} \overline{P}_{nm}(t) + A_{nm} t \overline{P}'_{nm}(t) - B_{nm} \overline{P}'_{n-1,m}(t), \tag{3.71}$$

and

$$\overline{P}'_{m-1,m}(t) \equiv 0, \ \overline{P}'_{mm}(t) = \sqrt{\frac{2(2m+1)}{(2m)!}} (-m) t (1-t^2)^{(m/2)-1}. \tag{3.72}$$

When the second derivatives $\overline{P}''_{nm}(t)$ are needed, one can directly use the Legendre equation (3.64).

An alternative to the recursive evaluation of individual Legendre function is the so-called *Clenshow summation method* that one can find in Tscherning and Poder (1981).

We come now to establish important functional properties of $\{Y_{nm}(\vartheta,\lambda)\}$.
The first, known as *orthogonality relation*, is given by

$$\frac{1}{4\pi}\int_{S_1} Y_{nm}(\vartheta,\lambda)Y_{\ell k}(\vartheta,\lambda)d\sigma = \delta_{\ell n}\delta_{mk}. \tag{3.73}$$

In fact (3.73) says that $\{Y_{nm}(\vartheta,\lambda)\}$ is an orthonormal system in $L^2(S_1)$. Note that in this $L^2(S_1)$ scalar product the factor 4π, which is a simple normalization factor, is conventional and introduced to simplify formulas. Moreover the theory developed in Part III, Chap. 13 leads to a fundamental property, which we state in the form of theorem (see Part III, Definition 19 and Theorem 3).

Theorem 2 (Completeness of $\{\mathbf{Y}_{nm}\}$ in $\mathbf{L^2(S_1)}$). *The sequence $\{Y_{nm}(\vartheta,\lambda)\}$ is a complete orthonormal system in $L^2(S_1)$.*

That $\{Y_{nm}\}$ is orthonormal has already been expressed by (3.73); that it is complete means that for every $f(\vartheta,\lambda)$ square integrable on S_1 we have the following Fourier representation

$$\begin{aligned} f(\vartheta,\lambda) &= \sum_{n=0}^{+\infty}\sum_{m=-n}^{n} f_{nm}Y_{nm}(\vartheta,\lambda) \\ f_{nm} &= \langle f, Y_{nm}\rangle_{L^2(S_1)} = \frac{1}{4\pi}\int_{S_1} f(\vartheta,\lambda)Y_{nm}(\vartheta,\lambda)d\sigma. \end{aligned} \tag{3.74}$$

The series in (3.74) is convergent in the sense of $L^2(S_1)$ and the following Perseval's identity holds (cf. Part III, Remark 4).

$$\|f\|^2_{L^2(S_1)} = \frac{1}{4\pi}\int_{S_1} f^2(\vartheta,\lambda)d\sigma = \sum_{n=0}^{+\infty}\sum_{m=-n}^{n} f^2_{nm}. \tag{3.75}$$

Remark 3. If we define a Hilbert space of harmonic functions

$$HL^2(S_R) \equiv \{u\,;\ \Delta u = 0, r > R\,;\ \int_{S_1} u^2 dS < +\infty\} \tag{3.76}$$

and we consider the series

$$\begin{aligned} u(r,\vartheta,\lambda) &= \sum_{n=0}^{+\infty}\sum_{m=-n}^{n} u_{nm}\left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta,\lambda) \\ &= \sum_{n=0}^{+\infty}\sum_{m=-n}^{n} u_{nm}S_{nm}(r,\vartheta,\lambda), \end{aligned} \tag{3.77}$$

which represents a typical element of $HL^2(S_R)$, fixing $r = R$ we see that

$$u(R, \vartheta, \lambda) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} u_{nm} Y_{nm}(\vartheta, \lambda) \; ; \tag{3.78}$$

on the other hand u_{nm} are then determined by (3.74), so that we can say that each function $u(r, \vartheta, \lambda)$ in $HL^2(S_R)$ is in one-to-one correspondence with its trace on $S_R, u(R, \vartheta, \lambda)$. In particular both functions have the same sequence of coefficients $\{u_{nm}\}$ and such coefficients have to satisfy the condition

$$\sum_{n=0}^{+\infty} \sum_{m=-n}^{n} u_{nm}^2 < +\infty$$

as otherwise u cannot belong to $HL^2(S_R)$.

With this identification of functions between $u \in HL^2(S_R) \Leftrightarrow u|_{S_R} \in L^2(S_1)$ implying also that

$$S_{\ell m} \in HL^2(S_R) \Leftrightarrow S_{\ell m}|_{S_R} \equiv Y_{\ell m} \in L^2(S_1),$$

we see that $\{S_{\ell m}\}$ can be considered as a complete orthonormal system in $HL^2(S_R)$. This allows the definition of easy rules for a calculus with harmonic functions in spherical domains.

Remark 4. Let us consider a surface S satisfying some regularity condition, such as the continuity of the normal vector $\mathbf{n}(P)$ (see Part III, Sect. 13.2) and such that the origin O is within the body B enclosed by S. If we call Ω the exterior domain, we can consider $HL^2(S)$ i.e. the Hilbert space of functions which are harmonic in Ω and on S are square integrable, i.e.

$$HL^2(S) \equiv \{u \; ; \; \Delta u = 0 \text{ in } \Omega \; ; \; \int_S u^2 dS < +\infty\}.$$

Note that equivalent norms, like those discussed in Sect. 2.2, could also be used here. In such a space we define the scalar product as $u, v \in HL^2(S)$

$$< u, v >= \int_S u(P) v(P) dS_P. \tag{3.79}$$

Now, it is clear that $\{S_{nm}(r, \vartheta, \lambda)\} \in HL^2(S)$ and we can consider the linear subspace

$$\text{Span}\{S_{nm}\} \equiv \left\{ u_N = \sum_{n=0}^{N} \sum_{m=-n}^{n} \lambda_{nm} S_{nm} \; ; \; \forall N, \forall \lambda_{nm} \in \mathcal{R} \right\} . \tag{3.80}$$

We note that under these conditions there will always be a sphere B_0 centered at O and such that $B_0 \subset B$. Indeed it is enough to take a sphere with radius R_B satisfying

$$R_B \leq \min_{P \in S} r_P.$$

A sphere like this is called in literature a *Bjerhammar sphere* and R_B a *Bjerhammar radius*. Here, when we use a set of solid spherical harmonics S_{nm}, we assume that the R used in their definition (cf. (3.57)) is equal to R_B.

We note too that, due to the non-spherical shape of S, in general $\{S_{nm}(r, \vartheta, \lambda)|_S\}$ is not any more an orthonormal sequence in $L^2(S)$, i.e. it is not orthonormal in $HL^2(S)$.

Nevertheless the property of completeness still holds true or, said in another way, Span$\{S_{nm}\}$ is dense in $HL^2(S)$.

Theorem 3 (Completeness of $\{\mathbf{S}_{nm}|_{\mathbf{S}}\}$ in $\mathbf{L^2(S)}$). *Let S be a surface satisfying a condition of continuity of the normal $\mathbf{n}(P)$ and smoothly mapped to the unit sphere, for instance a star-shaped surface; then $\{S_{nm}|_S\}$ is a complete sequence in $L^2(S)$. Accordingly $\{S_{nm}\}$ is a complete sequence in $HL^2(S)$.*

The proof of this theorem can be found in Part III, Sect. 13.4 under Theorem 5. The meaning of the statement of the theorem is precisely that, given any function $u(r, \vartheta, \lambda) \in HL^2(S)$ and any $\varepsilon > 0$, there are an integer N and constants $\{u_{nm}\,;\ |m| \leq n, n \leq N\}$ such that

$$\begin{aligned} &\|u - \sum_{n=0}^{N} \sum_{m=-n}^{n} u_{nm} S_{nm}(r, \vartheta, \lambda)\|_{HL^2(S)} \\ &= \left\{ \int_S [u(P) - \sum_{n=0}^{N} \sum_{m=-n}^{n} u_{nm} S_{nm}(P)]^2 dS_P \right\}^{\frac{1}{2}} \leq \varepsilon. \end{aligned} \tag{3.81}$$

This theorem constitutes the theoretical basis for the construction of global models of the anomalous potential T, as we shall see later on, in this chapter.

Remark 4, and Part III, Sect. 13.4, Theorem 5 recalled here, suggest that by providing the values of a square integrable function $f(P)$ on S, we could recover by means of a suitable harmonic series a representation of the function $u(P)$ which is harmonic in Ω and agrees with $f(P)$ on S. In other words we are tempted to take the limit for $N \to \infty$ in (3.81) and claim that we find in this way a harmonic series converging in the whole of Ω. This is not true and we shall give in the next section an elementary counterexample.

The reason for this relies on the fact that the coefficients $\{u_{nm}\}$ for which a minimum is attained in formula (3.81) do change when we change N and we should denote them, more carefully, as $\{u_{nm}^N\}$. On the basis of more advanced analyses, like those performed in Sects. 14.4 and 14.5 of Part III, one can claim that, as a matter

of fact, it is possible to take the limit

$$\lim_{N\to\infty} u_{nm}^N = \overline{u}_{nm} \tag{3.82}$$

and the limit coefficients are in fact related to the series $\sum_{n,m} \overline{u}_{nm} S_{nm}(r,\vartheta,\lambda)$, which is convergent for $r > \overline{R}$, with $\overline{R}$ a Brillouin radius. However, the same series is not convergent in general for $r < \overline{R}$ so, while the individual coefficients have the limit (3.82), the other limit

$$\lim_{N\to\infty} \sum_{n=0}^{N} \sum_{m=-n}^{n} u_{nm}^N S_{nm}(P) \equiv \lim_{N\to\infty} u_N(P)$$

in general does not exist when P is on the surface S. Or better, such a sequence $\{u_N\}$ is converging in $L^2(S)$ to $u|_S$, namely not in a pointwise way, but this limit function **cannot** be expressed as a convergent series of the form

$$u(P)\Big|_S = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \overline{u}_{nm} S_{nm}(P)\Big|_S .$$

Nevertheless there are cases in which this ugly phenomenon is not happening, namely when S itself is a sphere, of radius R, so that we can take R_0 to coincide with R. The effect of this choice is that $\{S_{nm}\}$ now becomes orthonormal and, as a consequence, the "best" set of coefficients minimizing the norm (3.81) does not depend anymore on N, so that taking the limit in this formula becomes much easier.

We show three examples of solutions of problems of determining $T(P)$ from boundary values on a sphere. Two of them are as a matter of fact closely related to boundary value problems (BVP's) of great geodetic significance. Yet they should be taken only as examples used to grasp, in a simple situation, the qualitative behaviour of solutions of BVP's: a sounder theory for this argument has to be found in Part III, Chap. 14, where its numerical implementation is discussed too.

Example 1 (Poisson). We assume that S is a sphere of radius R and we put $S_{nm} = \left(\frac{R}{r}\right)^{n+1} Y_{nm}$. We assume that the values of $T(P)$ are given all over S and the corresponding function $f(P)$ is in $L^2(S)$. Then the solution $T(P)$ of the Dirichlet problem

$$\begin{cases} \Delta T = 0 \text{ in } \Omega \\ T = f \quad \text{on } S \\ T \to 0 \quad r \to \infty, \end{cases} \tag{3.83}$$

is given by the Poisson integral, as proved in Part III, Sect. 12, i.e.

$$T(P) = \frac{1}{4\pi} \int \Pi_{Re}(P, Q) f(Q) d\sigma_Q \tag{3.84}$$
$$(d\sigma_Q = \sin \vartheta_Q d\vartheta_Q d\lambda_Q \; ; \; f(Q) = f(\vartheta_Q, \lambda_Q))$$

where

$$\Pi_{Re}(P, Q) = \frac{R(r_P^2 - R^2)}{\{r^2 + R^2 - 2rR\cos\psi_{PQ}\}^{3/2}} = \frac{R(r_P^2 - R^2)}{\ell_{PQ}^3}. \tag{3.85}$$

The index *Re* here means that the Poisson kernel is referred to the solution of an external problem for a sphere of radius R.

Example 2 (Hotine). Assume $S \equiv \{P \; : r_P = R\}$ and that on S we give $\delta g(P)$ (cf. (2.30), (2.99)). In this setup we use the spherical approximation and we define the problem

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ -\frac{\partial T}{\partial r} = \delta g & \text{on } S \\ T \to 0 & r \to \infty, \end{cases} \tag{3.86}$$

which is also known as a *Neumann problem* since we supply on S the derivative of T in the radial direction, which is normal to S in this case.

The explicit solution of (3.86) is given by means of the so-called *Hotine function*

$$H(P, Q) = \frac{2R}{\ell_{PQ}} - \log \frac{\ell_{PQ} + R - r_P \cos\psi_{PQ}}{r_P(1 - \cos\psi_{PQ})}, \tag{3.87}$$

by the integral relation

$$T(P) = \frac{R}{4\pi} \int H(P, Q) \delta g(Q) d\sigma_Q. \tag{3.88}$$

This is obtained as follows. Put

$$T = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} T_{nm} \left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta, \lambda) \; ; \tag{3.89}$$

then the first and last of (3.86) are satisfied. We can try to satisfy the second of (3.86), i.e.

$$-\left.\frac{\partial T}{\partial r}\right|_{r=R} = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \frac{n+1}{R} T_{nm} Y_{nm}(\vartheta, \lambda) = \delta g(\vartheta, \lambda). \tag{3.90}$$

By using the orthogonality relations (3.73) we derive

$$\frac{n+1}{R}T_{nm} = \frac{1}{4\pi}\int \delta g(\vartheta,\lambda)Y_{nm}(\vartheta,\lambda)d\sigma, \tag{3.91}$$

which, substituted back into (3.89), gives

$$\frac{1}{4\pi}\int\left[\sum_{n=0}^{+\infty}\sum_{m=-n}^{n}\frac{R}{n+1}\left(\frac{R}{r_P}\right)^{n+1}Y_{nm}(\vartheta_P,\lambda_P)Y_{nm}(\vartheta,\lambda)\right]\delta g(\vartheta,\lambda)d\sigma$$

$$= \frac{1}{4\pi}\int R\left[\sum_{n=0}^{+\infty}\frac{2n+1}{n+1}\left(\frac{R}{r_P}\right)^{n+1}P_n(\cos\psi_{PQ})\right]\delta g(Q)d\sigma_Q \tag{3.92}$$

The series in parenthesis is then added by splitting it a according to

$$\sum_{n=0}^{+\infty}\frac{2n+1}{n+1}s^{n+1}P_n(t) = 2\sum_{n=0}^{+\infty}s^{n+1}P_n(t) - \sum_{n=0}^{+\infty}\frac{s^{n+1}}{n+1}P_n(t) = H_1(s,t) - H_2(s,t).$$

Then recalling the definitions of generating functions (3.16) and (3.17) we find

$$H_1(s,t) = 2sG(s,t)\ ;\ \frac{\partial}{\partial s}H_2(s,t) = G(s,t). \tag{3.93}$$

Integrating the second of (3.93) between 0 and s, taking into account that $H_2(0,t) = 0$, and substituting back we get (3.87).

Example 3 (Stokes). In this case we assume to give on S the function $\Delta g(P)$ that we express in spherical approximation as in (2.100). So our problem is now

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ -\frac{\partial T}{\partial r} - \frac{2}{r}T = \Delta g(P) & \text{on } S \\ T \to 0 & r \to \infty \end{cases} \tag{3.94}$$

If we use the representation (3.89) for T, we find

$$-\frac{\partial T}{\partial r} - \frac{2}{r}T\bigg|_{r=R} = \sum_{n=0}^{+\infty}\sum_{m=-n}^{n}\frac{n-1}{R}T_{nm}Y_{nm}(\vartheta,\lambda) = \Delta g(\vartheta,\lambda). \tag{3.95}$$

The use of orthogonality relations now gives us

$$\frac{n-1}{R}T_{nm} = \frac{1}{4\pi}\int \Delta g(\vartheta,\lambda)Y_{nm}(\vartheta,\lambda)d\sigma. \tag{3.96}$$

Equation (3.95) tells us two things: if we put $n = 1$ in it we see that T_{1m} are not determined, but the equation can be satisfied only if

$$\Delta g_{1m} \equiv 0. \tag{3.97}$$

This means that if we give on S a function that does not satisfy (3.97), this is not in reality a gravity anomaly because all gravity anomalies do fulfil such a relation. In addition, if Δg has been generated from a normal potential with the same mass content as W, then we know in advance that $T_{00} = 0$. Furthermore, if the barycenter is placed at the origin we know in advance that $T_{1m} = 0$ (cf. (3.60)).

Summarizing, substituting back into (3.89) and using the summation theorem (3.54), we get

$$T(P) = \frac{1}{4\pi} \int R \left[\sum_{n=2}^{+\infty} \frac{2n+1}{n-1} \left(\frac{R}{r} \right)^{n+1} P_n(\cos \psi_{PQ}) \right] \Delta g(Q) d\sigma_Q$$
$$= \frac{R}{4\pi} \int S(P, Q) \Delta g(Q) d\sigma_Q \,. \tag{3.98}$$

Again the series in parenthesis can be split as

$$\sum_{n=2}^{+\infty} \frac{2n+1}{n-1} s^{n+1} P_n(t) = 2 \sum_{n=2}^{+\infty} s^{n+1} P_n(t) + 3 \sum_{n=2}^{+\infty} \frac{s^{n+1}}{n-1} P_n(t)$$
$$= 2S_1(s,t) + 3S_2(s,t),$$

The series are then added, using the relations

$$S_1 = s[G(s,t) - 1 - st]$$
$$s^2 D_s \left(\frac{1}{s^2} S_2 \right) = G(s,t) - 1 - st,$$

with $G(s,t)$ given by (3.16) and (3.17).

The calculus is laborious and it provides ultimately the Stokes function

$$S(P, Q) = \frac{2R}{\ell_{PQ}} + \frac{R}{r_P} - 3\frac{R\ell_{PQ}}{r_P^2} \tag{3.99}$$
$$- \frac{R^2}{r_P^2} \cos \psi \left[5 + 3 \log \frac{r_P - R \cos \psi_{PQ} + \ell_{PQ}}{2r_P} \right],$$

to be used in the Stokes's integral (3.98) (cf. Heiskanen and Moritz 1967).

Let us remark here that (3.98) provides the sought anomalous potential in the whole outer space $\{r \geq R\}$. In particular we can take P on the sphere itself by putting $r_P = R$ in (3.99). In this way we get the simple spherical Stokes formula yielding T, and hence the geoid undulation N, on the sphere. Noting that with $r_P = R$ we have

$$\frac{\ell_{PQ}}{r_P} = \sqrt{2(1 - \cos \psi)} = 2 \sin \frac{\psi}{2},$$

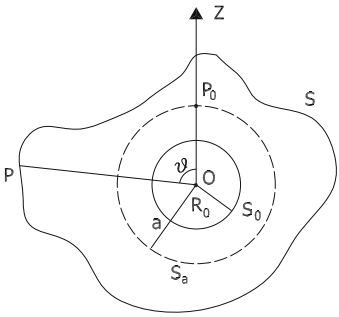


Fig. 3.3 The outlook of the grain of sand example

(3.99) and (3.98) become respectively

$$S(\psi) = \frac{1}{\sin\frac{\psi}{2}} + 1 - 6\sin\frac{\psi}{2} - \cos\psi\left[5 + 3\log\left(\sin\frac{\psi}{2} + \sin^2\frac{\psi}{2}\right)\right], \quad (3.100)$$

$$P \in S_R, \quad T(P) = \frac{R}{4\pi}\int S(\psi_{PQ})\Delta g(Q)d\sigma_Q. \quad (3.101)$$

3.5 Downward Continuation and Krarup's Theorem

Since, unfortunately, the geodetic literature is not exempt from errors on this item, we deem it useful to clarify the fundamental fact that not every function harmonic in Ω and square integrable on S can be continued down to a Bjerhammer sphere S_0 by some potential that is still harmonic in the layer between S_0 and S (cf. Fig. 3.3)

It is enough to prove it by a counterexample which, in spite of its simplicity, should give the reader the idea that it is much easier to find a potential that cannot be continued rather than the opposite. The example is taken from Moritz (1980).

Example 4 (Grain of sand). We refer to Fig. 3.3 and assume that R_0 is any radius such that $S_0 \subset B$. We can find then a number a, which is still a Bjerhammer radius, but $a > R_0$. Then we assume that the potential we want to discuss, is that generated by a "grain of sand" of mass m placed at P_0, $(r_{P_0} = a)$. This potential is

$$T(P) = \frac{Gm}{\ell_{P_0P}} \quad (3.102)$$

which is a bounded regular function on S because $\mathrm{dist}(P_0, S) > 0$.

For the sake of simplicity we define the Z axis so that P_0 belongs to it and so the angle ψ between $\mathbf{e}_{P_0}$ and any other direction is the same as the spherical co-latitude of this direction.

Since $r_P > a$ when $P \in S$, we can develop (3.102) into a convergent series of spherical harmonics, namely

$$T(P) = \frac{Gm}{a}\sum_{n=0}^{+\infty}\left(\frac{a}{r_P}\right)^{n+1} P_n(\cos\psi) \tag{3.103}$$

$$= \frac{Gm}{a}\sum_{n=0}^{+\infty}\left(\frac{a}{r_P}\right)^{n+1} \frac{1}{\sqrt{2n+1}} Y_{n,0}(\vartheta,\lambda)\ ;$$

in (3.103) the relations (3.65) and (3.50) have been used. Now assume that $T(P)$ can be continued down to S_0 and that it is square integrable on this sphere. Then we must have, denoting with $\overline{T}$ the function T continued to S_0,

$$\overline{T}(P) = \sum_{n=0}^{+\infty}\sum_{m=-n}^{n} \overline{T}_{nm}\left(\frac{R_0}{r}\right)^{n+1} Y_{nm}(\vartheta,\lambda). \tag{3.104}$$

Comparing (3.104) with (3.103) and putting $r_P = \overline{R}$, a large Brillouin radius, one gets

$$\overline{T}_{nm}\left(\frac{R_0}{\overline{R}}\right)^{n+1} = \left(\frac{a}{\overline{R}}\right)^{n+1} \frac{1}{\sqrt{2n+1}}\delta_{m0}$$

and then

$$\overline{T}_{nm} = \left(\frac{a}{R_0}\right)^{n+1} \frac{1}{\sqrt{2n+1}}\delta_{m0}. \tag{3.105}$$

Since $R_0 < a$, we find

$$\Sigma_{nm}\overline{T}^2_{nm} = \sum_{n=0}^{+\infty}\left(\frac{a}{R_0}\right)^{2n+2} \frac{1}{2n+1} = +\infty, \tag{3.106}$$

contrary to the hypothesis that $\overline{T}(P)$ is square integrable over S_0. So the hypothesis proves to be absurd. Since R_0 is any radius such that $R_0 < a$, we have proved that the grain of sand potential cannot be harmonically continued below the level of the grain.

Since we can place the grain at any point below S, we have that not only inside the masses, but even in part of the empty space (3.103) might not converge.

As simple as it is this counterexample permits to state a general rule that we establish in the form of a theorem.

Theorem 4. *Let $\overline{R}$ be the minimum Brillouin radius for the surface S, i.e $\{\overline{R} = \sup_{P\in S} r_P\}$; let us denote by $R_c(T)$ the radius of convergence of the harmonic series that represents a potential $T \in HL^2(S)$*

$$T(P) = \frac{Gm}{R}\sum_{n=0}^{+\infty}\sum_{m=-n}^{n} T_{nm}\left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta,\lambda). \tag{3.107}$$

Following Krarup (2006) we put, by definition,

$$R_c(T) = \inf\left\{\widetilde{R}\,;\,\Sigma_{nm}T_{nm}^2\left(\frac{R}{\widetilde{R}}\right)^{2n+2} < +\infty\right\} \tag{3.108}$$

and obviously any time that $R_c(T) < R$, the series (3.107) is uniformly convergent $\forall r \geq R$. Then we have

$$\sup_{T\in HL^2(S)} R_c(T) = \overline{R}, \tag{3.109}$$

that is: if we want a radius R such that the series (3.107) is convergent for all T in $HL^2(S)$ (in fact one can prove for all potentials harmonic in Ω) one has to put necessarily $R > \overline{R}$.

The reason why there is some confusion on this point in geodetic literature, is due to the fact that although not all T harmonic in Ω, and such that $T|_S = f(P) \in L^2(S)$, can be downward continued, yet it is always possible to make a *small* (in $L^2(S)$ sense) variation of f to obtain an $\overline{f}$, such that the potential F corresponding to $\overline{f}$ in the sense that $F|_S = \overline{f}$, is close to T in Ω, and can indeed be continued harmonically inside B, down to some predefined surface $\overline{S}$, all contained into B.

This is basically one of the possible simplified formulations of a fundamental theorem known in geodetic literature as the *Runge-Krarup theorem*.

Theorem 5 (Runge-Krarup). *Let $T(P)$ be any potential harmonic in Ω and such that $\int_S T^2(P)dS_P < +\infty$; let further $\overline{S}$ be a smooth surface, all included in B, and let us fix an $\varepsilon > 0$. Then we can find a potential $\overline{T}$, harmonic down to $\overline{S}$, which is close to $T(P)$ in the sense that*

$$\int_S [T(P) - \overline{T}(P)]^2 dS < \varepsilon. \tag{3.110}$$

In this very elementary formulation we don't need to prove the theorem, which holds true under much more general conditions, because we can simply observe that $\mathrm{Span}\{S_{nm}(r,\vartheta,\lambda)\}$ is as a matter of fact dense in both $HL^2(\overline{S})$ and $HL^2(S)$ and it consists of functions which are harmonic in the whole space, outside the origin.

When we choose $\overline{S}$ to be a Bjerhammar sphere S_B, with radius R_B, we have a situation which is very much in use in geodesy, where one has an approximate expression for T in terms of a function harmonic down to S_B. However if one tries to restrict ε in (3.110) one finds that the harmonic coefficients in the convergent series (3.107) do change too, and, most of the times, their limits for $\varepsilon \to 0$ does not provide anymore a convergent series.

Nevertheless when we provide only a finite number of observations on S, for instance mean values of T or of Δg over area blocks, we are always able to interpolate them exactly and this is the fact that has generated some confusion and led some authors to believe that a true downward continuation of T existed.

Remark 5. Imagine we take a surface $\widetilde{S}$ that initially coincides with S and then is progressively moved inside B towards the origin. It should be clear then that the set of potentials harmonic down to $\widetilde{S}$ becomes thinner and thinner, though it is always densely embedded into $HL^2(S)$. As a consequence achieving an ε-approximation of T as in (3.110) is always possible but it becomes more and more difficult while we move downward $\widetilde{S}$. For instance if we use a finite linear combination of functions harmonic down to $\widetilde{S}$, we might be forced to take a larger number of them in order to achieve the same level of approximation. It is for this reasons that when we perform a global approximation of T it is not so convenient to use a Bjerhammar sphere as $\widetilde{S}$, but it is preferable to use the earth ellipsoid $\mathcal{E}$. This is in fact much closer to S than any Bjerhammar sphere as the height of the highest mountain is less than $2 \cdot 10^{-3} a (a \cong 6{,}378\,\mathrm{m})$ while a Bjerhammar sphere, globally contained in B, has at most a radius equal to b, the semi-minor axis of $\mathcal{E}$, meaning that it is at least $\sim 20\,\mathrm{km}$ below the surface in equatorial regions.

3.6 Ellipsoidal Harmonics

In this section we shall develop a theory similar to that of Sect. 3.4, establishing a general representation of a potential harmonic outside the ellipsoid $\mathcal{E}$ and square integrable on it. This will be done by a formula which is the exact counterpart of the spherical harmonics series (3.77).

To this aim we go back to Example 4 and recall the definition of ellipsoidal coordinates $(q, \overline{\vartheta}, \lambda)$. In that example we have found the form of the Laplace equation in such coordinates (cf. (1.110), (1.111)) that we repeat here for the sake of readability:

$$(q^2 + E^2)\frac{\partial^2 u}{\partial q^2} + 2q\frac{\partial u}{\partial q} + \overline{\Delta}_\sigma u - \frac{E^2}{q^2 + E^2}\frac{\partial^2 u}{\partial \lambda^2} = 0 \tag{3.111}$$

$$\overline{\Delta}_\sigma = \frac{\partial^2}{\partial \overline{\vartheta}^2} + \mathrm{ctg}\overline{\vartheta}\frac{\partial}{\partial \overline{\vartheta}} + \frac{1}{\sin^2 \overline{\vartheta}}\frac{\partial^2}{\partial \lambda^2}. \tag{3.112}$$

A quite interesting feature of (3.111) is that the angular part of the Laplace operator is constructed by a combination of the Laplace Beltrami operator $\overline{\Delta}_\sigma$ and of $\frac{\partial^2}{\partial\lambda^2}$, with coefficients that do not depend on $(\overline{\vartheta}, \lambda)$.

If we remember (3.63) and (3.66) and the subsequent comments, we find that, when q is kept fixed, we have

$$\overline{\Delta}_\sigma Y_{nm}(\overline{\vartheta}, \lambda) - \frac{E^2}{q^2 + E^2}\frac{\partial^2}{\partial\lambda^2} Y_{nm}(\overline{\vartheta}, \lambda) = \left[-n(n+1) + \frac{E^2 m^2}{q^2 + E^2}\right] Y_{nm}(\overline{\vartheta}, \lambda). \tag{3.113}$$

This suggests the idea of separating the dependence of $u(q, \overline{\vartheta}, \lambda)$ from the angular variables by putting

$$u(q, \overline{\vartheta}, \lambda) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} u_{nm}(q) Y_{nm}(\overline{\vartheta}, \lambda). \tag{3.114}$$

In fact, by fixing q in $u(q, \overline{\vartheta}, \lambda)$ we find as a matter of fact a function of $(\overline{\vartheta}, \lambda)$, which can also be seen as a couple of coordinates of a point running on the unit sphere S_1. Since such a function is quite regular in $\{q > b\}$, we already know that the representation (3.114) has to hold q by q, and even more we know that the coefficients $u_{nm}(q)$ will be given by the orthogonality relation (see (3.57), (3.58))

$$u_{nm}(q) = \frac{1}{4\pi} \int_{S_1} u(q, \overline{\vartheta}, \lambda) Y_{nm}(\overline{\vartheta}, \lambda) d\sigma. \tag{3.115}$$

On the other hand if we substitute (3.114) into (3.111) and take (3.113) into account, by using the linear independence of $\{Y_{nm}\}$ we find that $u_{nm}(q)$ do have to satisfy the differential equation

$$(q^2 + E^2) u''_{nm} + 2q u'_{nm} - \left[n(n+1) - \frac{E^2 m^2}{q^2 + E^2}\right] u_{nm} = 0, \tag{3.116}$$

$$\left(u'_{nm} = \frac{du_{nm}}{dq},\ u''_{nm} = \frac{d^2 u_{nm}}{dq^2}\right).$$

Equations like (3.116) are well-known and studied in mathematical literature and we can even find a quite interesting relation to the Legendre equation (3.64); in fact if we put $q = -iEt,\ (i^2 = -1)$, into (3.116) we find that, as a function of t, $u_{nm}(t)$ has to satisfy exactly the Legendre equation

$$(1 - t^2) u''_{nm} - 2t u'_{nm} + \left[n(n+1) - \frac{m^2}{1 - t^2}\right] u_{nm} = 0. \tag{3.117}$$

Yet we cannot think of using the solutions of (3.117) which we already know, i.e. $P_{nm}(t)$, because these, extended to the complex plane, are not bounded for $|t| \to \infty$.

In fact, since $u(q, \overline{\vartheta}, \lambda)$ given by (3.114) has to be a regular potential at infinity, we certainly want solutions $u_{nm}(q)$ of (3.116) that do tend to zero when $r \to \infty$, i.e. when $q \to \infty$. This is also obvious because (cf. (1.103))

$$r^2 = q^2 + E^2 \sin^2 \vartheta.$$

Solutions of (3.117) with such characteristics are known as *Legendre associated functions of second kind* (cf. Heiskanen and Moritz 1967; Nikiforov and Uvarov 1988). They are usually denoted by $Q_{nm}(t)$. It is even possible to see that

$$Q_{nm}(t) \sim \frac{c}{|t|^{n+1}}, \ |t| \to \infty$$

which, expressed in terms of the variable q and then r, is nicely reproducing the asymptotic behaviour of spherical harmonics.

Summarizing, if we put

$$u_{nm}(q) = u^e_{nm} v_{nm}(q) = u^e_{nm} \frac{Q_{nm}(i\frac{q}{E})}{Q_{nm}(i\frac{b}{E})} \tag{3.118}$$

we find a set functions that do satisfy (3.116), tend to zero when $q \to \infty$ and, when we put $q = b$, yields

$$u_{nm}(b) = u^e_{nm}, v_{nm}(b) = 1 \tag{3.119}$$

By setting $q = b$ in (3.115), we see that

$$u^e_{nm} = \frac{1}{4\pi} \int_{S_1} u(b, \overline{\vartheta}, \lambda) Y_{nm}(\overline{\vartheta}, \lambda) d\sigma, \tag{3.120}$$

showing once more that if we know u on the boundary $\mathcal{E}$ we can compute u^e_{nm} from (3.120) and then recover $u(q, \overline{\vartheta}, \lambda)$ by using (3.118) into (3.114).

An important point is that one can see that $Q_{nm}(t)$ have a parity opposite to $n - |m|$ so that $Q_{nm}(i\frac{q}{E})$ is a pure imaginary number when $n - |m|$ is even and a real number when $n - |m|$ is odd. Accordingly, the ratio $\dfrac{Q_{nm}(i\frac{q}{E})}{Q_{nm}(i\frac{b}{E})}$ is always real, as it is necessary if we want our potential given by (3.114) to be real too. So if we define solid ellipsoidal harmonics as

$$S^e_{nm}(q, \overline{\vartheta}, \lambda) = \frac{Q_{nm}(i\frac{q}{E})}{Q_{nm}(i\frac{b}{E})} Y_{nm}(\overline{\vartheta}, \lambda) \tag{3.121}$$

we have established a general representation of a potential u harmonic outside $\mathcal{E}$, in the form

$$u(q, \overline{\vartheta}, \lambda) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} u^e_{nm} S^e_{nm}(q, \overline{\vartheta}, \lambda), \tag{3.122}$$

with the ellipsoidal coefficients $\{u^e_{nm}\}$ given by (3.120).

Remark 6. If we remember the Example 4, and the expression for line elements on the ellipsoid $\mathcal{E}$ corresponding to the choice $q = b$, see (1.107),

$$\begin{aligned} d\ell_\vartheta &= h_\vartheta d\vartheta = \sqrt{b^2 + E^2 \sin^2 \overline{\vartheta}}\, d\vartheta \\ d\ell_\lambda &= a \sin \overline{\vartheta}\, d\lambda \end{aligned} \tag{3.123}$$

we see that the area element of $\mathcal{E}$ is

$$\begin{aligned} dS_e &= ab\sqrt{1 + e'^2 \sin^2 \overline{\vartheta}}\, \sin \overline{\vartheta}\, d\overline{\vartheta}\, d\lambda \\ &= abW(\overline{\vartheta}) d\sigma, \end{aligned} \tag{3.124}$$

with $d\sigma$ the usual area element of S_1 and e'^2 the second eccentricity, $e'^2 = \frac{E^2}{b^2}$.

Due to the presence of the weight function $W(\overline{\vartheta})$, the sequence $\{Y_{nm}(\overline{\vartheta}, \lambda)\}$ is not orthonormal in $L^2(S_e)$, although it is complete in such a space. In fact if we map $\mathcal{E}$ onto the unit sphere S_1 through the coordinates $(\overline{\vartheta}, \lambda)$, (see Fig. 3.4) we see that

$$\frac{1}{4\pi} \int_{S_1} u^2(b, \overline{\vartheta}, \lambda) d\sigma = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} (u^e_{nm})^2. \tag{3.125}$$

At the same time, since $ab\, d\sigma \le dS_e \le a^2 d\sigma$, we have

$$\frac{1}{4\pi} \int_{S_1} u^2(b, \overline{\vartheta}, \lambda) d\sigma \le \frac{1}{a^2} \frac{1}{4\pi} \int_{S_e} u^2(b, \overline{\vartheta}, \lambda) dS_e \tag{3.126}$$

$$\frac{1}{4\pi} \int_{S_1} u^2(b, \overline{\vartheta}, \lambda) d\sigma \ge \frac{1}{ab} \frac{1}{4\pi} \int_{S_e} u^2(b, \overline{\vartheta}, \lambda) dS_e. \tag{3.127}$$

Relations like (3.126) and (3.127) prove that the ordinary norm in $L^2(S_e)$ is equivalent to the $L^2(S_1)$ norm, after the mapping $P_e \equiv (\overline{\vartheta}, \lambda) \leftrightarrow P_S \equiv (\overline{\vartheta}, \lambda)$ between $\mathcal{E}$ and S_1 has been used (cf. Fig. 3.4).

This implies that the $L^2(S_1)$ convergent series (3.122) is also a convergent series in $L^2(S_e)$. Therefore (3.125) is a necessary and sufficient condition for $u(b, \overline{\vartheta}, \lambda)$ to be in $L^2(S_e)$.

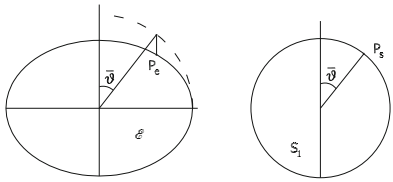


Fig. 3.4 The mapping $\mathcal{E} \leftrightarrow S_1$ in a meridian plane

In order to perform a calculus with ellipsoidal harmonics, implying also the numerical determination of $Q_{nm}(i\frac{q}{E})$ for any value of q, it would be nice to use recursive relations like those we found for spherical harmonics (see (3.67), (3.68)).

As a matter of fact one can prove that the same relations hold for $\overline{P}_{nm}$ and Q_{nm}, yet for the relevant arguments of $t = i\frac{q}{E}$ (note that $\frac{q}{E} > \frac{b}{E} \geq 12$), such relations become quite unstable and they cannot be used with n larger than $\sim$20. There are series developments of $Q_{nm}(z)$ in literature; however also the coefficients of the series become quite large when the degree rises over 1,000, as it is possible and necessary today.

One of the methods presently used is to apply an explicit and computable transformation from solid spherical harmonics to solid ellipsoidal harmonics and viceversa.

It is not appropriate to derive here the coefficients of this transformation, for which we send the interested reader to the literature (cf. Hobson 1955; Jekeli 1988). Yet we mention that it is indeed expected that a solid ellipsoidal harmonic $S^e_{\ell m}(q, \overline{\vartheta}, \lambda)$ (see (3.121)) could be expressed in terms of a series of spherical harmonics, $S_{nm}(r, \vartheta, \lambda)$ because after all it is a harmonic function outside a sphere with radius R of the order of the ellipsoid semiaxes.

Even more, since the longitude λ is the same for both ellipsoidal and spherical coordinate systems, we expect $Y_{\ell m}(\overline{\vartheta}, \lambda)$ to be a linear combination of $Y_{nm}(\vartheta, \lambda)$ *with the same order* m, because in this way both $S^e_{\ell m}$ and $S_{nm}(r, \vartheta, \lambda)$ depend on the same $\sin|m|\lambda$ or $\cos m\lambda$. Furthermore, since both $S^e_{\ell m}(q, \overline{\vartheta}, \lambda)$ and $S_{nm}(r, \vartheta, \lambda)$ have a definite parity as functions of $\overline{\vartheta}$ and ϑ respectively and since such parity is alternating (even and odd) with n, we can predict that $S^e_{\ell m}(q, \overline{\vartheta}, \lambda)$ can depend only on $S_{\ell \pm 2k, m}(r, \vartheta, \lambda)$. It turns out that the above linear combination has a particular form; more precisely, if we reason directly in terms of harmonic coefficients, there are constants

$$\lambda_{\ell m k}, \quad k = 0, 1, \ldots, I_{\ell m} = \left[\frac{\ell - |m|}{2}\right], \tag{3.128}$$

$[t]$ meaning the largest integer equal or smaller than the real number t, such that

$$u^e_{\ell m} = \sum_{k=0}^{I_{\ell m}} \lambda_{\ell m k} u^s_{\ell - 2k, m}, \tag{3.129}$$

where $u^e_{\ell m}, u^s_{nm}$ are respectively the harmonic coefficients of the potential u represented with ellipsoidal or spherical harmonics. Among other things (3.129) says that even if we have a potential u which is given by a finite sum of spherical harmonics, then the corresponding ellipsoidal representation will have coefficients different form zero for all degrees, naturally with a maximum value for the order m.

The relation (3.129) can be inverted in the form

$$\Lambda_{nmk}, \quad k = 0, 1, \dots I_{nm}, \tag{3.130}$$

$$u^s_{nm} = \sum_{k=0}^{I_{nm}} \Lambda_{nmk} u^e_{n-2k,m}. \tag{3.131}$$

The coefficients $\lambda_{nmk}, \Lambda_{nmk}$ can be computed by recursive relations, as described for instance in Jekeli (1988).

Although there are a number of methods to compute corrective terms to switch from the ellipsoidal to the spherical set up (see for instance Cruz (1986)), we report here only approximate formulas which exploit a perturbation in the eccentricity parameter e^2 and the fact that, for terrestrial applications in the topographic layer, we need only to compute Q_{nm} with q close to b, say with $|q - b| \leq 10^{-3}b$ (Sona 1995).

Such formulas can be summarized as

$$v_{nm} \cong \frac{1}{s^{n+1-\alpha}} \cong \frac{1}{s^{n+1}} \left[1 + e'^2 \frac{(n+1)(n+2) + m^2}{2n+1} (s-1) \right]; \tag{3.132}$$

the proof can be fond in Sect. A.3, where the value of α is given by (3.201) (Sona 1995).

The relative approximation of the simple formula (3.132) is in the range of 10^{-5} as far as we stay in the topographic layer and it is practically sufficient for most of our computations, when the maximum degree is at the level of hundreds, e.g. up to degree 360.

Remark 7. Now that we possess the full concept of ellipsoidal harmonics we can return to Sect. A.4 and observe that the determination of the normal potential was reduced to the research of a function $V_e(q, \overline{\vartheta})$, harmonic outside the ellipsoid $\mathcal{E}$ and satisfying on $\mathcal{E}$ the boundary condition (cf. (1.119))

$$V_e|_{\mathcal{E}} = U_0 - \frac{1}{2}\omega^2 a^2 \sin^2 \overline{\vartheta}. \tag{3.133}$$

The solution, explicitly constructed in Sect. 1.9, was given by (1.127).

Now, if we take into account that

$$\frac{2}{3} - \sin^2 \overline{\vartheta} = \cos^2 \overline{\vartheta} - \frac{1}{3} = \frac{2}{3} P_2(\cos \overline{\vartheta}),$$

(1.127) can be written as

$$V_e = \frac{GM}{E}\arctan\frac{E}{q} + \frac{1}{3}\frac{\omega^2 a^2}{Q(b)}Q(q)P_2(\cos\overline{\vartheta}). \tag{3.134}$$

A comparison with (3.121) and (3.122), recalling that $Y_{20}(\overline{\vartheta},\lambda) = \sqrt{5}P_2(\cos\overline{\vartheta})$ (see (3.50), (3.51), (3.52), (3.53)) shows directly that the gravitational part of the normal potential is just a combination of two ellipsoidal harmonics.

Since this will be useful in the sequel, we want to find here as well the representation of V_e in spherical harmonics. In fact we know a priori that, at least for $r > a$, V_e must have a convergent representation in terms of spherical harmonics. Considering the cylindrical symmetry of $V_e(r,\vartheta)$ we know that only the zonal coefficients of $Y_{nm}(\vartheta,\lambda)$, i.e. of $P_n(\cos\vartheta)$, must be different from zero.

Furthermore since V_e has to be symmetric with respect to the equatorial plane, only coefficients with even degree and zero order have to be different form zero. Traditionally, $V_e(r,\vartheta)$ is represented in the form (cf. Heiskanen and Moritz 1967)

$$\begin{aligned} V_e(r,\vartheta) &= \frac{GM}{r} - \frac{GM}{a}\sum_{n=1}^{+\infty} J_{2n}\left(\frac{a}{r}\right)^{2n+1} P_{2n}(\cos\vartheta) \\ &= \frac{GM}{r}\left[1 - \sum_{n=1}^{+\infty} J_{2n}\left(\frac{a}{r}\right)^{2n} P_{2n}(\cos\vartheta)\right]. \end{aligned} \tag{3.135}$$

In order to find a relation between J_{2n} and the constants used in (3.134) we take advantage of the fact that (cf. (1.103); Heiskanen and Moritz 1967, Sect. 2.9)

$$\vartheta = 0 \Rightarrow \overline{\vartheta} = 0,\ q = z = r\ ;$$

therefore one must have

$$\begin{aligned} &\frac{GM}{E}\arctan\frac{E}{r} + \frac{1}{3}\frac{\omega^2 a^2}{Q(b)}Q(r) \\ &\equiv \frac{GM}{r}\left[1 - \sum_{n=1}^{+\infty} J_{2n}\left(\frac{a}{r}\right)^{2n}\right] \end{aligned} \tag{3.136}$$

at least for every $r > a$.

By using the Taylor series

$$\arctan x = \sum_{n=0}^{+\infty}\frac{(-1)^n x^{2n+1}}{2n+1} \tag{3.137}$$

and recalling the formula (1.127) that defines $Q(r)$, one gets after some algebra for the first member of (3.136)

$$V_e(r,\vartheta) = \frac{GM}{r}\sum_{n=0}^{+\infty}\frac{(-1)^n}{2n+1}\left[1-\frac{\mu e'}{3Q(b)}\,\frac{4n}{3n+3}\right]\frac{E^{2n}}{r^{2n}} \tag{3.138}$$

where

$$\mu = \frac{\omega^2 a^2 b}{GM}$$
$$Q(b) = \left(\frac{3}{e'^2}+1\right)\arctan\frac{1}{e'} - \frac{3}{e'}$$
$$(e')^2 = \frac{a^2-b^2}{b^2}.$$

By comparing (3.138) with (3.136) we finally get

$$J_{2n} = (-1)^{n+1}\frac{(e^2)^n}{2n+1}\left[1-\frac{\mu e'}{3Q(b)}\,\frac{4n}{2n+3}\right], \tag{3.139}$$

where e^2 is as usual the squared eccentricity of the first kind.

To make (3.139) more manageable one can write it for $n = 1$, namely

$$J_2 = \frac{e^2}{3}\left[1-\frac{4\mu e'}{15Q(b)}\right], \tag{3.140}$$

derive $\frac{\mu e'}{Q(b)}$ from it and substitute back into (3.139) to get

$$J_{2n} = (-1)^{n+1}\frac{3(e^2)^n}{(2n+1)(2n+3)}\left[1-n+5\frac{J_2}{e^2}n\right]. \tag{3.141}$$

Equation 3.141, knowing that $J_2 \sim 10^{-3}$ and $e^2 \sim 6,7\cdot 10^{-3}$, gives quite a good representation of the velocity with which J_{2n} tend to zero.

3.7 Global Models as Approximate Solution of Boundary Value Problems

Generally speaking a global model of the gravity field anomalous potential $T(P)$, or global geopotential model, is a finite linear combination of functions $H_m(P)$ that are regular harmonic on S and in the whole outer space Ω

$$T_M(P) = \sum_{m=1}^{M} a_m H_m(P) \; ; \tag{3.142}$$

what makes of $T_M(P)$ a "global" model is that the coefficients $\{a_m\}$ in (3.142) are chosen in such a way as to reproduce as closely as possible global sets of observations, i.e. sets of data that cover geographically the whole of the earth surface, or most of it. So one first fundamental requirement on the sequence $\{H_m(P)\}$ is that when these base functions are restricted to S they form a complete system in the space to which we assume that the actual anomalous potential belongs. In the context of this book we assume that such space is $HH^{1,2}(S)$ (cf. Sect. 2.2), i.e. the space of functions which are regular harmonic in Ω and have a gradient which is square integrable on the boundary S.

This condition is to some extent natural in the sense that it generates functions like gravity disturbances δg, gravity anomalies Δg or deflections of the vertical (ξ, η) that can be defined on the boundary S and are square integrable there.

There are as a matter of fact many base functions that could be used to build global models T_M. For instance it is worth mentioning that potentials generated by point masses suitably distributed in layers at different levels inside B, is one such alternative which has been studied and used in literature (Bjerhammar 1987; Marchenko 1998).

Yet by far the most important type of global models, as of today, uses as base functions the solid spherical harmonics; so we shall put by definition

$$\begin{aligned} T_M(P) &= \frac{GM}{\overline{R}} \sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} T_{\ell m} S_{\ell m}(r, \vartheta, \lambda) \\ &= \frac{GM}{\overline{R}} \sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} T_{\ell m} \left(\frac{\overline{R}}{r} \right)^{\ell+1} Y_{\ell m}(\vartheta, \lambda). \end{aligned} \tag{3.143}$$

In (3.143) $\overline{R}$ is a purely conventional radius and it can be chosen of whatever value, although the ordinary choice is equal to the mean radius of the earth. Notice that the conventional factor $\frac{GM}{\overline{R}}$ in front of (3.143) allows to consider $T_{\ell m}$ as non-dimensional numbers.

That $\{S_{\ell m}(r, \vartheta, \lambda)\}$, restricted to S, are a complete, but non orthonormal, system in $L^2(S)$ has been illustrated in Sect. 3.4; that they form a complete system in $H^{1,2}(S)$, i.e. in the space of traces on S of potentials in $HH^{1,2}(S)$, is a true fact that will not be further investigated here. This statement however could be deduced from the results presented in Part III, Chap. 14.

The fact that spherical harmonics $\{S_{\ell m}\}$ instead of ellipsoidal harmonics $\{S^e_{\ell m}\}$ are used in (3.143) is due, in the author's opinion, to three reasons:

- the fact that numerical calculations in spherical harmonics are much simpler than the corresponding calculations in ellipsoidal harmonics where Legendre functions of the second kind have to be used,

- the belief, corroborated by numerical experiences as well as by the perturbative theory of Part III, Chap. 14, that it is possible to "correct" the data, e.g. gravity anomalies Δg, of the *ellipsoidal effects* by exploiting some prior knowledge of the gravity field so that the observation functionals are reduced to the more convenient *spherical approximation* form, (remember Sect. 2.6 concerning the definition of spherical approximation used here),
- the fact that it is possible to transform spherical harmonics into ellipsoidal harmonics as well as spherical harmonic coefficients $T_{\ell m}$ into ellipsoidal harmonic coefficients $T^e_{\ell m}$ and viceversa (cf. (3.129), (3.131)) so that we are likely not loosing information by the use of the model (3.143).

We have now to define what are the values of the (integer) parameters L, M, i.e. the minimum and maximum degree that we want to be represented by our model (3.143). We start to discuss L.

As explained many times, L should not be 0 or 1, because $T_{00} = 0$ by a suitable choice of the normal potential and $T_{1,m} = 0, (m = -1, 0, 1)$, by the choice of reference system. Both choices are consequences of satellite tracking results. As a matter of fact by those techniques and the more recent results of space gravimetry we could say the first 10 or 20 degrees to be so well-known that they could be considered as fixed and eliminated from (3.143) at least when the unknown $T_{\ell m}$ have to be determined; of course they have to be added back when we want to represent the full anomalous potential.

So we can agree that $L = 2$, when we consider (3.143) as a representation of T_M given the coefficients, but it could be higher when we decide to determine $T_{\ell m}$ from data. This is important for instance for the theory developed in Part III.

As for the choice of M, this is the result of a compromise between the distribution of the available data and our desire to obtain a better and better approximation of T. The correct term to describe the phenomenon we are going to investigate is "resolution". We shall represent it by means of the side Δ of a regular geographic grid at the knots of which we are able to provide the data necessary to determine T_M. For instance, if we have Δg data on S and we are able to produce a grid of mean values of Δg over blocks of dimension $0.5° \times 0.5°$, implying that over almost all the surface S we have data enough to form block averages on $0.5° \times 0.5°$ areas, we say that we have a resolution of $0.5°$. If we have holes in the data there are various techniques to fill them without destroying or biasing the original information present in other areas, at least when holes are not too large.

We translate that number into a linear scale by taking the length of a $0.5°$ arc at the equator, namely $\sim$55 km.

So if we think that we have enough data to produce a $5' \times 5'$, we say that we have $5'$ (or $\sim$9 km) resolution in the data. This means that we are able to provide 9,331,200 values and we do not try to see any tiny element in our data set below the size of 9 km.

There is an important relation between the number Δ described above and the maximum value of M that we can choose in (3.143) to represent T_M. This is in some sense similar to what happens on a circle. If we have $2N + 1$ points on a

circle, with a distance $\Delta = \frac{360^\circ}{2N+1}$ from one another, we can determine the Fourier coefficients of sines and cosines up to the frequency N, i.e. the maximum integer smaller than one half of the points where we have data. On the same time $\cos(N\vartheta)$ and $(\sin N\vartheta)$ have each $2N$ zeros on the circle, i.e. as many zeros as data (minus one).

What happens for the circle is that if we try to use sines or cosines with a frequency higher than N, these functions do reproduce the same values as a sine or cosine of lower frequency on the grid of points at distance Δ.

The reader can check this by the useful example

$$\cos 3k\frac{2\pi}{5} \equiv \cos 2k\frac{2\pi}{5} \quad k = 0, 1, 2, 3, 4 \tag{3.144}$$

corresponding to $N = 2, 2N + 1 = 5$.

This phenomenon is called *aliasing* and it means that when we use Δ grids we don't recognize in a function a behaviour regularly oscillating with zeros closer than Δ. More on this can be found in Chap. 10 of Part II.

By the way, the situation on the sphere is not so neat, because the distance between points of a grid regular in (ϑ, λ) is not constant; of course two points with the same colatitude and longitudes different by a certain angle Δ are closer if they are chosen in the polar regions rather than in the equatorial belt. This makes the spherical aliasing more difficult to study, though possible indeed (Albertella et al. 1992; Jekeli 1996; Driscoll and Healy 1994).

Yet both theory and numerical proofs have shown that by assuming as a rule of thumb the same formula as for the circle, namely

$$M \cong \frac{360^\circ}{2\Delta}, \tag{3.145}$$

we can avoid aliasing, i.e. the coefficients $T_{\ell m}$ of T_M are uniquely determinable. Indeed the true gravity field is not a band limited function, so the above statement holds only for the model T_M.

In fact for a certain maximum degree M we have the functions $Y_{M,0}(\vartheta)$ which are polynomials of degree M in $\cos\vartheta$, so they have M zeros in 180° and their mean distance is $\Delta \sim 180/M$ as in (3.145). Recall here that the variables ϑ and $t = \cos\vartheta$ are in a one-to-one correspondence when $0 \le \vartheta \le \pi$.

Similarly, with same value of M, we have the functions $Y_{M,M}(\vartheta, \lambda)$ and $Y_{M,-M}(\vartheta, \lambda)$ which are proportional to $\cos M\lambda$ and $\sin M\lambda$, i.e. they have $2M$ zeros over the full turn of 360°, again agreeing with (3.145). So we shall definitely adopt the rule (3.145), implying for instance that a resolution of 0.5° allows the determination of a model up to degree 360, while a grid of 5′ will permit a model up to degree 2,160, which is the limit that has been recently achieved (Pavlis et al. 2008).

Now the problem we have to face is: what are the data and what are the methods to be used in order to determine the coefficients $T_{\ell m}$ of a global model?

Having more data than unknowns, one is inclined to apply the least squares method.

However if the stochastic nature of the errors is not properly described, i.e. a simple sum of the squares of the residuals is minimized, then (as it happens in reality) a long wavelength error present in the data will be absorbed by the estimated coefficients without leaving any trace into the residuals. This type of errors will be identified only by comparison with an independent data set. This subject will be discussed in depth in Chap. 6.

As for the data, there are three principal sources of information used to generate T_M; (a) satellite tracking or satellite gravimetry, (b) satellite altimetry on ocean, (c) gravimetry on solid earth.

Let us examine them in short, separately:

(a) Space techniques have improved enormously the data available on the gravity field and, without going into details, coefficients up to degree 200–300 can be usefully determined in this way (see also Part III, Sect. 15.7). As such they can enter into the process of determining high resolution global models (e.g. $M = 2{,}160$) as a first useful guess or approximation. However one point has to be clear: when we use a finite amount of data, the coefficients determined by satellite measurements are not the *same* as those determined by ground measurements, because they respond to different optimization criteria. In particular while the satellite coefficients have a clear relation to physical moments of the mass distribution (see Sect. 3.2), the coefficients determined from ground data are only derived on the basis of suitable mathematical criteria. This is clearly illustrated by the fact that if from satellite data we were able to cover a Brillouin sphere (out of all the masses) with noiseless observations of some suitable functional of T, then we could recover the coefficients $T_{\ell m}$ in an exact way, up to any prefixed degree and order M while if we covered the earth surface with known errorless functionals of T we could only set up an approximation procedure where the $T_{\ell m}$ estimates change in principle as functions of the maximum degree M used in the model T_M,

(b) Satellite altimetry provides, by radar measurements repeated in time, a quite accurate evaluation (in the range of a few centimeters) of the stationary surface of oceans, cleaned from waveforms, tidal effects and various seasonal phenomena. The resulting data then are first the sum of geoid and a height component called *sea surface topography*, or *dynamic height*; so called because the difference between sea surface and geoid is sustained by dynamic effects related to steady currents. In terms of an equation, if we call h_0 the oceanic mean surface, N the geoid and η the dynamic height, we have

$$h_0 = N + \eta = \frac{T}{\gamma} + \eta. \tag{3.146}$$

It follows that, if oceanographers provide us with a sufficiently accurate dynamic model of η, we can derive the relation on oceanic areas from (3.146)

$$T = \gamma(h_0 - \eta). \tag{3.147}$$

Alternatively η can be parametrized and estimated from data, Rapp (1997b),

(c) The gravimetric observations on continental areas (but nowadays also from airborne gravimetry), combined with altimetric observations, have already been analyzed in Sect. 2.3. Ultimately they lead to the linearized equation for free air anomalies (cf. (2.37))

$$\Delta g = -\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma} T \qquad \left(\gamma' = \frac{\partial \gamma}{\partial h}\right). \tag{3.148}$$

When the ellipsoidal height together with g are observed, e.g. by a GPS receiver, (3.148) has to be substituted by the simpler gravity disturbance equation

$$\delta g = -\frac{\partial T}{\partial h}. \tag{3.149}$$

Although times are clearly evolving from the use of (3.148) to that of (3.149), yet at present the large majority of available data are in the form of free air anomalies and this is in fact the data set still used to produce global models.

So in principle the determination of T can be formulated, at least as a limit case when we have data covering the whole boundary, as the solution of the boundary value problem (e.g. cf. Sansò 1997)

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ -\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma} T = \Delta g & \text{on } S_L \\ T = \gamma(h_0 - \eta) & \text{on } S_O \\ T \to 0, \ r \to \infty \end{cases} \tag{3.150}$$

$$(S_L = \text{Land part of } S, \quad S_O = \text{Ocean part of } S).$$

Indeed in reality, instead of continuous data we have block mean values of Δg on land and block mean values of T on the ocean. The resolution nowadays achievable is $5' \times 5'$, corresponding to a global solution of maximum degree 2,160.

The standard method to get this solution could be least squares, in Hilbert space sense, as described in Part III, Sect. 14.4.

And in fact this solution has been implemented, e.g. up to degree 90 (see Rapp 1997a). This implies the solution of a normal system with 8,100 unknowns and no special structure of the normal matrix. However at this point we already have very good models up to degree 360, which can be used for an intermediate step that dramatically simplifies our problem. In fact, it will be shown in Chap. 9 of Part II that, if the long wavelength content of T is subtracted from ocean observations, with the residual part T_{res} one can perform a very good prediction of mean block values

of the corresponding Δg_{res} and finally we add back to this a Δg_{prior}, consistent with T_{prior}, to obtain $\Delta g = \Delta g_{\text{prior}} + \Delta g_{\text{res}}$.

So we are left with a much simpler problem, namely to find the approximate solution of the BVP.

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ -\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma} T = \Delta g & \text{on } S \\ T \to 0, \ r \to \infty \end{cases} \tag{3.151}$$

by means of a model T_M of the type (3.143). This can be (and has been) done in two ways when very high degrees ($>10^3$) are involved. The procedures will be described in more detail in Part III, Sects. 14.4 and 14.5, relating them to better known methods of mathematical analysis.

On the other hand the specific implementation of theoretical ideas is going to be fully presented in Part II. Basically it relies on the concept that, after a least squares approximation up to some intermediate degree, the higher degrees are determined by downward continuing Δg to the ellipsoid $\mathcal{E}$ and then using the orthogonality relations (3.120).

3.8 Commission and Omission Errors. Kaula's Rule

The problem we want to face now is how to assess the quality of our model, i.e. to answer to the question: *how well is T_M fitting T?* To reason on such a matter we first note that if we had really performed a least squares solution, the r.m.s of residuals with respect to observations would be the natural quality index. However, if the stochastic nature of the errors is not perfectly described, i.e. a simple sum of the squares of the residuals is minimized, then (as it happens in reality) a long wavelength present in the data will be absorbed by the estimated coefficients, (see also Chap. 6).

On the other hand we want to point out here that the residuals will contain two types of errors: one is the error of the measurements which propagates into the estimates of the coefficients $T_{\ell m}$ up to the maximum degree M, the other is the model error due to the fact that the true T has a part that cannot be modeled in any event by a finite sum of harmonics. Our purpose is exactly to explain how to distinguish between the two and to evaluate them.

In order to make quantitative our reasoning we need to use a simplified situation, namely we shall use the Galerkin method (Mikhlin 1964; Kirsch 1996) assuming that the true T can really be continued down to the sphere $\overline{S}$, with radius $\overline{R}$, so that we are entitled to write

$$T(P) = \frac{GM}{\overline{R}} \sum_{\ell=L}^{+\infty} \sum_{m=-\ell}^{\ell} T_{\ell m} S_{\ell m}(r, \vartheta, \lambda) \tag{3.152}$$

$$S_{\ell m}(r,\vartheta,\lambda)=\left(\frac{\overline{R}}{r}\right)^{\ell+1}Y_{\ell m}(\vartheta,\lambda), \tag{3.153}$$

$$T_M(P)=\frac{GM}{\overline{R}}\sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell}\widehat{T}_{\ell m}S_{\ell m}(r,\vartheta,\lambda), \tag{3.154}$$

$$\widehat{T}_{\ell m}=T_{\ell m}+\tau_{\ell m}, \tag{3.155}$$

where $\tau_{\ell m}$ are the estimation errors of the $T_{\ell m}$. Note should be taken that we start both (3.152) and (3.154) from L, corresponding to the idea that the lower harmonics, say the first 24 degrees to fix the ideas, are perfectly known and subtracted everywhere, from data as well as from models.

With such formulas we can compute the norm of the residual anomalous potential

$$\begin{aligned}T-T_M=T_{\rm res}=&\frac{GM}{\overline{R}}\sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell}\tau_{\ell m}S_{\ell m}(r,\vartheta,\lambda)\\&+\frac{GM}{\overline{R}}\sum_{\ell=M+1}^{+\infty}\sum_{m=-\ell}^{\ell}T_{\ell m}S_{\ell m}(r,\vartheta,\lambda).\end{aligned} \tag{3.156}$$

The norm of $T_{\rm res}$ is taken in the sense of $L^2(\overline{S})$, i.e.

$$\begin{aligned}\|T_{\rm res}\|^2_{L^2(\overline{S})}=&\left(\frac{GM}{\overline{R}}\right)^2\sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell}\tau^2_{\ell m}\\&+\left(\frac{GM}{\overline{R}}\right)^2\sum_{\ell=M+1}^{+\infty}\sum_{m=-\ell}^{\ell}T^2_{\ell m}.\end{aligned} \tag{3.157}$$

As we can see this norm does depend from the random variables $\tau_{\ell m}$ which ultimately depend on the noise measurement, so it is only natural to take as an index of the total error the average of (3.157) with respect of the variability of the noise.

Since $E\{\tau_{\ell m}\}=0$, as we shall see in a minute, we have

$$E\{\tau^2_{\ell m}\}=\sigma^2(\tau_{\ell m})\,; \tag{3.158}$$

then taking the expectation of (3.157) we find

$$\begin{aligned}\mathcal{E}^2_{\rm tot}=E\{\|T_{\rm res}\|^2_{L^2(\overline{S})}\}=&\left(\frac{GM}{\overline{R}}\right)^2\sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell}\sigma^2(\tau_{\ell m})\\&+\left(\frac{GM}{\overline{R}}\right)^2\sum_{\ell=M+1}^{+\infty}\sum_{m=-\ell}^{\ell}T^2_{\ell m}.\end{aligned} \tag{3.159}$$

The first term in the R.H.S. of (3.159) is called *commission error* while the second is called *omission error*, because the first depends on the errors that we commit by measuring, while the second depends from the degrees that we omit from the model.

We denote them by

$$\mathcal{CE}^2 = \left(\frac{GM}{\overline{R}}\right)^2 \sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} \sigma^2(\tau_{\ell m}^2) \tag{3.160}$$

$$\mathcal{OE}^2 = \left(\frac{GM}{\overline{R}}\right)^2 \sum_{\ell=M+1}^{+\infty} \sum_{m=-\ell}^{\ell} T_{\ell m}^2. \tag{3.161}$$

In order to proceed to the evaluation of $\mathcal{CE}$ and $\mathcal{OE}$ we need on one side to describe the propagation of the measurement noise to $\tau_{\ell m}$, on the other side we must have some guess on the order of magnitude of $T_{\ell m}$ for $\ell > M$.

(a) *Error propagation.* Assume that the data used to derive $\widehat{T}_{\ell m}$ are free air gravity anomalies, already reduced to the sphere $\overline{S}$, averaged on squared geographic blocks B_{rs} of size $\Delta \times \Delta$, i.e.

$$(\overline{\Delta g}_{\text{obs}})_{rs} = \frac{1}{N_{rs}} \Sigma_{P_n \in B_{rs}} \Delta g_{\text{obs}}(P_n) \tag{3.162}$$
$$(N_{rs} = \text{number of } P_n \in B_{rs})$$

where

$$\begin{cases} \Delta g_{\text{obs}}(P_n) = \Delta g(P_n) + \nu_n \\ E\{\nu_n\} = 0,\ E\{\nu_n \nu_j\} = \delta_{nj} \sigma_g^2(\overline{P}_{rs}), \end{cases} \tag{3.163}$$

with $\overline{P}_{rs}$ the centers of B_{rs}.

Indeed the hypothesis that $\Delta g_{\text{obs}}(P_j)$ have all the same error variances when $P_n \in B_{rs}$ might be a simplification, although it reflects the real fact that usually gravity measurements in a certain area are performed all together with the same instruments, so that the hypothesis is at least plausible.

We shall assume further that there exists a smooth function $\mu(P)$, that we could call *area data density*, such that, denoting with $|B_{rs}| \cong \sin \vartheta_{rs} \Delta^2$ the size of B_{rs}, the relation holds

$$N_{rs} = \mu(\overline{P}_{rs}) \cdot |B_{rs}|. \tag{3.164}$$

Note that, if N_{tot} is the total number of observations involved, then

$$N_{\text{tot}} = \Sigma N_{rs} \cong \int \mu(P) d\sigma. \tag{3.165}$$

Now we write a simple approximate orthogonality relation on $\overline{S}$, namely

$$\left(\frac{GM}{\overline{R}^2}\right)(\ell-1)\widehat{T}_{\ell m} \cong \frac{1}{4\pi}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})(\overline{\Delta g}_{\text{obs}})_{rs}\,|B_{rs}|\ , \qquad (3.166)$$
$$|m| \le \ell,\ L \le \ell \le M.$$

If we use (3.162) and (3.163) in (3.166) we can write

$$\widehat{T}_{\ell m} = \left(\frac{GM}{\overline{R}^2}\right)^{-1}(\ell-1)^{-1}\left\{\frac{1}{4\pi}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})(\overline{\Delta g})_{rs}\,|B_{rs}|\right.$$
$$\left.+\frac{1}{4\pi}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})\delta\overline{\Delta g}_{rs}\,|B_{rs}|\right\} = T_{\ell m} + \tau_{\ell m} \qquad (3.167)$$

where we have put

$$\delta\overline{\Delta g}_{rs} = \frac{1}{N_{rs}}\Sigma_{P_n \in B_{rs}} \nu_n. \qquad (3.168)$$

In this way we have found the direct relation between the measurement errors ν_n and the estimation error $\tau_{\ell m}$, i.e.

$$\tau_{\ell m} = \left(\frac{GM}{\overline{R}^2}\right)^{-1}(\ell-1)^{-1}\frac{1}{4\pi}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})\delta\overline{\Delta g}_{rs}\,|B_{rs}|\,. \qquad (3.169)$$

Due to (3.163) we see that

$$E\{\tau_{\ell m}\} = 0,$$

as already anticipated. Moreover, using (3.163) and (3.168), we have

$$E\{\delta\overline{\Delta g}_{rs}\delta\overline{\Delta g}_{uv}\} = \delta_{ru}\delta_{sv}\frac{1}{N_{rs}}\sigma_g^2(\overline{P}_{rs}). \qquad (3.170)$$

So the noise propagation through (3.169) gives, exploiting (3.164) and (3.170),

$$\sigma^2(\tau_{\ell m}) = \left(\frac{GM}{\overline{R}^2}\right)^{-2}(\ell-1)^{-2}\frac{1}{16\pi^2}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})^2\sigma_g^2(\overline{P}_{rs})\frac{|B_{rs}|^2}{N_{rs}}$$
$$= \left(\frac{GM}{\overline{R}^2}\right)^{-2}(\ell-1)^{-2}\frac{1}{16\pi^2}\Sigma_{rs} Y_{\ell m}(\overline{P}_{rs})^2\sigma_g^2(\overline{P}_{rs})\frac{|B_{rs}|}{\mu(\overline{P}_{rs})}$$
$$\cong \left(\frac{GM}{\overline{R}^2}\right)^{-2}(\ell-1)^{-2}\frac{1}{16\pi^2}\int Y_{\ell m}(P)^2\frac{\sigma_g^2(P)}{\mu(P)}d\sigma. \qquad (3.171)$$

The expression (3.171), though rough, provides a quite comfortable formula for the approximate computation of the estimation error variances and therefore of the commission error (3.160).

Example 5. Let us see how (3.171) and (3.160) work in a quite simplified case. For instance assume that one has 10^6 point free air anomalies, uniformly distributed on the sphere, with a constant noise

$$\sigma_g(P) = \sigma_g = 5\,\text{mGal}.$$

Note that in this case $\mu(P)$ is constant too, namely, from (3.165),

$$\mu = \frac{N_{\text{tot}}}{4\pi} = \frac{10^6}{4\pi}.$$

With these values we find in (3.171)

$$\begin{aligned}\sigma^2(\tau_{\ell m}) &= \left(\frac{GM}{\overline{R}^2}\right)^{-2}(\ell-1)^{-2}\frac{\sigma_g^2}{N_{\text{tot}}}\frac{1}{4\pi}\int {Y_{\ell m}}^2 d\sigma \\ &= \left(\frac{GM}{\overline{R}^2}\right)^{-2}(\ell-1)^{-2}\frac{\sigma_g^2}{N_{\text{tot}}}.\end{aligned}$$

Using this estimate in (3.160) we receive

$$C\mathcal{E}^2 = \frac{\overline{R}^2\sigma_g^2}{N_{\text{tot}}}\sum_{\ell=L}^{M}\frac{2\ell+1}{(\ell-1)^2} \cong \frac{\overline{R}^2\sigma_g^2}{N_{\text{tot}}}2\log\frac{M}{L}.$$

In this formula we have exploited the approximation

$$\sum_{L}^{M}\frac{1}{\ell} \sim \log\frac{M}{L}$$

which, in the useful range of L and M, is good to better than 3%.

So if we assume that $L = 25$ and $M = 360$ we get

$$C\mathcal{E} = 2.3\cdot 10^{-3}\overline{R}\sigma_g.$$

In order to make this number readable we transform, roughly, the commission error of the anomalous potential $C\mathcal{E}(T)$ into a commission error in geoid

$$C\mathcal{E}(N) \cong \frac{C\mathcal{E}(T)}{\gamma_0}$$

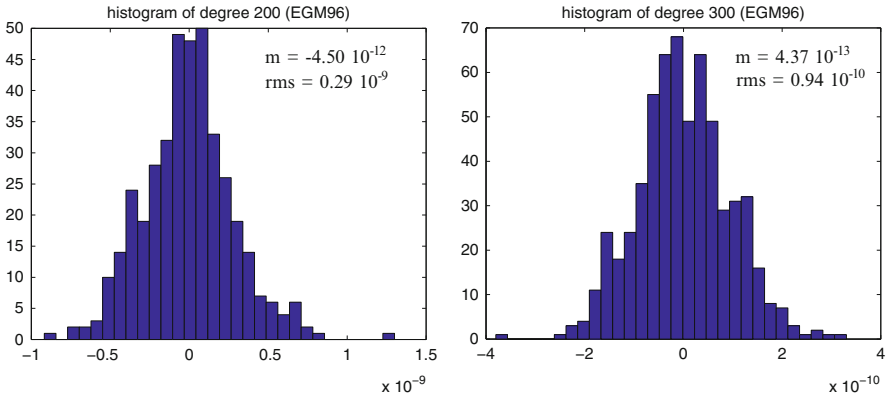


Fig. 3.5 Histograms of harmonic coefficients $\{T_{\ell m}\}$ of EGM96 (**a**) for degree 200, (**b**) for degree 300

when $\gamma_0 \sim 10^3$ Gal, i.e.

$$C\mathcal{E}(N) = 11.5 \cdot 10^{-9}\overline{R} \cong 7.4\,\text{cm}$$

which seems quite a sensible result.

(b) *Guess omission error function (Kaula's rule)*

In order to evaluate the function $O\mathcal{E}_M(T)$ given by (3.161) we would need to know $T_{\ell m}$, for all orders $\ell > M$. This is indeed not possible, so we have to give a guess for $O\mathcal{E}_M$ and this can be based for instance on a simple statistical reasoning. We first of all observe that the estimated values of $T_{\ell m}$ display quite a regular statistical behaviour when the degree increases. For instance, if we take the histograms of the coefficients of degree 200 and 300 of the global model EGM96 (cf. Lemoine et al. 1998), we find the bell-shaped figures plotted in Fig. 3.5.

As we can see the distribution shows a remarkable regularity and we could say that the coefficients are of the order of $0.3 \cdot 10^{-9}$ for degree 200 and 10^{-10} for degree 300.

The idea is now that, although the individual estimated $\widehat{T}_{\ell m}$ do contain a variable part due to the estimation error, in reality the r.m.s degree per degree which is computed from hundreds of coefficients, is quite stable and reliable.

For this and other reasons, to be discussed more in depth in Sect. A.1, the concept of *degree variances* has been introduced in geodesy; this is defined as

$$\sigma_\ell^2(T) = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} {T_{\ell m}}^2. \tag{3.172}$$

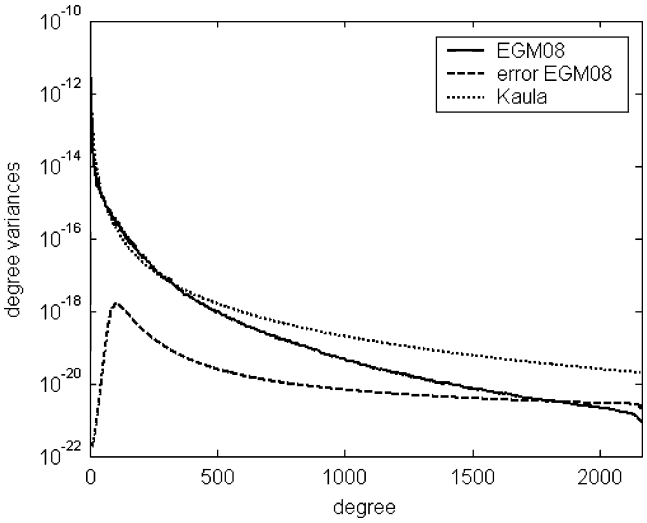


Fig. 3.6 Degree variances of EGM08 (*full line*), EGM08 error degree variances (*dashed line*), Kaula's rule (*dotted line*)

To be precise (3.172) are defined as variances of the individual coefficients of degree ℓ; close to this is the concept of *full power degree variances*

$$\overline{\sigma}_\ell^2 = (2\ell+1)\sigma_\ell^2 = \sum_{m=-\ell}^{\ell} T_{\ell m}{}^2. \tag{3.173}$$

We note that $\sigma_\ell^2(T)$ represent the mean squared L^2 norm of an individual harmonic in degree ℓ, while $\overline{\sigma}_\ell^2(T)$ represent the squared L^2 norm of the whole degree ℓ; whence the names.

It is interesting to plot $\overline{\sigma}_\ell^2$ against ℓ for instance for the most recent model EGM08 (Pavlis et al. 2008); this is plotted in Fig. 3.6. As we can see again we find quite a regular pattern of this function and Kaula, with much less knowledge of $\overline{\sigma}_\ell^2$ than today, has proposed a simple analytical law, nowadays known as *Kaula's rule* (see Kaula 2000), to express such degree variances apart from the tiny irregularity visible in Fig. 3.6,

$$\sigma_\ell(T) = \frac{10^{-5}}{\ell^2}$$

or

$$\overline{\sigma}_\ell^2(T) = \frac{(2\ell+1)}{\ell^4} 10^{-10}, \text{(Kaula's rule)}. \tag{3.174}$$

Those values of $\overline{\sigma}_\ell^2(T)$ are plotted in Fig. 3.6 as a *dotted line*; as we can see it seems that it gives a reasonable interpolation of the empirical values for medium degrees, although there is a clear misfit at degrees higher than 300, indicating that the decay of (3.174) is sensibly slower than the true one. This is an important point because even from the theoretical point of view the law (3.174) is not satisfactory. In fact, since

$$\Delta g_{\ell m} = (\ell - 1) T_{\ell m} \tag{3.175}$$

we find with (3.174)

$$\overline{\sigma}_\ell^2(\Delta g) = \sum_{m=-\ell}^{\ell} \Delta g_{\ell m}^2 = (\ell - 1)^2 \frac{10^{-10}}{\ell^4}(2\ell + 1). \tag{3.176}$$

As we can see from (3.176) we would have $\overline{\sigma}_\ell^2(\Delta g) = O\left(\frac{1}{\ell}\right)$ implying that

$$\begin{aligned} \|\Delta g\|_{L^2(\overline{S})}^2 &= \left(\frac{GM}{\overline{R}}\right)^2 \sum_{\ell=2}^{+\infty} \sum_{m=-\ell}^{\ell} \Delta g_{\ell m}^2 \\ &= \left(\frac{GM}{\overline{R}}\right)^2 \sum_{\ell=2}^{+\infty} \overline{\sigma}_\ell^2(\Delta g) = +\infty. \end{aligned} \tag{3.177}$$

This is not complying with our models requiring that Δg at the boundary has to be at least square integrable. So if we return to Fig. 3.6 we could think that $\overline{\sigma}_\ell(T)$ could be interpolated with some function of ℓ that converges to zero more rapidly.

In Fig. 3.7 and in Fig. 3.8 we display an improved version of Kaula's rule of the form

$$\overline{\sigma}_\ell^2 = \frac{3.9 \cdot 10^{-8}(0.999443)^\ell}{(\ell - 1)(\ell - 2)(\ell + 4)(\ell + 17)}. \tag{3.178}$$

This in turn is a slight generalization of a Tscherning-Rapp model, also displayed in Fig. 3.8, that we will discuss in detail in Chap. 5. The interpolation is here performed between degrees 180 and 1,800. By the way, by using Kaula's rule into the formula for the omission error for the geoid we find

$$\begin{aligned} O\mathcal{E}_M(N) &= \frac{O\mathcal{E}_M(T)}{\gamma_0} = \overline{R}\left(10^{-10} \sum_{\ell=M+1}^{+\infty} \frac{2\ell + 1}{\ell^4}\right)^{(1/2)} \\ &\cong \overline{R} \cdot \frac{10^{-5}}{M + 1} \quad \text{(Kaula's rule)}. \end{aligned} \tag{3.179}$$

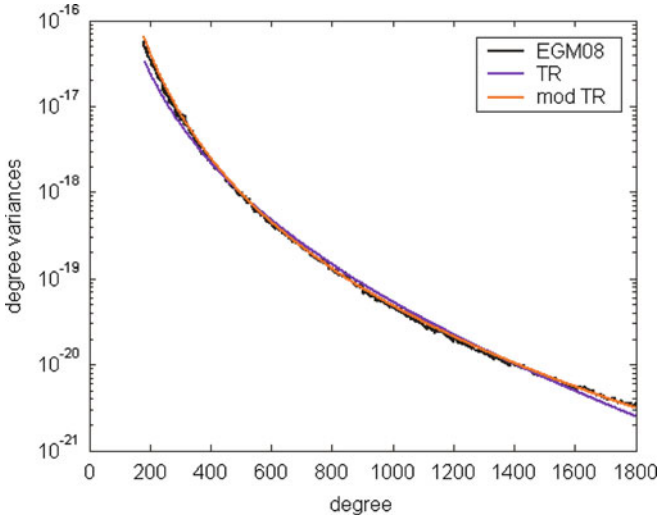


Fig. 3.7 Degree variances of EGM08 between degrees 180 and 1,800 and the best fitting curves according to the models (3.178) and (3.181)

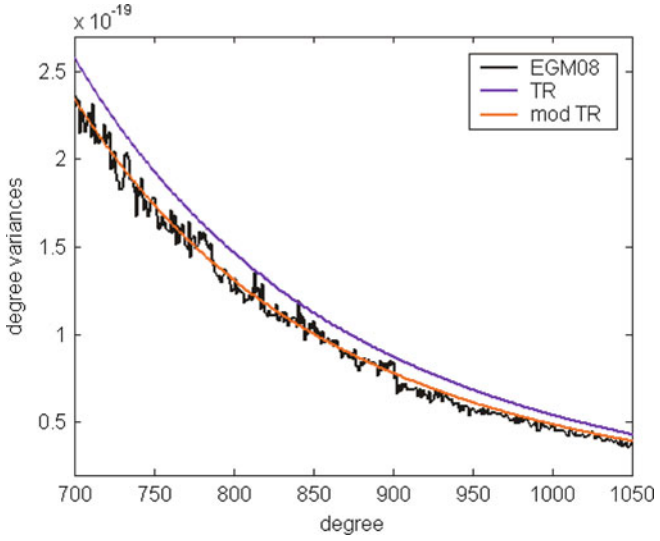


Fig. 3.8 Zoom of Fig. 3.7 TR is the Tscherning-Rapp model (3.181), while mod TR is its modified version (3.178)

To add the series in (3.179) the approximate formula $\sum_{\ell=M+1}^{+\infty} \frac{2}{\ell^3} \sim \frac{1}{(M+1)^2}$ has been used. An analogous reasoning for the improved formula (3.178) leads to

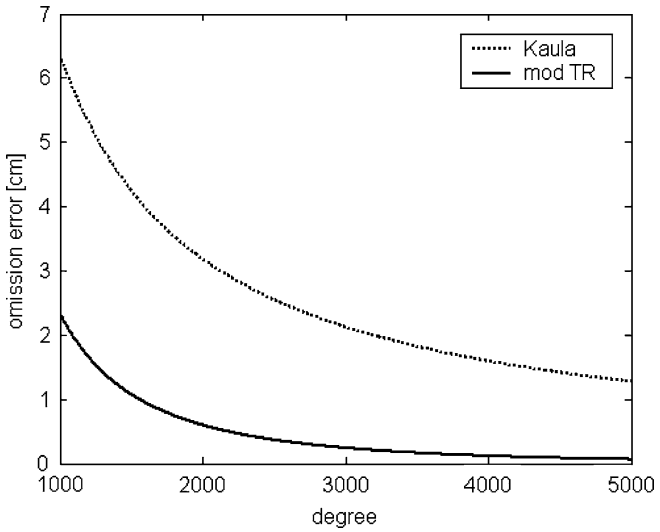


Fig. 3.9 The omission error in terms of geoid undulation according to the two laws (3.179), $O\mathcal{E}1$ (Kaula's rule) and (3.180), $O\mathcal{E}2$

$$O\mathcal{E}_M(N) = \overline{R} \cdot 1.975 \cdot 10^{-4} \left(\sum_{\ell=M+1}^{+\infty} \frac{(0.999443)^{\ell}}{(\ell-1)(\ell-2)(\ell+4)(\ell+17)} \right)^{(1/2)}. \quad (3.180)$$

The two curves (3.179) and (3.180) between degree 1,000 and degree 5,000 are plotted in Fig. 3.9, where they are denoted $O\mathcal{E}1$ and $O\mathcal{E}2$ respectively. For instance at degree 2,000 we have the two values

$$O\mathcal{E}1(2{,}000) = 3.18\,\text{cm}, \;\; O\mathcal{E}2(2{,}000) = 0.60\,\text{cm}.$$

To complete this discussion on global models and the interpolation of their degree variances with a smooth function of ℓ, we think it is useful to show the spectrum of the most recent model EGM08, with interpolations performed by means of the best fitting original Tscherning-Rapp model

$$\overline{\sigma}_{\ell}^2 = \frac{2.8 \cdot 10^{-10}(0.998365)^{\ell}}{(\ell-1)(\ell+2)(\ell+4)} \quad (3.181)$$

and by the improved model (3.178). As one can see from Fig. 3.7, they both perform very well, although there is a certain improvement in using (3.178) with respect to (3.181), as one can better appreciate in Fig. 3.8.

3.9 Exercises

Exercise 1. Prove that the following identities (see (3.24), (3.36)) hold

$$(1+s^2-2st)D_sG(s,t) \equiv (t-s)G(s,t)$$
$$sG(s,t)+2s^2D_sG(s,t) \equiv (1-s^2)D_tG(s,t),$$

where $G(s,t)$ is a Legendre polynomials generating function

$$G(s,t) = \frac{1}{\sqrt{1+s^2-2st}}$$

Exercise 2. In order to compute the Hotine function one needs the term $H_2(s,t)$ (see Example 2). After observing that $H_2(0,t) \equiv 0$ (cf. (3.97)), verify that

$$H_2(s,t) = \int_0^s G(\sigma,t)d\sigma = \log\frac{s-t+G^{-1}(s,t)}{1-t}$$

and that this coincides with the log term in (3.92).

Exercise 3. By using the arguments of Sect. 3.2, prove that the gravitational potential generated by a body B with mass density $\rho(Q)$, outside a Brillouin sphere can be put into form

$$T(P) = T_0(P)+T_1(P)+T_2(P)+O\left(\frac{1}{r_P^4}\right) =$$
$$= \frac{GM}{r_P}+\frac{GM\mathbf{r}_P^t\mathbf{b}}{r_P^3}+\frac{GM}{r_P^5}\left\{\frac{3}{2}\mathbf{r}_P^t I\,\mathbf{r}_P-\frac{1}{2}(TrI)r_P^2\right\}+O\left(\frac{1}{r_P^4}\right)$$

where

$$M = \int_B \rho(q)db_Q,\ \mathbf{b} = \frac{1}{M}\int_B \mathbf{r}_Q\rho(Q)dB_Q,$$

and I, the tensor of the moment of inertia, is given by

$$I = \frac{1}{M}\int_B \mathbf{r}_Q\mathbf{r}_Q^t\rho(Q)dB_Q\ ;$$

here $\mathbf{r}_P = [x_P, y_P, z_P]^t$ and similarly $\mathbf{r}_Q$.

Exercise 4. In order to compute the Stokes function one needs the term $S_2(s,t)$ (see Example 3). Observing that, according to its definition, $\frac{1}{s^2}S_2(s,t) \to 0$ when $s \to 0$, prove that

$$\frac{1}{s^2}S_2(s,t) = \int_0^s \frac{1}{\sigma^2}[G(\sigma,t) - 1 - \sigma t]d\sigma$$
$$= \frac{1}{s} - \frac{G^{-1}(s.t)}{s} - t - t\log\frac{1 - st + G^{-1}(s,t)}{2}.$$

(**Hint:** use the change of variable $\frac{1}{\sigma} = \tau$ and note that

$$\int \frac{1}{\sigma^2}G(\sigma,t)d\sigma = -\int \tau G(\tau,t)d\tau = -\int(\tau - t)G(\tau,t)d\tau +$$
$$-t\int G(\tau,t)d\tau,$$

which is then easy to integrate, using also the result of Exercises 1 and 2. Note that the constant in the indefinite integral has to be assigned in such a way that $\frac{1}{s^2}S_2(s,t) \to 0$ for $s \to 0$).

Exercise 5. By using the formulas for Example 5 for the commission error, in terms of geoid undulation, and (3.179) and (3.180) for the omission error, compute tentatively the total estimation error for a model with $M = 600$; $L = 2$; $N_{\text{tot}} = 4 \cdot 10^6$ number of available gravity anomalies, uniformly distributed, $\sigma_g = 5\,\text{mGal}$.

Note that at $M = 600$ the formulas give for the two models of omission error

$$O\mathcal{E}1(600) \sim 10,6\,\text{cm},\ O\mathcal{E}2(600) \sim 3,9\,\text{cm}.$$

Verify with the formulas of example 5 that $O\mathcal{E}(N) \sim 5.5\,\text{cm}$.

Verify that with $O\mathcal{E}1$ the total error is $\mathcal{E}_{\text{tot}} \sim 11.9\,\text{cm}$, while with $O\mathcal{E}2$ is $\mathcal{E}_{\text{tot}} \sim 6.7\,\text{cm}$.

Appendix

A.1

We want to prove the formula for the reproducing property of $P_n(\cos\psi)$, (3.39).

Let us first consider the third Green identity (1.61). We take as surface S the unit sphere S_1, we take any point P in Ω, $r_P > 1$, and we write it for the function (cf. Fig. 3.10)

$$v(P) = \frac{1}{\ell_{P_0P}}\ ;\ (r_{P_0} < 1).$$

Since P_0 is an arbitrary but fixed point in the unit ball, $v(P)$ is harmonic in Ω and therefore the Green identity applies. So we have

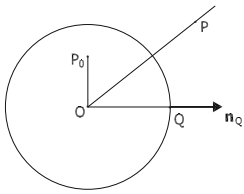


Fig. 3.10 The relative position of P_0, Q, P with respect to S_1

$$\frac{1}{\ell_{P_0P}} \equiv \frac{1}{4\pi}\int_{S_1}\left\{\frac{1}{\ell_{P_0Q}}\frac{\partial}{\partial n_Q}\frac{1}{\ell_{PQ}} - \left(\frac{\partial}{\partial n_Q}\frac{1}{\ell_{P_0Q}}\right)\frac{1}{\ell_{PQ}}\right\} dS_Q \tag{3.182}$$

where $dS_Q = d\sigma_Q$, $Q \in S$, (i.e. $r_Q = 1$). Note that $\frac{\partial}{\partial n_Q} \equiv \frac{\partial}{\partial r_Q}$ for this case. Now we can compute

$$\frac{1}{\ell_{P_0Q}} = \sum_{n=0}^{+\infty}\frac{r_{P_0}^n}{r_Q^{n+1}}P_n(\cos\psi_{P_0Q}) \tag{3.183}$$

$$\frac{\partial}{\partial r_Q}\frac{1}{\ell_{P_0Q}} = \sum_{n=0}^{+\infty} -(n+1)\frac{r_{P_0}^n}{r_Q^{n+2}}P_n(\cos\psi_{P_0Q}) \tag{3.184}$$

$$\frac{1}{\ell_{PQ}} = \sum_{n=0}^{+\infty}\frac{r_Q^n}{r_P^{n+1}}P_n(\cos\psi_{PQ}) \tag{3.185}$$

$$\frac{\partial}{\partial r_Q}\frac{1}{\ell_{PQ}} = \sum_{n=0}^{+\infty} n\frac{r_Q^{n-1}}{r_P^{n+1}}P_n(\cos\psi_{PQ}) \tag{3.186}$$

In all the above formulas we can put $r_Q = 1$ and substitute into (3.182), getting

$$\begin{aligned}\frac{1}{\ell_{P_0P}} &= \sum_{n=0}^{+\infty}\frac{r_{P_0}^n}{r_P^{n+1}}P_n(\cos\psi_{P_0P}) \qquad (3.187)\\ &= \sum_{\ell,n=0}^{+\infty}\frac{r_{P_0}^\ell}{r_P^{n+1}}(\ell+n+1)\frac{1}{4\pi}\int_{S_1}P_\ell(\cos\psi_{P_0Q})P_n(\cos\psi_{PQ})d\sigma_Q.\end{aligned}$$

Equation 3.187 has to hold $\forall r_{P_0} < 1$ and $\forall r_P > 1$; this is enough to maintain that

$$\delta_{\ell n}P_n(\cos\psi_{P_0P}) = (\ell+n+1)\frac{1}{4\pi}\int P_\ell(\cos\psi_{P_0Q})P_n(\cos\psi_{PQ})d\sigma_Q. \tag{3.188}$$

as it was to be proved.

A.2

We want to prove that spherical harmonics are L^2 orthogonal on S_1 (cf. (3.194)).

We start from (3.188) in which we substitute the summation formula (3.54); we obtain

$$\frac{\delta_{\ell n}}{(2n+1)} \sum_{m=-n}^{n} Y_{nm}(\vartheta_{P_0}, \lambda_{P_0}) Y_{nm}(\vartheta_P, \lambda_P) \tag{3.189}$$

$$= \frac{(2n+1)}{(2n+1)(2\ell+1)} \sum_{m=-n}^{n} \sum_{k=-\ell}^{\ell} Y_{nm}(\vartheta_{P_0}, \lambda_{P_0}) Y_{\ell k}(\vartheta_P, \lambda_P) \cdot$$

$$\cdot \left(\frac{1}{4\pi} \int Y_{nm}(\vartheta_Q, \lambda_Q) Y_{\ell k}(\vartheta_Q, \lambda_Q) d\sigma_Q \right)$$

Since

$$\frac{1}{4\pi} \int Y_{nm}(\vartheta, \lambda) Y_{\ell k}(\vartheta, \lambda) d\sigma = \frac{1}{4\pi} \int_0^{\pi} d\vartheta \sin\vartheta \overline{P}_{nm}(\vartheta) \overline{P}_{\ell k}(\vartheta) \cdot$$

$$\int_0^{2\pi} f_m(\lambda) f_k(\lambda) d\lambda \tag{3.190}$$

where, due to the well-known Fourier orthogonality,

$$\frac{1}{2\pi} \int_0^{2\pi} f_m(\lambda) f_k(\lambda) d\lambda = 0, m \neq k, \tag{3.191}$$

assuming $n \geq \ell$, the relation (3.189) can be written as

$$\delta_{\ell n} \sum_{m=-n}^{n} Y_{nm}(\vartheta_{P_0}, \lambda_{P_0}) Y_{nm}(\vartheta_P, \lambda_P) \tag{3.192}$$

$$= \sum_{m=-n}^{n} Y_{nm}(\vartheta_{P_0}, \lambda_{P_0}) Y_{\ell,m}(\vartheta_P, \lambda_P)$$

$$\cdot \left(\frac{1}{4\pi} \int Y_{nm}(\vartheta, \lambda) Y_{\ell m}(\vartheta, \lambda) d\sigma \right).$$

If we consider this relation as an identity in λ_{P_0}, λ_P and we further notice that in such variables (3.192) is just a Fourier's truncated series, we find that it is equivalent to

$$\frac{1}{4\pi} \int_{S_1} Y_{nm}(\vartheta, \lambda) Y_{\ell m}(\vartheta, \lambda) d\sigma = \delta_{\ell n}. \tag{3.193}$$

By combining (3.193) with (3.190) and (3.191) we finally arrive at the orthonormality relations

$$\frac{1}{4\pi}\int_{S_1} Y_{nm}(\vartheta,\lambda)Y_{\ell k}(\vartheta,\lambda)d\sigma = \delta_{\ell n}\delta_{mk}. \tag{3.194}$$

A.3

We want to justify the approximate formula (3.132). We use here the notation of Sect. 3.6.

The idea is first to perform the change of variable $q = b \cdot s$ in (3.111), so that denoting with $e'^2 = \frac{E^2}{b^2}$ the second eccentricity, we get the equation

$$(s^2 + e'^2)v''_{nm} + 2sv'_{nm} - \left[n(n+1) - \frac{e'^2 m^2}{s + e'^2}\right] v_{nm} = 0. \tag{3.195}$$

Note should be taken that in (3.195) we continue to use the notation $v' = \frac{du}{ds}$, as we did before for $\frac{d}{dq}$, however no confusion should rise for that. Next we write (3.195) in the equivalent form

$$\begin{aligned}(1 + e'^2)s^2 v''_{nm} + 2sv'_{nm} - n(n+1)v_{nm} + \frac{e'^2}{1+e'^2}m^2 v_{nm} \\ = e'^2(s^2-1)v''_{nm} + \frac{e'^2(s^2-1)}{(1+e'^2)(s^2+e'^2)}m^2 v_{nm}\,;\end{aligned} \tag{3.196}$$

it is easy to verify that (3.195) and (3.196) are one and the same equation.

However now in the left hand side of (3.196) we have a homogeneous differential operator applied to $v_{nm}(s)$, while the right hand side can be considered as a higher order perturbation. In fact note that, while we stay in the topographic layer,

$$s^2 - 1 \cong \frac{q-b}{b} \cdot 2 \le 2 \cdot 10^{-3}.$$

This means that for instance $e'^2(s^2-1)v''_{nm}$ is 10^{-5} smaller than $s^2 v''_{nm}$ and even $2\cdot 10^{-3}$ smaller than $e'^2 s^2 v''_{nm}$, so that it is natural to neglect it in a first approximation solution. A similar consideration holds for the second term in the right hand side, compared with $\frac{e'^2}{1+e'^2}m^2 v_{nm}$.

So we are reconducted now to solve the equation

$$(1 + e'^2)s^2 v''_{nm} + 2sv'_{nm} - n(n+1)v_{nm} + \frac{e'^2}{1+e'^2}m^2 v_{nm} = 0, \tag{3.197}$$

where the two differential operators $s^2\frac{d^2}{ds^2}, s\frac{d}{ds}$ are both homogeneous, similarly to what happens separating the radial variable in the spherical Laplace equation.

It is only natural then to try a solution of (3.197) in the form

$$v_{nm} = \frac{1}{s^{n+1-\alpha}}. \tag{3.198}$$

We note that in this way we automatically satisfy the conditions

$$v_{nm}(1) = 1,\ v_{nm}(s) \to 0,\ s \to \infty\,; \tag{3.199}$$

the first of (3.199) is a condition on the value of u_{nm} at the ellipsoid $\mathcal{E}$, since $s=1$ corresponds to $q = b$, while the second of (3.199), implying the regularity at infinity, is true only if

$$\alpha < n + 1. \tag{3.200}$$

It is just a matter of simple algebra to substitute (3.121) into (3.120) and find that α has to be one of the two roots

$$\alpha_{nm} = \frac{(2n+1) + e'^2(2n+3) \pm \sqrt{[(2n+1) + e'^2(2n+3)]^2 - 4e'^2(1+e'^2)a_{nm}}}{2(1+e'^2)}$$

$$(a_{nm} = (n+1)(n+2) + m^2). \tag{3.201}$$

If condition (3.200) has to be satisfied, the root with the minus sign has to be chosen. It is interesting to note that α_{nm} can be developed up to the first order in e'^2 in the form

$$\alpha_{nm} = e'^2\frac{(n+1)(n+2)+m^2}{2n+1} + O(e'^4). \tag{3.202}$$

In particular for large n and putting $m = n$ into (3.202) we see that

$$\alpha_{nm} \leq e'^2\frac{2n^2}{2n+1} \sim e'^2 n \tag{3.203}$$

showing that (3.200) is certainly satisfied for all n and m.

Finally we can further develop (3.198) considering that close to the earth surface

$$s^{\alpha_{nm}} = 1 + \alpha_{nm}(s-1) + O(s-1)^2$$

so that we get

$$v_{nm} = \frac{1}{s^{n+1-\alpha}} \cong \frac{1}{s^{n+1}}\left[1 + e'^2\frac{(n+1)(n+2)+m^2}{2n+1}(s-1)\right]. \tag{3.204}$$

A.4

In this appendix we like to introduce the convolution calculus on the sphere, because this will be used in the next chapter, particularly in the form of a moving average calculus.

We define a convolution of the function $f(P)$ with an isotropic kernel $F(\psi)$ on the sphere by means of the formula

$$g(P) = F * f = \frac{1}{4\pi} \int F(\psi_{PQ}) f(Q) d\sigma_Q. \tag{3.205}$$

Therefore, the moving average operator on a moving cap C_Δ of spherical radius Δ is a convolution with kernel

$$M(\psi) = \frac{4\pi}{C_\Delta} \vartheta_H(\Delta - \psi) = \begin{cases} \frac{4\pi}{C_\Delta} & \psi < \Delta \\ 0 & \psi > \Delta, \end{cases} \tag{3.206}$$

$\vartheta_H(t)$ being the ordinary Heavyside function, and C_Δ denoting the measure of the cap of angular radius Δ, given by

$$C_\Delta = 2\pi(1 - \cos \Delta). \tag{3.207}$$

We want to prove that if we put

$$F_n = \int_0^\pi F(\psi) P_n(\cos \psi) \sin \psi d\psi \tag{3.208}$$

then

$$g = F * f = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \left(\frac{1}{2} F_n \right) f_{nm} Y_{nm}(\vartheta_Q, \lambda_Q). \tag{3.209}$$

We start by noting that (3.208), recalling (3.46), implies

$$F(\psi) = \Sigma F_n \frac{(2n+1)}{2} P_n(\cos \psi).$$

We use the summation rule (3.54) and substitute it into (3.205) to get (3.209).

Accordingly, the moving average operator

$$M_\Delta(f) = \frac{1}{4\pi} \int M(\psi_{PQ}) f(Q) d\sigma_Q \tag{3.210}$$

with $M(\psi)$ given by (3.206) has spectral factors $\frac{1}{2} M_n$ given by

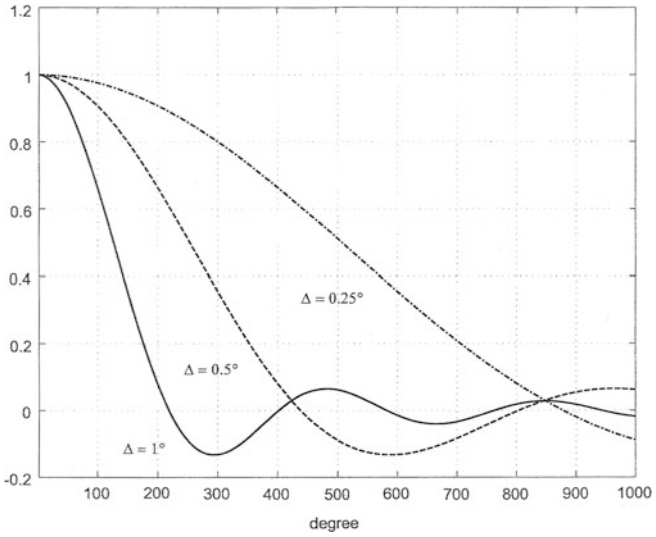


Fig. 3.11 Pellinen-Meissel's coefficients with $\Delta = 1^\circ, 0.5^\circ, 0.25^\circ$ respectively

$$\begin{aligned}\frac{1}{2}M_n(\Delta) &= \frac{1}{2}\int_o^\pi M(\psi)P_n(\cos\psi)\sin\psi d\psi \\ &= \frac{4\pi}{2C_\Delta}\int_0^\Delta P_n(\cos\psi)\sin\psi d\psi \\ &= \frac{4\pi}{2C_\Delta}\int_{\cos\Delta}^1 P_n(t)dt \\ &= \frac{1}{(1-\cos\Delta)(2n+1)}[P_{n-1}(\cos\Delta) - P_{n+1}(\cos\Delta)]\,;\end{aligned} \tag{3.211}$$

such coefficients are known in literature as *Pellinen's* or *Meissel's coefficents* (Colombo 1981a).

The key step in (3.211) is proved by using (3.37) under the integral sign. In Fig. 3.11 we represent the Pellinen-Meissel coefficients as functions of n for $\Delta = 1^\circ$, $\Delta = 0.5^\circ$ and $\Delta = 2.5^\circ$.

Chapter 4
The Local Modelling of the Gravity Field: The Terrain Effects

4.1 Outline of the Chapter

Summarizing the results of Chap. 3, we could say that up to now we have learnt how to produce an approximate anomalous potential in the form of a truncated series of spherical or ellipsoidal harmonics, namely a global potential model.

Now we focus on the other side of the spectrum of T, namely the very high-frequency components. The purposes of the Chapter are: (1) to clarify that if we want to determine the gravity field and the geoid with an appropriate spatial resolution, for instance on a 1 km by 1 km grid on the earth surface, we need then a detailed model of the geometry of the surface, i.e. a digital terrain model (DTM) with say a 100 m horizontal resolution. This because we will never be able to reach this resolution with ground gravity measurements covering the whole earth surface, while a proper DTM can be and has been derived by satellite observations, (2) to clarify that most of the high-frequency part of the potential T comes exactly from the shape of the masses modelled by the topographic surface, because high-frequency signals from internal density variations (e.g. those due to the topography of core-mantle boundary) are naturally strongly smoothed by the harmonic upward continuation, (3) to find the proper analytical computable expression of the potential due to topographic masses (on such matters one can consult Forsberg 1988, 2008, 2010).

The item (1) is discussed in Sect. 4.2 and in particular it is illustrated by means of the elementary Example 1.

The argument (2) is taken up in Sect. 4.3, where a simplified earth-like model is constructed, and it is proved that the spectrum of the potential is directly related to the shape of the topography through simple spectral relations, such as (4.16) and (4.18).

A coarse evaluation of the order of magnitude of the implied effects immediately advocates the existence of a compensation of the excess topographic masses by means of some *isostatic* mechanism. A very classical argument this, illustrated in Remark 2.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_4,

The same spectral relations recalled above make us understand however that the features of topography at all wavelengths produce in fact the corresponding features in the gravity field.

On the other hand such features at wavelengths of 20 km, or longer, are already included into the model potential described in Chap. 3.

In order to avoid counting twice the same effect, one has not to compute the whole effect of the masses above the geoid, but only that of the masses included between the actual surface S and a smoothed version of it, $\widetilde{S}$, where only long-wavelength features of S are represented. This is the so-called *residual terrain correction* discussed in Sect. 4.4.

The outcomes of this paragraph are some integral relations expressing the residual terrain corrections; in practice these integrals have to be discretized in some way to pass to a numerical implementation. The item is discussed in Sect. 4.5, where in particular the problem of the distance at which the integration has to be performed is highlighted, and different numerical procedures are discussed.

In Sect. 4.6 the formulas of Sect. 4.5 are compared with classical Bouguer formulas of full corrections to gravity anomalies, derived on the basis of a planar model, which however has not a sound theoretical basis in that it is not providing a suitable theory for the computation of the potential T. It is found though that formulas for the gravity anomalies correctly derived from an ellipsoidal set up through suitable simplifications and approximations, do coincide in the end with the classical Bouguer formulas, when we compute residual differences, explaining thus the success of the latter.

Some considerations and proposals for future research close the chapter.

4.2 High Accuracy and High Resolution Local Gravity Model

Up to now we have represented the gravity field potential as the sum of the normal and the anomalous potential, $W = U + T$, and then we have started studying T by further splitting it into a global model T_M, plus a residual part describing more local features, T_L; in formula one can write

$$W = U + T_M + T_L. \tag{4.1}$$

As we realize, this approach is, so to say, a kind of homemade multiresolution analysis; the subject has been treated in due mathematical rigour for instance in (Freeden and Schreiner 2009), though here we follow our more traditional and intuitive approach

As we have seen, T_M is expressed in terms of spherical harmonics $\{Y_{\ell m}(\vartheta, \lambda)\}$, up to some maximum degree which nowadays can be as high as 2,160. Spherical harmonics are oscillating functions bearing a certain resemblance to the Fourier's basis $(\sin nt, \cos nt)$, so that (by using the rule of thumb (3.145)), we can roughly

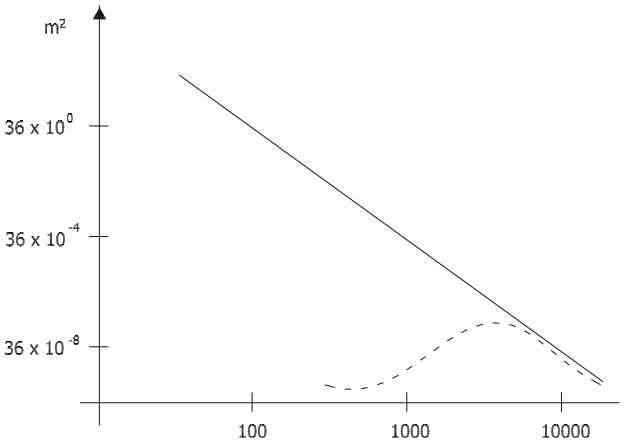


Fig. 4.1 ——— Kaula's rule spectrum for ζ, - - - - - - local component of ζ_L spectrum; abscissa degree, ordinate $\sigma_\ell^2(\zeta_L) = \frac{1}{2L+1} \sum_{m=-L}^{L} (\zeta_L)_{\ell m}^2$

state that by T_M we can describe the mean behaviour of T, down to a wavelength of 20 km. Yet, as pointed out in Sect. 3.8, this representation will be in any case affected by a *commission* error which, expressed in term of height anomaly (cf. Example 5), easily amounts to 10 cm or more; in fact it is only thanks to the most recent gravity satellite missions that one can hope to reach the level of commission error of 1 cm up to degree 200. Furthermore, by exploiting the very rough Kaula's rule estimate, one can see that over degree 200 one has, always in terms of height anomaly, a mean square omission error of 30 cm, meaning that in critical areas one can easily expect 1 m or even more of omission error. In fact, due to the specific constitution of the earth, there are areas, small if compared to the whole surface of the globe, where however we have intense gravity variations in relation to specific geophysical features like plate boundaries, mountain belts etc.

So we would expect the component $\zeta_L = \frac{T_L}{\gamma} = \frac{W-U-T_M}{\gamma}$ to reach locally the amplitude of meters, with a power spectrum mostly energetic at wavelengths between 200 km down to 2 km. A qualitative spectral behaviour of ζ_L is displayed in Fig. 4.1

To fix the ideas we shall say that we have a high-accuracy local gravity field model T_L if we can determine the corresponding height anomaly ζ_L with an error of the order of 1 cm, at any point in a given local area. If we consider that the quasi-geoid can have irregular oscillations of several centimeters over a distance of very few kilometers, we understand that a grid representation with a resolution high enough to comply with the above requirements, has to have a side of the order of 1 km.

In an equivalent way one could state that we want to be able to predict a gravity anomaly Δg with an accuracy of 1 mGal, on a grid of 1 km side or finer in the given local area.

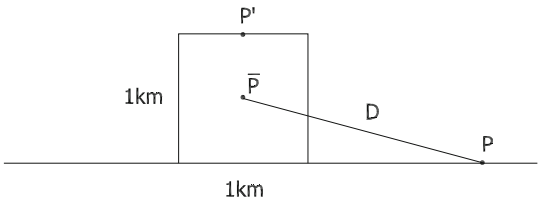


Fig. 4.2 Geometry of the gravimetric effects of a cubic mountain

This figure is derived by assuming that the main components of T and of Δg are at degree 600 (wavelength ∼40 km) and using the relations derived in Sect. 4.3.

Already at that point we can understand that with an ordinary gravity material, when Δg is observed every few kilometers (or with a much worse coverage when we are in mountaineous areas), we will never be able to build a model with the above characteristics of accuracy and resolution.

As a matter of fact, the part of the gravity signal which depends on close masses, shaped by the tiny elements of the earth surface, has to be modelled separately and used in our processing according to the remove-restore concept already illustrated in Sect. 2.5.

A simple example with a computation of orders of magnitude will be enough to convince us to go along with this necessary program.

Example 1. Assume a simple planar approximation of the reference gravity field, as it is acceptable if we move in an area of a few kilometers of radius and we just perform computations of orders of magnitude.

If one has a cubic mountain of $1 \times 1 \times 1$ km size with density $\delta_0 = 2.67\,\mathrm{g\,cm^{-3}}$ (as it is the average density in the crust), one can compute the gravimetric effects of the mountain on the geoid $\delta\zeta = \frac{\delta T}{\gamma}$ and on the gravity anomaly $\delta\Delta g$ at points P at different distances form the center $\overline{P}$ of the mountain, as in Fig. 4.2

We will use a simplified model in which all the mass of the mountain is concentrated in the barycenter, because this gives results comparable with the exact formulas already provided in Exercise 2.

So we first compute the *topographic mass* as

$$M_T = 2.67\,\mathrm{g\,cm^{-3}} \cdot 10^{15}\,\mathrm{cm^3} = 2.67 \cdot 10^{15}\,\mathrm{g}$$

and then we compute (with $G \sim 6.67 \cdot 10^{-5}\,\mathrm{mGal\,cm^2 g^{-1}}$)

$$\delta\zeta(P) = \frac{\delta T}{\gamma} \cong \frac{1}{\gamma}\frac{GM_T}{D} = \frac{1.8}{D(\mathrm{km})}\,\mathrm{cm},$$

where D is in kilometers and the result in cm. If we go similarly to the attraction, in the direction opposite to z to find the variation of the reference gravity component which is pointing downward, we get (with $H_{\overline{P}} = 0.5$ km, as in Fig. 4.2 and $H_P = 0$)

$$\delta\Delta g \cong -GM_T\frac{H_{\overline{P}}}{D^3} \cong -\frac{9}{D^3(\mathrm{km})}\,\mathrm{mGal}$$

with D in kilometers and the result in mGals.

As we can see already at $D = 2\,\text{km}$ both effects, $\delta\zeta$ and $\delta\Delta g$, are at the limit of the required prediction error; but this means also that if we have measurements every 4 km and the mountain is central with respect to measure points, we shall not feel its effect in the observations, and accordingly we shall smooth it out even when making predictions on the top of the mountain, P' in Fig. 4.2. However at P' the variations of ζ and Δg, due to the presence of the mountain, are respectively (roughly)

$$\delta\zeta = 3.6\,\text{cm}, \quad \delta\Delta g = -72\,\text{mGal},$$

(figures computed with a more precise model would be 3.2 cm and −88 mGal) neither of which is negligible.

The above example tells us that, whenever a digital model of the terrain is available, with higher resolution than the gravimetric data set, it should be independently used to account for the high frequency component of T_L, part of which would be otherwise completely lost in the local gravity field modelling, preventing us from reaching the target, e.g. the 1 cm-error level in geoid determination.

On the other hand the knowledge of the geometrical shape of the topography, in the past obtained by lengthy and costly leveling or photogrammetric operations, has now been determined by satellite borne SAR, with horizontal resolution of 100 m and an accuracy of $\sigma(H) \cong 10\,\text{m}$ (see Farr et al. 2007 or Bamler 1999).

This provides us with a lot of knowledge on the high frequency part of T_L and in general solves our data problem (see for instance the discussion in Sansò 1995).

Although we have used several times spectral arguments and in spite of the fact that the standard tool for spectral analysis of functions is the use of spherical (ellipsoidal) harmonics, yet a good representation of T_L will never use the S.H. basis. In fact these functions oscillate in very large areas on the sphere so that if we determine on the basis of local data only a S.H. coefficient different from zero, even at very high degree, the corresponding harmonic will spread the local behavior everywhere on the sphere. This is very similar to what happens with Fourier series. So, as it has been done in Fourier analysis theory, it is convenient here too to introduce suitable harmonic kernels that go to zero fast enough outside the area where we have data to avoid an improper propagation. This will be done by introducing a suitable statistical reasoning into the approximation procedure that will ultimately lead us to a solution which compares one to one to the result of Part III, that is derived on a purely analytical basis from the theory of Hilbert spaces with reproducing kernels.

As a matter of fact, many other approaches have followed the same line of thought, starting from the historical method of modifying the Stokes kernel to make it more short-tailed, dating back to Molodensky (cf. Heiskanen and Moritz 1967) arriving to the more recent multiresolution methods, employing a kind of spherical wavelets as proposed in Freeden and Schreiner (2009).

Remark 1. A special mention has to be made of the problem of removing from marine gravity data the gravimetric signal coming from the shape of the sea floor. In fact one could consider this problem as a pure topographic correction with

inverted, only negative, heights (depths) and a density which is now about $1\,\mathrm{g\,cm^{-3}}$ (the density of water) in contrast to the crust density $\delta_0 \sim 2.67\,\mathrm{g\,cm^{-3}}$.

The major difference here is that in this case we compute the bathymetric correction on the smooth base of this *inverted* topography and not at points with strongly varying depth. As a consequence, the deeper is the sea floor the smoother is its signal on the sea surface. This is good because it is difficult to have a high resolution bathymetry in deep oceans. We shall return to this problem more precisely at the end of the chapter.

Summarizing, we could say that once our gravity potential has been reduced to the local component we have first of all to regularize it as much as possible by applying a suitable correction taking into account the digital model of the terrain; we shall see in the next sections that this has always the effect of smoothing the field and how to perform (in principle) such a computation.

4.3 The Smoothing Role of Terrain Correction (TC)

In this section we shall develop a formula expressing globally the effects of the boundary S, with its height variations, on the exterior gravity field in terms of potential variations, T_t, and gravity anomaly variations, Δg_t.

The index t will be used in this section to mean *quantity related to a topographic effect*.

Calculations will be performed with a model that is realistic but not close enough to reality to consider the result as directly applicable to true data. Our purpose in fact is just to elucidate the smoothing effect of TC, not, for the moment, to find a formula appropriate for numerical implementation.

Then consider a body B composed by an inner part which is a ball B_0 bounded by a sphere S_{R_0} with radius R_0 and an outer part, a crust C, overlain on S_0 with a thickness $H(Q_0)$ always positive (see Fig. 4.3).

The mass density $\delta(Q)$ within B_0 will generate a gravity field T_0, implying a geoid of some 30 m (r.m.s.) as it is for the true earth. The mass density in C is constant, $\delta_0 = 2.67\,\mathrm{g\,cm^{-3}}$, similar to the average figure for the earth crust. Please note that in this section and in the next the symbol $H(Q_0)$ is not necessarily used for orthometric height, but rather for a function of $P_0(\vartheta, \lambda)$ that expresses analytically the height of S (radial in this case) over the sphere S_0.

The total field outside S will be

$$T = T_0 + T_C, \tag{4.2}$$

where T_C is given by the Newton integral in spherical coordinates, fixing the computation point P on an outer sphere S_R,

$$T_C(P) = G\delta_0 \int d\sigma \int_{R_0}^{R_0+H_Q} \frac{r^2 dr}{\ell_{PQ}}, \tag{4.3}$$

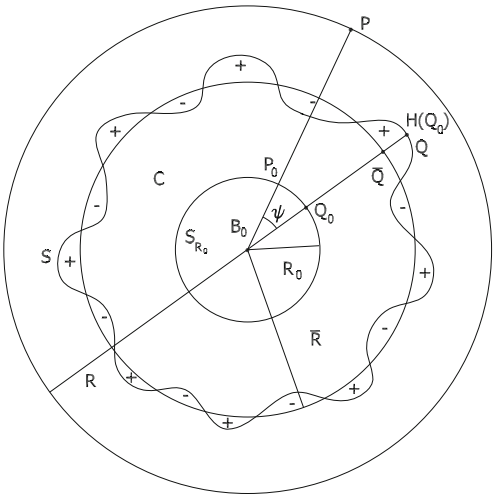


Fig. 4.3 The body $B = B_0 \cup C$: the crust surface S with height $H(Q)$ over S_{R_0} and $H(Q) - \overline{H}$ over $S_{\overline{R}}$, $(\overline{H} = \overline{R} - R_0)$; $\pm$ regions of positive and negative apparent density $\pm\delta_0$

$$\ell_{PQ} = [R^2 + r^2 - 2Rr\cos\psi_{PQ}]^{(1/2)}.$$

Note that the index C (recalling *crust* and *correction*) is substituting here, for the moment, the index t.

If we call $\overline{H}$ the mean height of S, i.e.

$$\overline{H} = \frac{1}{4\pi}\int H(\overline{Q})d\sigma_{\overline{Q}} \tag{4.4}$$

and we put

$$\delta H = H(Q) - \overline{H}, \quad \overline{R} = R_0 + \overline{H} \tag{4.5}$$

we see that the inner integral in (4.3) can be split into two

$$\int_{R_0}^{R_0+H} \frac{r^2 dr}{\ell_{PQ}} = \int_{R_0}^{\overline{R}} \frac{r^2 dr}{\ell_{PQ}} + \int_{\overline{R}}^{\overline{R}+\delta H} \frac{r^2 dr}{\ell_{PQ}}. \tag{4.6}$$

Correspondingly the potential T_C is split into two potentials, one of which is (outside $S_{\overline{R}}$) just a monopole potential

$$T_C = \overline{T} + T_t, \tag{4.7}$$

with

$$\overline{T} = \frac{G\frac{4}{3}\pi(\overline{R}^3 - R_0^3)\delta_0}{R} \tag{4.8}$$

and

$$T_t = G\delta_0 \int d\sigma \int_{\overline{R}}^{\overline{R}+\delta H} \frac{r^2 dr}{\ell_{PQ}}. \tag{4.9}$$

Note that in (4.9) δH will be positive or negative in different regions, i.e. $T_t(P)$ is the potential of a body C_t with apparent density $+\delta_0$ in some regions and $-\delta_0$ in others (see Fig. 4.3).

So the effect of the shape of S goes into an average term (4.8) and an oscillating term (4.9). We want to study T_t as function of δH and to do that we *linearize* the internal integral in (4.9) with the approximation

$$\int_{\overline{R}}^{\overline{R}+\delta H} \frac{r^2 dr}{\ell_{PQ}} \cong \frac{\overline{R}^2}{\ell_{P\overline{Q}}} \delta H(\overline{Q}), \tag{4.10}$$

which is nothing but a Taylor expansion stopped at the first order in δH.

We note that in this way we have substituted the exact expression (4.9)

$$T_t(P) = G \int_{C_t} \frac{\delta_a(Q)}{\ell_{PQ}} dC,$$

$$\delta_a(Q) = \delta_0 \cdot \begin{cases} +1 & \text{when } \delta H > 0 \\ -1 & \text{when } \delta H < 0, \end{cases} \tag{4.11}$$

with the other expression

$$T_t(P) \cong G\delta_0 \int_\sigma \frac{\delta H(\overline{Q})}{\ell_{P\overline{Q}}} \overline{R}^2 d\sigma \tag{4.12}$$

representing the potential of a single layer on $\overline{S}$, with surface density $\delta_0 \delta H(\overline{Q})$; i.e. we have squeezed the mass column of base $d\overline{S}$ and height δH onto $d\overline{S}$, and for that reason the approximation (4.12) is also known in literature as *coating method* (see Heiskanen and Moritz 1967).

Now, since $R > \overline{R}$, we can use in (4.12) the development

$$\frac{1}{\ell_{P\overline{Q}}} = \frac{1}{\overline{R}} \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \left(\frac{\overline{R}}{R}\right)^{n+1} (2n+1)^{-1} Y_{nm}(\vartheta_P, \lambda_P) Y_{nm}(\vartheta_{\overline{Q}}, \lambda_{\overline{Q}}) \tag{4.13}$$

and putting

$$\delta H_{nm} = \frac{1}{4\pi} \int_\sigma \delta H(\vartheta, \lambda) Y_{nm}(\vartheta, \lambda) d\sigma \tag{4.14}$$

we get

$$T_t(P) \cong 4\pi G\delta_0 \overline{R} \sum_{n=1}^{+\infty} \sum_{m=-n}^{n} \left(\frac{\overline{R}}{R}\right)^{n+1} \delta H_{nm}(2n+1)^{-1} Y_{nm}(\vartheta_P, \lambda_P). \quad (4.15)$$

Note that in (4.15) the summation on n starts from 1 because $\delta H_{00} = 0$ as a consequence of (4.4) and (4.5).

In spectral form (4.15) writes

$$(T_t)_{nm} = (4\pi G\delta_0 \overline{R}) \left(\frac{\overline{R}}{R}\right)^{n+1} \frac{\delta H_{nm}}{2n+1}; \quad (4.16)$$

here we recognize that with the coating approximation every coefficient δH_{nm} of the topographic height function δH, is upward continued by the factor $\left(\frac{\overline{R}}{R}\right)^{n+1}$ and smoothed by the typical effect of Newton's kernel $(2n+1)^{-1}$. But basically to every δH_{nm} corresponds a $(T_t)_{nm}$. Even more interesting is to compute from (4.15) the corresponding gravity anomaly $\Delta g_t(P)$ in spherical approximation with the formula (cf. (2.100))

$$\begin{aligned} \Delta g_t &= -\frac{\partial T_t}{\partial R} - \frac{2}{R} T_t \quad (4.17) \\ &= 4\pi G\delta_0 \sum_{n=1}^{+\infty} \sum_{m=-n}^{n} \left(\frac{\overline{R}}{R}\right)^{n+2} \delta H_{nm} \frac{n-1}{2n+1} Y_{nm}(\vartheta_p, \lambda_p). \end{aligned}$$

Since we are interested in the effect of the topography for medium-high degrees, (e.g. $n > 90$) where by the way $\{\delta H_{nm}\}$ become more important, we can make the further approximation

$$\frac{n-1}{2n+1} \sim \frac{1}{2}$$

to get the approximate spectral relation

$$(\Delta g_t)_{nm} \cong 2\pi G\delta_0 \left(\frac{\overline{R}}{R}\right)^{n+2} \delta H_{nm} \quad (4.18)$$

From (4.18) we read that, apart from a constant and the usual upward continuation factor (which for gravity anomalies is $\left(\frac{\overline{R}}{R}\right)^{n+2}$), every δH_{nm} is translated into a corresponding $(\Delta g_t)_{nm}$.

Therefore if we have a field $\Delta g(P)$ observed on S_R and we subtract from it Δg_t we are left with $\Delta g_0 + \Delta \overline{g}$, i.e. the anomalies corresponding to $T_0 + \overline{T}$ (cf. (4.7));

but $\Delta\overline{g}$ is constant and Δg_0 is much smoother than Δg_t because it comes from a lower level $R_0 < \overline{R}$ and therefore it will be smoothed by a factor $\left(\frac{R_0}{R}\right)^{n+2}$ instead of $\left(\frac{\overline{R}}{R}\right)^{n+2}$.

In other words by subtracting Δg_t from Δg we get a smoothed outer field of gravity anomalies.

Remark 2. Take for a moment $R = \overline{R}$ in (4.18), i.e. let us disregard the upward continuation effect, to evaluate the mean square value of Δg_t at the mean level of the topography $\overline{R}$. We find

$$\sigma(\Delta g_t) = \left\{\Sigma_{n,m}\Delta g_{t^{nm}}^2\right\}^{(1/2)} = 2\pi G\delta_0\sigma(\delta H);$$

assuming a root mean square value for δH of $\sigma(\delta H) = 300\,\mathrm{m}$ and with the known values of G, δ_0 we get

$$\sigma(\Delta g_t) \cong 33\,\mathrm{mGal}.$$

This figure is already equal to the root mean square value of Δg for the full anomalous potential T, including the large part coming from inner masses. This shows that for the real earth there must be some mechanism which naturally tends to damp down the variations of Δg_t. Several models to explain this have been developed in geophysics (cf. Turcotte and Schubert 2001); the simplest (dating back to the mid-nineteenth century) is probably the so-called Airy-Heiskanen isostatic system where it was supposed that the load of the topographic features higher than the mean elevation $\overline{H}$ is compensated by a hydrostatic pressure from the mantel on the crust due to a *root* of lower density intruding into the mantel and mirroring the topography.

This is schematically explained in Fig. 4.4; we see that a hydrostatic equilibrium of the column is obtained if the weight of the upper column, $\delta_0 \cdot dS \cdot \delta H \cdot g$ is compensated by an archimedean force due to the fact that the upper mantel has a mean density $\delta_m = 3.27\,\mathrm{g\,cm^{-3}}$, larger than δ_0; this force is expressed by $(\delta_m - \delta_0)dS\delta H_r g$, and equating the two, one gets

$$\delta H_r = \frac{\delta_0}{\delta_m - \delta_0}\delta H \cong 4.45\delta H$$

As we can see, to an excess of mass due to the mountain corresponds a defect of mass in the root that partly compensates the increase of Δg. A simple mechanism like that is not anymore accepted in modern geophysics where the dynamics of different layers is also taken into account (cf. Sabadini and Vermeersen 2004). Nevertheless it is known that in the average a certain compensation is in fact realized by the body of the earth explaining the actual mean amplitude of the gravity anomalies.

In addition one has to consider that most of these effects have a long-wavelength character and therefore are accounted for by a global model (Sünkel and Tscherning 1981).

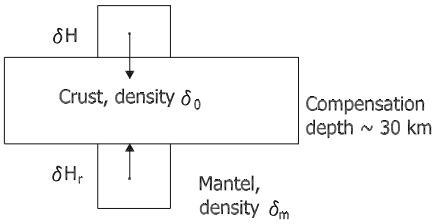


Fig. 4.4 A schematic view of the isostatic compensation mechanism of Hieskanen-Airy

Summarizing we could say that the smoothing effect of terrain correction has been demonstrated, though since we want to apply this to data observed directly on the earth surface S and not on an outer sphere, we still have to work out formulas appropriate to the numerical implementation.

4.4 From Terrain Correction (TC) to Residual Terrain Correction (RTC)

In order to let the example of Sect. 4.3 to become realistic and applicable, we need first of all to substitute the sphere S_0 with the earth ellipsoid $\mathcal{E}$, and we have to express the Newton integral for the layer (cf. Fig. 4.5)

$$C \equiv \{0 \leq h_Q \leq H(Q_0)\}, \tag{4.19}$$

surrounded by the surface $S \equiv \{h = H(Q_0)\}$.

Let us mention here that a lot of work has been done in geodesy on similar items, also in different contexts like the discussion of Helmert method. We refer here for instance to Heck (2003b) and to Sjöberg (2000) as well as to and Martinec (1998) and Sünkel (1986).

To be precise, from the geophysical point of view C is not *the crust*, but just part of it, since the crust is extending below $\mathcal{E}$, down to the Mohorovicic discontinuity (see Table 3.1).

Please note that in this section we are still using $H(Q_0) = H(\vartheta, \lambda)$ as a function defined on $\mathcal{E}$, expressing the ellipsoidal height of S; so, with reference to Fig. 4.5, we can write $h_{Q'} \equiv H(Q_0)$.

Then, in terms of anomalous potential T, the terrain contribution of C reads simply

$$T_C(P) = G \int_C \frac{\delta(Q)}{\ell_{PQ}} dV_Q. \tag{4.20}$$

We are going to apply a number of approximations to (4.20), some of which are in the range of a relative error of $10^{-2}/10^{-3}$. Since, as already our elementary Example 1 has shown, we can easily have a terrain perturbation in the gravity anomaly larger than 100 mGal, approximations at the 10^{-2} level would not be acceptable according to our criteria.

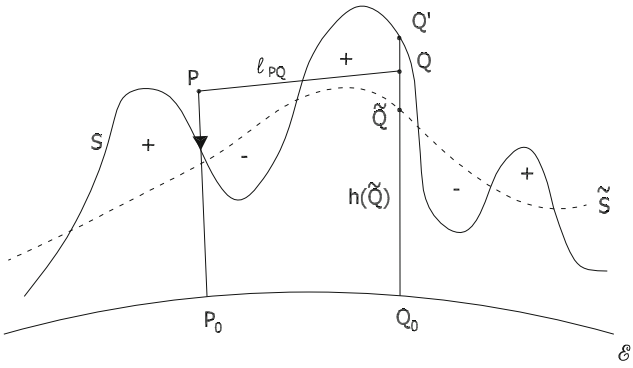


Fig. 4.5 S is the earth surface with equation $h_{Q'} = H(Q_0)$; $\widetilde{S}$ is a *smoothed* surface with equation $h_{\widetilde{Q}} = \widetilde{H}(Q_0)$; the *arrow* on P means that we first compute potential and gravity at any point P outside S, then we let P go to S along the normal of $\mathcal{E}$

So we have to *reduce* the size of T_C before we start our processing.

In Sect. 4.3 we have subtracted to H its global mean value $\overline{H}$, among other things because in this way we do not introduce a coefficient $T_{00} \neq 0$, which would contradict our definition of anomalous potential. Here however we understand that a global value $\overline{H}$ is usually not efficient in reducing T_C in a local area; on the contrary, $\overline{H}$ would give rise to a $\overline{T}$ that in most cases would appear locally as a bias. This leads us to the idea of introducing a kind of local mean height surface $\widetilde{S}$, with equation $h_{\widetilde{Q}} = \widetilde{H}(Q_0)$, enjoying the following properties:

1. $\widetilde{H}$ should be smooth in the sense that, for instance, by developing it in spherical harmonics we should find that contributions above a threshold degree N_C are negligible, e.g. one should have

$$\left\{ \sum_{n=N_C+1}^{+\infty} \sum_{m=-n}^{n} \widetilde{H}_{nm}^2 \right\}^{(1/2)} < 10\,\text{m}, \tag{4.21}$$

2. It should be *local*, i.e. it should be given, together with $H(Q_0)$, on an area which is larger, but not too much, than the area where we want to model the gravity field; for instance if we choose $N_C = 360$ we could go as far as one wavelength at degree 360, namely 110 km, outside the area where we want to make computations,
3. It should properly interpolate H, in the local area; this means as a minimum that, if we call $\widetilde{C}$ the body enclosed by $\widetilde{H}$, one should have (cf. Fig. 4.5)

$$\int_C \delta(Q)dV - \int_{\widetilde{C}} \delta(Q)dV = \int_{C \div \widetilde{C}} \delta_a(Q)dV = 0 \tag{4.22}$$

$$\left(\begin{array}{l}\delta_a(Q) = \text{ apparent density } = \left\{\begin{array}{l}+\delta(Q) \text{ when } H > \widetilde{H}\\ -\delta(Q) \text{ when } H < \widetilde{H}\end{array}\right.\\ C \div \widetilde{C} = \text{symmetric difference} = (C\backslash\widetilde{C}) \cup (\widetilde{C}\backslash C)\end{array}\right)$$

In practice a good choice of $\widetilde{H}$ could be a moving average of H, over a disk of radius Δ comparable to the long wavelength from which we start disregarding the terrain effects, i.e. considering in this case, for a rough computation, $\mathcal{E}$ as a sphere,

$$\widetilde{H}(Q_0) = M_\Delta\{H\} = \frac{1}{S_\Delta}\int_{S_\Delta} H(Q)dS_Q, \tag{4.23}$$
$$S_\Delta \equiv \{Q;\ \psi_{QQ_0} \leq \Delta\}.$$

Now think of the gravimetric effect of $\widetilde{C}$; in particular we claim that if $\widetilde{S}$ satisfies the above conditions, the effect of $\widetilde{C}$ above degree N_C is negligible.

To see that, we can again use formulas (4.16) and (4.17), with $R = \overline{R}$. If we choose for instance $N_C = 360$ and $\widetilde{H}$ satisfies (4.21) we immediately find that the high frequency contribution (above degree N_C) of $\widetilde{H}$ is smaller than 1.6 cm in geoid and 1.1 mGal in gravity anomaly.

Taking into account that we shall further apply an approximation procedure to what remains at a local level, these numbers are completely acceptable. Based on this remark we can decide that the local high frequency component of the gravity field due to terrain effects can be accounted for as the difference between the effect of C and that of $\widetilde{C}$. We call this the *residual terrain correction* and we put

$$\begin{aligned}T_{RC} &= T_C - T_{\widetilde{C}} = G\int_C \frac{\delta(Q)}{\ell_{PQ}}dV - G\int_{\widetilde{C}} \frac{\delta(Q)}{\ell_{PQ}}dV\\ &= G\int\int_{\widetilde{H}}^{H} \frac{\delta(Q)}{\ell_{PQ}}dSdh = G\int_{C\div\widetilde{C}} \frac{\delta_a(Q)}{\ell_{PQ}}dV;\end{aligned} \tag{4.24}$$

for the sake of brevity we have introduced in the last integral, as in (4.23) the symmetric difference $C \div \widetilde{C} = (C\backslash\widetilde{C}) \cup (\widetilde{C}\backslash C)$. Let us notice that in (4.24) one has $dV = \pm dSdh$ according to whether dh is positive or negative, i.e. H is larger or smaller than $\widetilde{H}$. This is the reason why, if we want to write T_{RC} as a volume integral (last term in (4.24)), one has to introduce the apparent density

$$\delta_a(Q) = \left\{\begin{array}{l}\delta(Q),\ H > \widetilde{H}\\ -\delta(Q),\ H < \widetilde{H}.\end{array}\right. \tag{4.25}$$

The first big advantage of going from T_C to T_{RC} is that now the size of T_{RC} is quite significantly reduced; even in mountainous areas, typical for T_{RC}/γ is a figure of some decimeters and for the corresponding Δg_{RC} of $10 \sim 20$ mGals. As a first consequence we are allowed to compute T_{RC} in spherical approximation, namely, from (4.24) we can write

$$T_{RC}(P) = G \int d\sigma \int_{\widetilde{H}}^{H} \frac{\delta(Q)(R_0 + h)^2}{\ell_{PQ}} dh. \tag{4.26}$$

Furthermore, since

$$(R_0 + h)^2 = R_0^2 \left(1 + 2\frac{h}{R_0} + \frac{h^2}{R_0^2}\right)$$

we can substitute R_0^2 instead of $(R_0 + h)^2$ with an error smaller than $3 \cdot 10^{-3}$ (remember that everywhere on the earth surface $\frac{h}{R_0} < 1.5 \cdot 10^{-3}$ and, apart from a very tiny portion of the surface, it is $h < 10^{-3} R_0$). In addition, in the volume between $\widetilde{H}$ and H, it is much more realistic to assume that $\delta(Q) = \delta_0 \cong 2.67\,\mathrm{g\,cm^{-3}}$, than for the whole column going down to $\mathcal{E}$, in particular in mountaineous areas.

This is also useful to clarify what density one should use to fill in the holes when $\widetilde{H} > H$ (see regions tagged with $-$ in Fig. 4.5). So (4.26) becomes

$$T_{RC}(P) = G\delta_0 R_0^2 \int d\sigma \int_{\widetilde{H}}^{H} \frac{dh}{\ell_{PQ}}. \tag{4.27}$$

Now we elaborate on the Newton kernel $1/\ell_{PQ}$.

We first note that the identity

$$\ell_{PQ} \equiv [(r_P - r_Q)^2 + 2r_P r_Q (1 - \cos\psi)]^{(1/2)} \tag{4.28}$$

holds. Then, since $r = R_0 + h$, we can write

$$r_P - r_Q = h_P - h_Q$$

and, in addition

$$r_P r_Q \cdot 2(1 - \cos\psi) \cong r_P r_Q \psi^2 = r_P r_Q \left(\frac{D_{0PQ}}{R_0}\right)^2 \cong D_{0PQ}^2 \tag{4.29}$$

where we have put (referring to Fig. 4.5)

$$D_{0PQ} = R_0 \psi \cong D_{P_0 Q_0}.$$

Going back to (4.28) we find

$$\ell_{PQ} \cong \left[D_{0PQ}^2 + (h_P - h_Q)^2\right]^{(1/2)} = \ell_{0PQ} \tag{4.30}$$

and (4.27) can be written as

$$T_{RC} = G\delta_0 R_0^2 \int d\sigma \int_{\widetilde{H}}^{H} \frac{dh}{\ell_{0PQ}}. \tag{4.31}$$

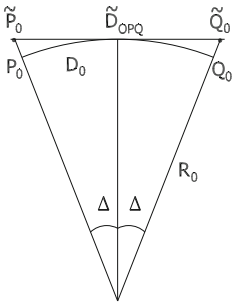


Fig. 4.6 Comparison of D_0 with the distance along the tangent plane $\widetilde{D}_0$

It has to be understood that in (4.31) the integral over the unit sphere, in $d\sigma$, has to be extended to the local area where we perform our computation and where we have data. As already pointed out this has to be a little larger than the one where gravity data are given. Outside this area one can imagine that $H = \widetilde{H}$ so that the inner integral vanishes.

From (4.31) we can derive the corresponding residual terrain effect to the gravity anomaly, Δg_{RC}.

To do that we observe that $\frac{\gamma'}{\gamma} T_{RC}$, when $\frac{T_{RC}}{\gamma}$ is as large as 1 m, is about 0.3 mGal, so that this term is usually neglected, and what is computed is

$$\begin{aligned} \Delta g_{RC} &= -\frac{\partial T_{RC}}{\partial h_P} \\ &= G\delta_0 R_0^2 \int d\sigma \int_{\widetilde{H}}^{H} \frac{h_P - h}{\ell_{0PQ}^3} dh. \end{aligned} \tag{4.32}$$

From spherical to planar approximation. Let us notice that computing horizontal distances on a sphere or on a tangent plane, up to an angular distance Δ from the tangence point, introduces a very small error.

In fact (cf. Fig. 4.6)

$$\begin{aligned} D_0 &= R_0 \cdot 2\Delta \\ \widetilde{D}_0 &= 2R_0 tg\Delta \cong 2R_0 \left(\Delta - \frac{1}{3}\Delta^3\right) = D_0 \left(1 - \frac{1}{3}\Delta^2\right); \end{aligned} \tag{4.33}$$

then even for a large distance with $\Delta = 10°$ we have a relative error

$$\frac{D_0 - \widetilde{D}_0}{D_0} \sim 2.6 \cdot 10^{-4},$$

which we know we can neglect in the present context.

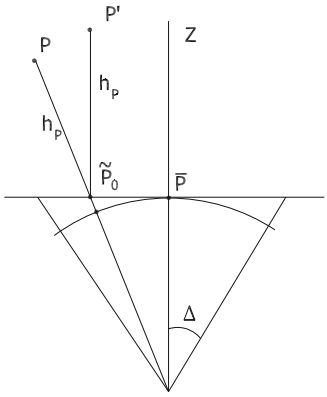


Fig. 4.7 Mapping the residual terrain correction integral on the tangent plane

The same reasoning applies to the area element $R_0^2 d\sigma$ on the sphere with respect to the corresponding area element on the tangent plane. Therefore if we put

$$\widetilde{D}_{0PQ} = D_{\widetilde{P}_0\widetilde{Q}_0}, \quad \widetilde{\ell}_{0PQ} = \left[\widetilde{D}_{0PQ}^2 + (h_P - h_Q)\right]^2$$

and we introduce a couple of Cartesian coordinates (x, y) on the tangent plane, we can rewrite (4.31) and (4.32) as

$$T_{RC} = G\delta_0 \int dS \int_{\widetilde{H}}^{H} \frac{dh}{\widetilde{\ell}_{0PQ}} \tag{4.34}$$

$$= G\delta_0 \int d\xi d\eta \int_{\widetilde{H}(\xi,\eta)}^{H(\xi,\eta)} \frac{dh}{[(\xi - x_P)^2 + (\eta - y_P)^2 + (h_P - h)^2]^{(1/2)}}$$

$$\Delta g_{RC} = G\delta_0 \int dS \int_{\widetilde{H}(Q)}^{H(Q)} \frac{(h_P - h)}{\widetilde{\ell}_{0PQ}^3} dh \tag{4.35}$$

$$= G\delta_0 \int d\xi d\eta \int_{\widetilde{H}(\xi,\eta)}^{H(\xi,\eta)} \frac{(h_P - h)}{[(\xi - x_P)^2 + (\eta - y_P)^2 + (h_P - h)^2]^{3/2}} dh.$$

As we can see (4.34) and (4.35) are purely Cartesian formulas, which, to be precise, must be used in the following way: assume you want to compute $T_{RC}, \Delta g_{RC}$ in an area laying within a disk of center $\overline{P}$ and angular radius Δ; then using a Cartesian system tangent to the sphere in $\overline{P}$ one projects all points P onto $\widetilde{P}_0$ on the tangent plane and takes the same height h_P in the Z direction (cf. Fig. 4.7).

This explains why the planar approximation, which we could derive directly in a geometry where the main part of the gravity field is parallel to the Z axis, is so widely applied and works so well.

There is no need to say that there are in literature several possibilities of implementing the computation of T_{RC} and Δg_{RC} in spherical approximation and even directly in ellipsoidal coordinates.

Remark 3. It is interesting to observe that if the digital terrain model is given in terms of orthometric heights, which we momentaneuously denote as OH_P to distinguish them from the height function $H(P_0)$ used in these sections, formulas (4.34) and (4.35) still hold with H and $\widetilde{H}$ computed from this digital terrain model. In fact, from

$$H(Q_0) = OH(Q_0) + N(Q_0)$$

we see, with obvious notation, that

$$H(Q_0) - \widetilde{H}(Q_0) = OH(Q_0) - O\widetilde{H}(Q_0) + N(Q_0) - \widetilde{N}(Q_0).$$

Since the geoid $N(Q_0)$ is already very smooth, $N(Q_0) - \widetilde{N}(Q_0)$ can amount at most to 1–2 m and therefore it can be neglected, since this figure is often comparable with the error affecting our knowledge of $H(Q_0)$.

We conclude the section by observing that in (4.34) and (4.35) we have finally to put $h_P = H(P_0)$ if we want to compute T_{RC} and Δg_{RC} at a point on the earth surface. Naturally this is not always the case, for instance when we use such formula for aerial gravimetry.

Finally it is worth to underline that T_{RC}, Δg_{RC} as expressed here are the effects of the residual topography $H - \widetilde{H}$ on T and Δg; therefore when we have to apply them for the purpose of smoothing we have to compute residual quantities as

$$\begin{aligned} T_r &= T_L - T_{RC} = T - T_M - T_{RC}, \\ \Delta g_r &= \Delta g_L - \Delta g_{RC} = \Delta g - \Delta g_M - \Delta g_{RC}. \end{aligned} \tag{4.36}$$

It is the computation of T_r from Δg_r that will occupy us in the next chapter.

Another point in favour of using RTC instead of terrain correction is that the use of the former does not imply any change in the total masses. Also the barycenter can hopefully be supposed to be little affected because we have a combination of positive and negative masses close to one another.

4.5 Strategies for the Implementation of Terrain Effects

The calculation of terrain effects is usually a significant numerical task, requiring the largest use of computer time in the remove-restore steps of the computation of a gravimetric geoid. This item is treated in greater detail in the book, Part II, Chaps. 8 and 10.

To make an example, just think that in an area of $1{,}000 \times 1{,}000$ km with a digital terrain model with 100 m of horizontal resolution, one has to implement the numerical integration of formulas (4.34) and (4.35) for $\sim 10^8$ computation points, handling 10^8 height data.

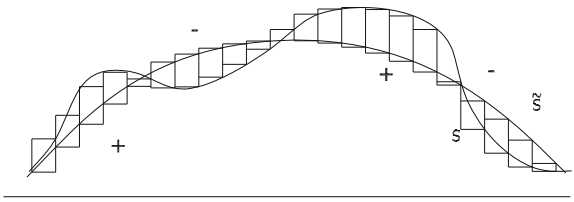


Fig. 4.8 The discretization of terrain effects by prisms with $\pm\delta_0$ apparent density

The detailed strategies for this implementation will be described in the second part of the book; here we like only to present the principles on which such implementations are based and discuss the relevant approximations.

Essentially there are two approaches to the computation of terrain effects:

(a) A simple discretization of the terrain in term of prisms and the subsequent computation of integrals as sums,
(b) The reduction through suitable approximations of (4.34) and (4.35) into two convolution integrals, which can then be efficiently computed by means of Fourier transform methods.

Here are the main points of the two approaches:

(a) The principle is absolutely clear; the residual topographic body $C \div \widetilde{C}$ is approximated by prisms; the effect of each prism in T_{RC} and Δg_{RC} is known at any point P in space and therefore we can compute our effect by adding those of each prism.

Analytical formulas for the potential of the prism have already been established in Exercise 9 of Chap. 1 and even simpler formulas can be found in Sect. A.1 at the end of this Chapter.

It is even possibile, to produce a better discretization algorithm, to take into account the spherical or ellipsoidal shape of the reference surface and use accordingly spherical/ellipsoidal prisms (Heck and Seitz 2007).

Apart from the sign of the correction that must follow the $\pm$ of the residual topography (cf. Fig. 4.8), an important point on which we have to focus is that if our area is very large, for a given point P we don't need to compute the correction due to all the prisms, some of which are very far from P and presumably produce an insignificant contribution.

Numerical experience says that computing the residual correction for an area around P of 1–2° for Δg_{RC} and 2–3° for T_{RC} is usually sufficient for our purposes.

In lack of a formal proof, we present an example with such a strong topography that should result convincing to everybody.

Example 2. We fix a point P on the plane and a topography starting only at a (plane) distance $\overline{D}$ from P. We assume that $\overline{D} > 100\,\text{km}$ and $H(Q_0)$ is done by prisms for which we shall use our approximation of Example 1, namely

$$-\Delta g(P\,;\,Q_0) = G\delta_0(L \cdot W \cdot H)\frac{H/2}{D^3} \tag{4.37}$$

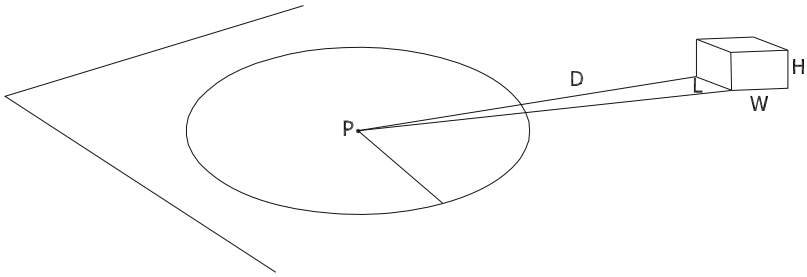


Fig. 4.9 The geometry of the residual correction of a far zone, $D > \overline{D}$

where L is the transversal size of the cube, W its width in radial direction in the plane, H the height of the prism (see Fig. 4.9).

The peculiar point of this example is to represent a quite varying topography by assuming that L and W are constant while H is a random variable with zero average and r.m.s.

$$\sigma_H = [E\{H^2\}]^{(1/2)} \tag{4.38}$$

So by averaging (4.37) we get

$$E\{-\Delta g(P, Q_0)\} = \frac{1}{2} G\delta_0 LW \frac{\sigma_H^2}{D^3}; \tag{4.39}$$

if we add the contributions of all prisms at distance D, which will be approximately

$$n = \frac{2\pi D}{L},$$

we have the contribution of all prisms between distance D and $D + W$, i.e.

$$\begin{aligned} E\{-\Delta g(P, D, W)\} &= \frac{1}{2} \frac{2\pi D}{L} G\delta_0 LW \frac{\sigma_H^2}{D^3} \\ &= \pi G\delta_0 \sigma_H^2 \frac{W}{D^2}. \end{aligned} \tag{4.40}$$

If we consider $W = dD$ and integrate from $\overline{D}$ to infinity we finally obtain

$$E\{-\Delta g(P),\ (D > \overline{D})\} = \pi G\delta_0 \frac{\sigma_H^2}{\overline{D}}. \tag{4.41}$$

With $\sigma_H = 500$ m and $\overline{D} = 125$ km this gives $\{\Delta g\} \cong 0.11$ mGal; with $\sigma_H = 1{,}000$ m and $\overline{D} = 220$ km it gives 0.25 mGal. Due to the extreme values adopted for the parameters we think that the example is quite convincing. Note that a similar

computation for T cannot be performed because the planar approximation of T_{RC} diverges if the upper limit of D_0 is tending to infinity; an effect well-known from the theory of Bouguer integrals (cf. Heiskanen and Moritz 1967) meaning only that the planar approximation is not valid for T when we take too large an area. So the limit of 2–3° in computing T_{RC} is just due to numerical experience.

(b) In this approach we want to take advantage of the fact that in a typical terrain model, when P_0, Q_0 are separated by a large enough distance

$$\widetilde{D}_0 > \overline{D}, \tag{4.42}$$

we verify that the inclination I of the line of sight between P and Q is quite small, for instance

$$|tgI| = \left|\frac{h_p - h}{\widetilde{D}_0}\right| < 10^{-1}. \tag{4.43}$$

corresponding to an inclination of less then 6°. Note that $\sup_{\widetilde{D}_0>0} |tgI|$ is a characteristic parameter of the topography of the local area.

Now consider that one can write

$$\frac{1}{\widetilde{\ell}_{0PQ}} = \frac{1}{[\widetilde{D}_0^2 + (h_P - h)^2]^{(1/2)}} = \frac{1}{\widetilde{D}_0[1 + tg^2 I]^{(1/2)}} = \frac{\cos I}{\widetilde{D}_0} \tag{4.44}$$

$$\frac{1}{\widetilde{\ell}^3_{0PQ}} = \frac{1}{[\widetilde{D}_0^2 + (h_P - h)^2]^{3/2}} = \frac{\cos^3 I}{\widetilde{D}_0^3}. \tag{4.45}$$

Since, with the bound (4.43), we have

$$|\cos I - 1| < 5 \cdot 10^{-3}$$
$$|\cos^3 I - 1| < 1.5 \cdot 10^{-2},$$

we can accept to substitute with 1 the cosine terms in (4.44) and (4.45), and we can proceed to suitably transform (4.34) and (4.35).

We call

$$\chi_{P_0}(Q_0, \overline{D}) = \begin{cases} 1 & (D_{P_0Q_0} < \overline{D}) \\ 0 & (D_{P_0Q_0} > \overline{D}), \end{cases} \tag{4.46}$$

the characteristic function of a planar disk of radius $\overline{D}$ around P_0. We note that $\chi_{P_0}(Q_0, \overline{D})$ is a function of $D_{P_0Q_0} = [(\xi - x_{P_0})^2 + (\eta - y_{P_0})^2]^{(1/2)}$ only.

Furthermore we call

$$\chi^c_{P_0}(Q_0, \overline{D}) = 1 - \chi_{P_0}(Q_0, \overline{D}) \tag{4.47}$$

i.e. the characteristic function of the complementary region of the plane, namely that lying outside the above disk.

By definition we have identically

$$\chi_{P_0}(Q_0, \overline{D}) + \chi^c_{P_0}(Q_0, \overline{D}) \equiv 1, \tag{4.48}$$

so that (4.34) can now be written as

$$\begin{aligned} T_{RC}(P) &= G\delta_0 \int d\xi d\eta \chi_{P_0}(Q_0, \overline{D}) \int_{\widetilde{H}}^{H} \frac{dh}{\widetilde{\ell}_{0PQ}} \\ &\quad + G\delta_0 \int d\xi dy \chi^c_{P_0}(Q_0, \overline{D}) \int_{\widetilde{H}}^{H} \frac{dh}{\widetilde{\ell}_{0PQ}} \\ &= T_{RC\,\mathrm{int}}(P) + T_{RC\,\mathrm{ext}}(P). \end{aligned} \tag{4.49}$$

On the other hand when $\chi^c_{P_0}(Q_0, \overline{D}) \neq 0$ we can use (4.44) with the approximation $\cos I = 1$, so that we have

$$T_{RC\,\mathrm{ext}}(P) = G\delta_0 \int d\xi d\eta \chi^c_{P_0}(Q_0, \overline{D}) \frac{[H(\xi, \eta) - \widetilde{H}(\xi, \eta)]}{[(\xi - x_P)^2 + (\eta - y_P)^2]^{(1/2)}}. \tag{4.50}$$

As we can see (4.50) is in the form of a convolution of $\delta H(\xi, \eta) = H(\xi, \eta) - \widetilde{H}(\xi, \eta)$ with the kernel

$$K(\xi - x_P, \eta - y_P) = \frac{\chi^c_{P_0}(Q_0, \overline{D})}{\widetilde{D}_{0PQ}}; \tag{4.51}$$

such convolution integrals can be very conveniently treated numerically with the Discrete Fourier Transform (DFT), as it will be explained in detail in the second part of the book, Chap. 10.

As for the first part of (4.49), the inner integral, there are two strategies: either we compute it numerically performing explicitly the inner integral, or we continue the development of ℓ^{-1}_{0PQ}, for instance putting

$$\frac{1}{\ell_{0PQ}} = \frac{1}{\widetilde{D}_{0PQ}} - \frac{1}{2} \frac{(h_P - h)^2}{\widetilde{D}^3_{0PQ}} \tag{4.52}$$

which neglects only fourth order terms in $tgI = \frac{h_P - h}{\widetilde{D}_{0PQ}}$ and then allows to reduce the threshold $\overline{D}$, so much so that in some cases the inner integral is completely neglected. As for the first approach, one can take advantage of integration formula

$$J(P,Q_0)=\int_{\widetilde{H}}^{H}\frac{dh}{\left[\widetilde{D}_{0PQ}^2+(h_P-h)^2\right]^{(1/2)}}$$
$$=\log\frac{H(Q_0)-h_P+\sqrt{\widetilde{D}_{0PQ}^2+(H(Q_0)-h_P)^2}}{\widetilde{H}(Q_0)-h_P+\sqrt{\widetilde{D}_{0PQ}^2+(\widetilde{H}(Q_0)-h_P)^2}} \tag{4.53}$$

and then discretize the integral of this function in the inner zone $\widetilde{D}_{0PQ}<\overline{D}$.

Summarizing either one writes, using the symbols (4.51) and (4.53),

$$T_{RC}(P)=G\delta_0\int d\xi d\eta\chi_{P_0}(Q_0,\overline{D})J(P,Q_0)$$
$$+G\delta_0\int d\xi d\eta K(\xi-x_P,\eta-x_P)\delta H(\xi,\eta) \tag{4.54}$$

and computes by discretization the inner integral and by DFT the outer integral, or one writes, using (4.52) and performing the integral on dh,

$$T_{RC}(P)=G\delta_0\int d\xi d\eta\frac{\delta H(\xi,\eta)}{\widetilde{D}_{0PQ}}$$
$$+\frac{G\delta_0}{6}\int d\xi d\eta\frac{[(h_P-H(Q_0))^3-(h_P-\widetilde{H}(Q_0))^3]}{\widetilde{D}_{0PQ}^3}. \tag{4.55}$$

Sometimes formulas (4.54) and (4.55) are combined, simply to reduce the computational work in the inner zone, though avoiding the use of diverging integrals as in (4.55).

In fact, note that (4.55) becomes a sum of convolution integrals only after developing the powers in the second term.

This difficulty shows that preserving in any case an inner zone (maybe small) as in (4.54) is not only more precise from the numerical point of view, but also neater as for the theoretical meaning of integrals.

By applying the same reasoning to (4.35) one gets first a formula similar to (4.54). Consider that

$$\int_{\widetilde{H}}^{H}\frac{h_P-h}{\widetilde{\ell}_{0PQ}^3}dh=\frac{1}{\left[\widetilde{D}_{0PQ}^2+(h_P-H(Q_0)^2\right]^{(1/2)}}+$$
$$-\frac{1}{\left[\widetilde{D}_{0PQ}^2+(h_P-\widetilde{H}(Q_0)^2\right]^{(1/2)}}=\Gamma(P,Q_0) \tag{4.56}$$

and that for $\widetilde{D}_{0PQ}>\overline{D}$ one can put

$$\int_{\widetilde{H}}^{H} \frac{h_P - h}{\widetilde{\ell}^3_{0PQ}} dh = \frac{1}{\widetilde{D}^3_{0PQ}} \int_{\widetilde{H}}^{H} (h_P - h) dh$$
$$= \frac{1}{\widetilde{D}^3_{0PQ}} \left[\frac{1}{2} (h_P - \widetilde{H})^2 - \frac{1}{2} (h_P - H)^2 \right] = \frac{1}{\widetilde{D}^3_{0PQ}} (H - \widetilde{H}) \left(h_P - \frac{H + \widetilde{H}}{2} \right). \tag{4.57}$$

Defining

$$B(\xi - x_P, \eta - y_P) = \frac{\chi^c_{P_0}(Q_0, \overline{D})}{\widetilde{D}^3_{0PQ}}, \tag{4.58}$$

one can write

$$\Delta g_{RC}(P) = G\delta_0 \int d\xi d\eta \chi_{P_0}(Q_0, \overline{D}) \Gamma(P, Q_0)$$
$$+ G\delta_0 \int d\xi d\eta B(\xi - x_P, \eta - y_P) \delta H(\xi, \eta) \left[h_P - \frac{H(\xi, \eta) + \widetilde{H}(\xi, \eta)}{2} \right], \tag{4.59}$$

which is the sought equation.

Sometimes the equation (4.59) is directly applied without the inner zone integral or the development of $\dfrac{h_P - h}{\widetilde{\ell}^3_{0PQ}}$ is pushed to higher order terms, similarly to (4.55); in this case however we incur again theoretical difficulties and we shall not pursue further this line.

A final remark is that any approximation of the vanishing or smoothed effects of masses far away from the computation point should be done in principle only on a numerical basis. In fact, strictly speaking, introducing a moving average form of formulas (4.34) modifies the harmonic character of the potential.

4.6 Comparisons and Interpretations

First of all we want to compare our results, in particular (4.59), with the classical theory of Bouguer correction, so much used in geophysical interpretation.

Typically the Bouguer correction is derived in the following way: we start by assuming a planar reference geometry and we compute the Bouguer terrain effect at a point P (see Fig. 4.10) by splitting it into the effect of a slab of density δ_0 up to the height h_P plus the effect of the differential topography with height $H_Q - h_P$.

For the slab part, as we suggest to the reader to prove in the subsequent Exercise 1, we have

$$\Delta g_{\text{slab}} = 2\pi G \delta_0 h_P. \tag{4.60}$$

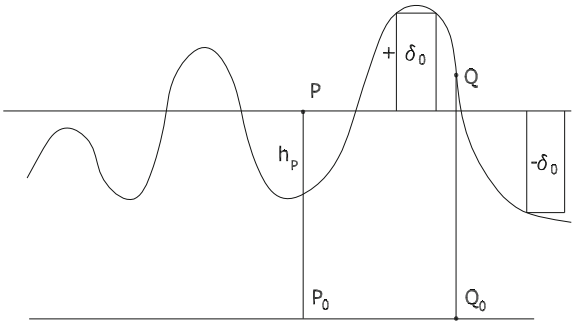


Fig. 4.10 The geometry of Bouguer correction

What is peculiar of the slab geometry is that the attraction (4.60) is the same at every point in space, outside the plane $h = h_P$ (cf. Exercise 1).

For the differential part we have

$$\Delta g_{\text{diff}}(P) = G\delta_0 \int d\xi d\eta \int_{h_P}^{H_Q} \frac{h_P - h}{\widetilde{\ell}_{0PQ}^3} dh. \tag{4.61}$$

If we take, as we are doing, $h_P > H_P$, the integral never becomes singular so we can proceed.

At this point we assume that the topography has a small inclination and we decide, therefore, that the approximation

$$\frac{1}{\widetilde{\ell}_{0PQ}^3} \cong \frac{1}{\widetilde{D}_{0PQ}^3}$$

can be accepted. Naturally, this is not true when P_0 is close to Q_0, however, since the orginal integral is not singular, we can always exclude a small neighborhood of P_0 without affecting too much the computation.

So, using the above relation and performing the integral on dh we obtain

$$\Delta g_{\text{diff}}(P) = -G\delta_0 \frac{1}{2} \int d\xi d\eta \frac{(h_P - H_Q)^2}{\widetilde{D}_{0PQ}^3}. \tag{4.62}$$

In particular we are allowed now to put $h_P = H_Q$ in the integral in (4.62) because this is not strongly singular if the inclination of topography, $\left(\frac{H_P - H_Q}{\widetilde{D}_{0PQ}}\right)$, is bounded, as we assume.

Note that the effect of the differential topography is always negative, i.e. the corresponding correction is always positive; this is a distinctive characteristic of Bouguer correction.

Adding (4.60) and (4.62), with $h_P > H_P$, we get the complete Bouguer effect (see also Heiskanen and Moritz 1967, Chap. 3)

$$\Delta g_B = 2\pi G\delta_0 h_P - G\delta_0 \frac{1}{2}\int d\xi d\eta \frac{[H_P - H(\xi,\eta)]^2}{[(\xi - x_P)^2 + (\eta - y_P)^2]^{3/2}}. \tag{4.63}$$

Now write the same relation (4.63) for the reference surface $\widetilde{H}(\xi,\eta)$,

$$\Delta \widetilde{g}_B = 2\pi G\delta_0 h_P - G\delta_0 \frac{1}{2}\int d\xi d\eta \frac{[H_P - \widetilde{H}(\xi,\eta)]^2}{\widetilde{D}^3_{0PQ}}. \tag{4.64}$$

If we go to the difference of the two effects we find

$$\begin{aligned}\Delta g_B - \Delta \widetilde{g}_B &= G\delta_0 \frac{1}{2}\int d\xi d\eta \frac{[H_P - \widetilde{H}_Q]^2 - [H_P - H_Q]^2}{\widetilde{D}^3_{0PQ}} \\ &= G\delta_0 \int d\xi d\eta \frac{(H_Q - \widetilde{H}_Q)\left(H_P - \frac{H_Q + \widetilde{H}_Q}{2}\right)}{\widetilde{D}^3_{0PQ}}.\end{aligned} \tag{4.65}$$

The reasoning of the Bouguer formula is not very clean from the theoretical point of view, because the linear term in h_P cannot correspond to any regular potential in the half-space $h_P > 0$. Moreover when we subtract the two convergent integrals (4.63) and (4.64) we arrive at formula (4.65) where the integral is not convergent anymore. Yet we did it to show that (4.65) is basically the same as (4.59), without the inner zone part.

It has to be underlined however that in contrast to the simplistic reasoning of the Bouguer theory, our development of $T_{RC}, \Delta g_{RC}$ has been much more rigorous in the sense that at each step the degree of approximation has been suitably controlled.

Now that the comparison of the RTC theory with that of Bouguer has been accomplished, we have to tackle an important issue, namely to give a justification for our choice of the reference $\widetilde{S}$ surface.

We will perform a rough reasoning providing an answer which can be accepted as a general rule only with the understanding that it is grossly approximated, so that its implementation requires specific numerical investigations.

Let us go back to our rough model of Sect. 4.3; there we learnt that the coefficients $\{H_{nm}\}$ enter, with a proportionality constant, into Δg. However when we compute a global model, the data Δg which we use to compute Δg_{nm} are obtained first by downward continuing the actual data and then by averaging them on the ellipsoid.

In reality all that is done in one step by using prediction methods that will be studied in the next chapter. However here we shall work with the singular model (4.67) making much clearer our procedure.

Since the first order term is by far the largest in the downward continuation, we can write

$$\Delta g_e(P_0) = \Delta g_{\text{obs}}(P) - \frac{\partial \Delta g_M(P)}{\partial h} H_P, \tag{4.66}$$

where P_0 is on the ellipsoid, while P is the point on the surface where Δg has been observed. Remember that in (4.66) the vertical gradient of Δg is computed not for the true Δg (because we are not able) but rather for the model Δg_M through an iterative procedure.

We have already defined in (4.23) the moving average operator M_Δ, which in practice is substituted by averaging on blocks of geographic squares $\left\{\overline{\varphi} - \frac{\Delta}{2} \le \varphi \le \overline{\varphi} + \frac{\Delta}{2},\ \overline{\lambda} - \frac{\Delta}{2} \le \lambda \le \overline{\lambda} + \frac{\Delta}{2}\right\}$.

As it happens to all spherical filters, also M_Δ has a distinct behaviour on the harmonic coefficients of any function to which it is applied; basically it is a low-pass filter tending to leave the low degrees almost unchanged, while it tends to depress all wavelengths shorter than 2Δ, i.e. degrees $n > \frac{180^\circ}{\Delta^\circ}$.

On this point one can find more particulars in Sect. A.4.

So if we apply $M_\Delta\{\ \}$ to (4.66) we find

$$\Delta\overline{g}(P_0) = M_\Delta\{\Delta g_e(P_0)\} = M_\Delta\{\Delta g_{\text{obs}}(P)\} - M_\Delta\left\{\frac{\partial \Delta g_M(P)}{\partial h} H_P\right\}. \quad (4.67)$$

Now assume that our model Δg_M has a maximum degree N and our moving average has a radius Δ such that

$$\Delta \le \frac{180^\circ}{N}; \quad (4.68)$$

then we expect $\Delta g_M(P)$ as well as $\frac{\partial \Delta g_M}{\partial h}$ to be almost unaffected by M_Δ so that we can write

$$M_\Delta\left\{\frac{\partial \Delta g_M}{\partial h} H_P\right\} \cong \frac{\partial \Delta g_M}{\partial h} M_\Delta\{H_P\} = \frac{\partial \Delta g_M}{\partial h}\widetilde{H}(P_0). \quad (4.69)$$

Going back to (4.67) we see that

$$\Delta\overline{g}(P_0) \cong M_\Delta\{\Delta g_{\text{obs}}(P)\} - \frac{\partial \Delta g_M}{\partial h}\widetilde{H}(P_0), \quad (4.70)$$

i.e. Δg_M is ultimately evaluated from block averages (4.70) derived by downward continuation from the surface $\widetilde{S} \equiv \{h = \widetilde{H}(P_0)\}$. Said in another way, Δg_M will tend to represent the true field up to degree N including the effects of the blurred DTM function $\widetilde{H}(P_0)$.

Since in our remove-restore process we are going to split the actual Δg into

$$\Delta g = \Delta g_M + \Delta g_{RC} + \Delta g_r \quad (4.71)$$

and then, after computing T_r from Δg_r, we reconstruct T as

$$T_r + T_{RC} + T_M = T, \quad (4.72)$$

it is clear that, if Δg_{RC} represents the terrain effects of the topography residual with respect to the *same surface* $\widetilde{S}$, the Δg_r will be not only smoother because we reduced the high frequency part, but also smaller, for instance in mean quadratic sense, so making easier our last step of going from Δg_r to T_r. The conclusion is that we can state as a rule of thumb that our reference surface for the residual terrain correction has to be $\widetilde{H} = M_\Delta\{H\}$ with

$$\Delta \cong \frac{180^\circ}{N}, \tag{4.73}$$

where N is the maximum degree of the gravity anomaly model Δg_M.

Remark 4. It is clear that in all our reasonings there are many approximations, the strongest of which is to assume (4.69) to hold. So (4.73) can be used just as a starting point and, computing Δg_{RC} for different values of Δ, we can choose for instance the one that leaves the smallest residual Δg_r.

4.7 An Open Issue

Given the discussion of Sect. 4.4, one may wonder whether there could be a more direct and consistent way to make the surface $\widetilde{S}$ unique and compliant with the decomposition

$$\Delta g = \Delta g_M + \Delta g_{RC} + \Delta g_r. \tag{4.74}$$

This is as a matter of fact object of debate and might become an accepted practice in future, so we just outline the idea as a possibility. The main point is to define $\widetilde{S}$ and the corresponding residual field effects $\Delta g_{RC}, T_{RC}$ as a first step and globally, and only afterwards to estimate a global model with respect to the new data set given on the well-defined reference surface $\widetilde{S}$. In fact, let us start from the definition of $\widetilde{S} = \{h = \widetilde{H}(P_0)\}, \widetilde{H}(P_0) = M_\Delta\{H\}$, where Δ is chosen according to the rule (4.73) in relation to the maximum degree N of the model we want to estimate afterwards. Once $\widetilde{S}$ is defined, the regions with $\delta H > 0$ and $\delta H < 0$ are defined too (see Fig. 4.11); they are labelled with $+$ and $-$ respectively. We proceed to compute residual terrain corrections and move from P to $\widetilde{P}$ in the following way: we first compute the terrain effect of the regions labelled $+$ at P, $\Delta g_{RC+}(P)$, and apply this correction to all points; then we move all points to $\widetilde{S}$ using a prior model Δg_{M0} or better its vertical derivative $\frac{\partial \Delta g_{M0}}{\partial h}$; note that in this way all points are moving in free air, as P and P' in Fig. 4.11; then we compute at points $\widetilde{P}$ the further residual gravity effect of the regions labelled with a $-$, $\Delta g_{RC-}(\widetilde{P})$, and correct to obtain the final data on $\widetilde{S}$.

In Fig. 4.12 we give a schematic view of this operation.

In this way we produce a new field of gravity anomalies $\Delta\widetilde{g}$ on $\widetilde{S}$.

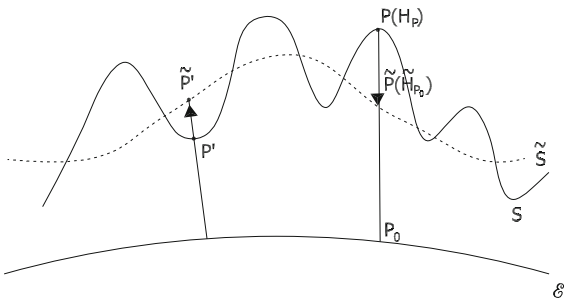


Fig. 4.11 Geometry of the reduction of gravity data from S to $\widetilde{S}$

$$
\begin{aligned}
\Delta g(P) \rightarrow\ & \Delta g(P) - \Delta g_{RC+}(P) \\
& \downarrow \\
& [\Delta g(P) - \Delta g_{RC+}(P)] + \left(\frac{\partial \Delta g_{M0}}{\partial h}\right) \delta H_{P_0} \\
& \downarrow \\
& [\Delta g(P) - \Delta g_{RC+}(P)] + \left(\frac{\partial \Delta g_{M0}}{\partial h}\right) \delta H_{P_0} - \Delta g_{RC-}(\widetilde{P}) \rightarrow \Delta \widetilde{g}(\widetilde{P})
\end{aligned}
$$

Fig. 4.12 The flow of corrections and change of boundary $P \rightarrow \widetilde{P}$ to produce the new data set $\Delta \widetilde{g}(\widetilde{P})$

Now $\widetilde{S}$ is, by its own definition, smooth enough to allow for a meaningful application of the moving average operator $M_\Delta\{\cdot\}$ to the data $\Delta \widetilde{g}$. In other words we can compute $M_\Delta\{\Delta \widetilde{g}\}$ and we can assign this value to the corresponding center $\widetilde{P}$ on $\widetilde{S}$, while if we computed moving averages of the original data $\Delta g(P)$ we would not know to what points in space these averages correspond.

At this point we could proceed to estimate a best approximation model $\widetilde{T}_M$ by an iterative solution of the least squares principle as described for instance in Part III, Sect. 14.5.

As explained in Part III, Chap. 15, Sect. 15.5, this is not the truncated development of the true gravity field, but only a best approximation in terms of ellipsoidal harmonics up to degree N of a true gravity field, which has not a convergent series representation at the level of the earth surface.

Once $\widetilde{T}_M$ has been computed we have available a procedure to approximate $T(P)$ at any P in space obtained by inverting the reasoning as represented in Fig. 4.13.

A perfectly analogous scheme can be constructed, starting from $\Delta \widetilde{g}_M(\widetilde{P})$ and ending with $\Delta \widetilde{g}(P)$, which is necessary to compute the final residual gravity anomalies.

It is essential, in applying a scheme like that, that all the quantities $\delta H, \Delta g_{RC\pm}, T_{RC\pm}$ be taken with their proper sign. The final result $\widetilde{T}(P)$ or $\Delta \widetilde{g}(P)$ will then be a "consistent" combination of approximation of $T(P), \Delta g(P)$ on both the long wavelengths, thanks to $\widetilde{T}_M, \Delta \widetilde{g}_M$, and the short wavelengths, thanks to $T_{RC\pm}, \Delta g_{RC\pm}$ respectively.

$$\begin{aligned}
&\widetilde{T}_M(\widetilde{P}) \rightarrow \widetilde{T}_M(\widetilde{P}) + T_{RC-}(\widetilde{P}) \\
&\qquad\qquad \downarrow \\
&\widetilde{T}_M(\widetilde{P}) + T_{RC-}(\widetilde{P}) - \left(\frac{\partial \widetilde{T}_M}{\partial h}\right) \delta H_{P_0} \\
&\qquad\qquad \downarrow \\
&\widetilde{T}_M(P) + T_{RC-}(\widetilde{P}) - \left(\frac{\partial \widetilde{T}_M}{\partial h}\right) \delta H_{P_0} + T_{RC+}(P) \rightarrow \widetilde{T}(P)
\end{aligned}$$

Fig. 4.13 The scheme of the restore procedure with the shift from $\widetilde{P}$ to P

Accordingly we can at the end compute the residual data

$$\Delta g_r(P) = \Delta g(P) - \Delta \widetilde{g}(P) \tag{4.75}$$

and apply some suitable method to transform this into an estimate of

$$T_r(P) = T(P) - \widetilde{T}(P), \tag{4.76}$$

that will finally allow us to determine $T(P)$ and the corresponding height anomaly $\zeta(P)$ everywhere in the outer space and on S.

We just note that in this way theoretically the only "terrain" effects left in $\Delta g_r(P)$ are those which derive from the masses between $\mathcal{E}$ and $\widetilde{S}$ that cannot be described by a model $\widetilde{T}_M$ up to degree N; such effects in any way have to be smooth because they refer to a geometry with a smooth boundary, i.e. $\widetilde{S}$; therefore the application of Runge-Krarup theorem (see Sect. 3.5) is favoured. As a final Remark we underline that the Molodensky principle of having the observation points where they are is not violated by this approach because once the approximate fields $\widetilde{T}$, $\Delta \widetilde{g}$ are determined, the residual fields are computed at the right point P in space.

4.8 Exercises

Exercise 1. Since it is really important we propose to the reader to prove that an infinite slab of density δ_0 and width h creates everywhere outside the slab an attraction given by

$$\Delta g = 2\pi G \delta_0 \cdot h.$$

With reference to Fig. 4.14, we propose to do that in two different ways:

(a) Compute directly the integral, for any point P at altitude $h_P > h$,

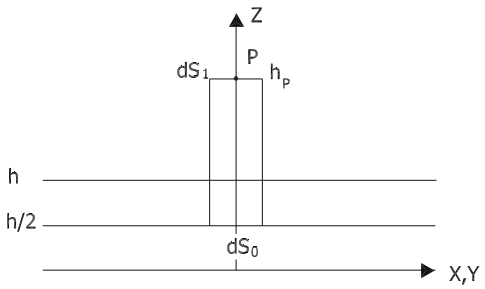


Fig. 4.14 Geometry of a homogeneous slab

$$\Delta g(h_P) = G\delta_0 \int d\xi d\eta \int_0^h dz \frac{h_P - z}{[\xi^2 + \eta^2 + (h_P - z)^2]^{3/2}}$$

(**Hint:** use cylindric coordinates $\xi^2 + \eta^2 = \rho^2$, $d\xi d\eta = \rho d\rho d\alpha$),

(b) Apply Gauss' theorem to an infinitesimal cylinder of base dS and note that Δg has to be zero on dS_0 and pointing inward on dS_1, as well as to be tangent to the lateral wall of the cylinder

Exercise 2. Consider the simple Bouguer formula (4.63) and apply it to a conic mountain, computing Δg_B on the peak of the mountain. Assume that H_0 is the height of the top and b the radius of the circular base and $tgI = \frac{H_0}{b}$ the slope of the mountain, and prove that

$$\Delta g_B = 2\pi G\delta_0 H_0 \left[1 - \frac{1}{2} tgI\right].$$

Comment on the fact that clearly such a formula cannot be meaningful when tgI is of the order of 1.

Furthermore compare the present result with that of Exercise 5 of Chap. 1 to show that even if $tgI \sim \sin I$ for small I, still we have an error of the order of $\frac{1}{2}\sin I$ in the Bouguer formula.

Exercise 3. Consider the case of a parallelopiped of constant density δ_0 and of sides $2a, 2b, 2c$ already treated in Exercise 3 of Chap. 3. According to that result, if we place the origin of the coordinates at the center of the prism, we have

$$T(P) = T_0(P) + T_2(P) = \\ = \frac{GM}{r_P} + \frac{GM}{r_P^5}\left[\frac{3}{2}\mathbf{r}_P^t I \mathbf{r}_P - \frac{1}{2}(TrI)r_P^2\right] + O\left(\frac{1}{r_P^4}\right)$$

with

$$I = \frac{\delta_0}{M}\int_D \mathbf{r}_Q \mathbf{r}_Q^t dB = \frac{1}{V}\int_D \mathbf{r}_Q \mathbf{r}_Q^t dB,$$
$$D \equiv \{-a \le x \le a,\ -b \le y \le b,\ -c \le z \le c\}, \quad V = 2a \cdot 2b \cdot 2c.$$

Prove that in this case, in Cartesian coordinates,

$$I = \frac{1}{3}\begin{vmatrix} a^2 & 0 & 0 \\ 0 & b^2 & 0 \\ 0 & 0 & c^2 \end{vmatrix}$$

and that therefore we have

$$T_2(P) = \frac{GM}{r_P^5}\,\frac{1}{6}\cdot\left[x_P^2(2a^2 - b^2 - c^2) + y_P^2(-a^2 + 2b^2 - c^2) + z_P^2(-a^2 - b^2 + 2c^2)\right].$$

Such a formula can be usefully applied to express $T(P)$ at points P for which $r_P \gg \sqrt{a^2 + b^2 + c^2}$.

Appendix

A.1

In this Appendix we like to prove that there are various formulas more numerically convenient than that found in Exercise 9 of Chap. 1 to express the potential of a parallelopiped

$$D \equiv \{-a \le x \le a,\ -b \le y \le b,\ -c \le z \le c\}.$$

Among them, one often met in literature is (MacMillan 1958)

$$2T = G\delta_0 \,|||2xy\log(z+R) + 2xz\log(y+R) + 2zy\log(x+R)$$

$$- x^2 \arctan\frac{yz}{xR} - y^2 \arctan\frac{xz}{yR} - z^2 \arctan\frac{xy}{zR}\Bigg|_{x_2}^{x_1}\Bigg|_{y_2}^{y_1}\Bigg|_{z_2}^{z_1}. \tag{4.77}$$

In (4.77), given the convention we follow to put the origin of the Cartesian axes at the center of the prism and the axes themselves parallel to the edges, and calling as in Exercise 6, Chap. 15,

$$A_\pm = a \pm x,\ B_\pm = b \pm y,\ C_\pm = c \pm z \tag{4.78}$$

as well as

$$R_{\pm\pm\pm} = \sqrt{A_\pm^2 + B_\pm^2 + C_\pm^2}, \tag{4.79}$$

the limits $x_i, y_i, z_i,\ i = 1, 2$, are given by the conventions

$$x_1 = A_-,\ y_1 = B_-,\ z_1 = C_-$$
$$x_2 = -A_+,\ y_2 = -B_+,\ z_2 = -C_+. \tag{4.80}$$

Indeed to prove the equivalence of (4.77) with formulas of Exercise 9, Chap. 1, it is enough to prove equality for one logaritmic term and one arctangent term, since the others will follow by symmetry.

So let us take for instance in (4.77) the terms in x, y, neglecting the factor $G\delta_0$, namely

$$\begin{aligned} 2T_1 &= 2x_1 y_1 \log(Z_1 + R_{---}) - 2x_1 y_1 \log(Z_2 + R_{--+}) \\ &= 2A_- B_- \log \frac{R_{---} + C_-}{R_{--+} - C_+}. \end{aligned} \tag{4.81}$$

If we look at Exercise 9, Chap. 1, we find in fact two logaritmic terms that multiply $A_- B_-$, one coming from $A_- B_-[I(A_-, B_-, C_-) - I(A_-, B_-, -C_+)]$, the other coming from $B_- A_-[I(B_-, A_-, C_-) - I(B_-, A_-, -C_+)]$. Since the logaritmic part of $I(A_-, B_-, \pm C_\mp)$ is symmetric with respect to the exchange of A_- with B_-, the two terms above are exactly the same and we will have in $2T$ a term like

$$\begin{aligned} &2A_- B_- \left[\log \frac{R_{---} + C_-}{\sqrt{A_-^2 + B_-^2}} - \log \frac{R_{--+} - C_+}{\sqrt{A_-^2 + B_-^2}} \right] = \\ &2A_- B_- \log \frac{R_{---} + C_-}{R_{--+} - C_+}, \end{aligned} \tag{4.82}$$

which is equal to $2T_1$ in (4.81).

As for the terms in arctan we can take in (4.77) only those multiplied by A_-^2, which give rise to the expression, neglecting $G\delta_0$,

$$\begin{aligned} 2T_2 &= -A_-^2 \left[\left(\arctan \frac{y_1 z_1}{x_1 R_{---}} - \arctan \frac{y_1 z_2}{x_1 R_{--+}} \right) + \right. \\ &\quad \left. - \left(\arctan \frac{y_2 z_1}{x_1 R_{-+-}} - \arctan \frac{y_2 z_2}{x_1 R_{-++}} \right) \right] \\ &= -A_-^2 \left[- \arctan \frac{x_1 R_{---}}{y_1 z_1} + \arctan \frac{x_1 R_{--+}}{y_1 z_2} \right. \\ &\quad \left. + \arctan \frac{x_1 R_{-+-}}{y_2 z_1} - \arctan \frac{x_1 R_{-++}}{y_2 z_2} \right] \\ &= A_-^2 \left[\arctan \frac{A_- R_{---}}{B_- C_-} + \arctan \frac{A_- R_{--+}}{B_- C_+} \right. \\ &\quad \left. + \arctan \frac{A_- R_{-+-}}{B_+ C_-} + \arctan \frac{A_- R_{-++}}{B_+ C_+} \right]. \end{aligned} \tag{4.83}$$

In the formula above, use of the identity $\arctan X = \frac{\pi}{2} - \arctan \frac{1}{X}$ has been done. Similarly from Exercise 9, Chap. 1, we derive the terms that multiply A_-^2 from $A_- F(A_- B_-, B_+, C_-, C_+)$, namely

$$\begin{aligned} -A_-^2 \Bigg[& \arctan \frac{B_- R_{---}}{A_- C_-} + \arctan \frac{B_- R_{--+}}{A_- C_+} \\ & + \arctan \frac{B_+ R_{-+-}}{A_- C_-} + \arctan \frac{B_+ R_{-++}}{A_- C_+} \\ & + \arctan \frac{C_- R_{---}}{A_- B_-} + \arctan \frac{R_{-+-} C_-}{A_- B_+} \\ & + \arctan \frac{C_+ R_{--+}}{A_- B_-} + \arctan \frac{R_{-++} C_+}{A_- B_+} \Bigg]. \end{aligned} \tag{4.84}$$

Now, exploiting the identity

$$\arctan X + \arctan Y = \arctan \frac{X + Y}{1 - XY},$$

we can show that every two terms in (4.84) add to give a corresponding term in (4.83), all containing the expression R with the same signature. We do that for the terms including R_{---}. In fact we have

$$\begin{aligned} & - \left[\arctan \frac{B_- R_{---}}{A_- C_-} + \arctan \frac{C_- R_{---}}{A_- B_-} \right] \\ & = - \arctan \frac{R_{---}}{A_-} \frac{\frac{B_-}{C_-} + \frac{C_-}{B_-}}{1 - \frac{R_{---}^2}{A_-^2}} \\ & = \arctan \frac{R_{---} A_-}{B_- C_-}, \end{aligned}$$

as it was to be proved.

Chapter 5
The Local Modelling of the Gravity Field by Collocation

5.1 Outline of the Chapter

The chapter aims at solving the problem of estimating the residual anomalous potential T_r from all available information, in particular in a certain area. Remember that here residual means that the long wavelength part as well as the short wavelength part of T have been at least reduced by means of the deterministic modelling described in Chaps. 3 and 4.

These models are then applied to data (remove step); from reduced observations we need to find T_r and then the models are added back to this (restore step).

Since the residual part of the potential is small (one has in terms of anomalous height $O\left(\frac{T_r}{\gamma}\right) \cong 2\,\mathrm{m}$), the application of spherical approximation is justified.

This notwithstanding such an approximation remains the harsh limitation of the theory presented in this chapter. This point is explained in Sect. 5.2.

The theory, known in geodesy as *collocation theory*, is introduced here as an optimization problem where a suitable mean square error has to be minimized in a class of estimators invariant under a certain transformation group, acting on the set Ω where the unknown function is defined. Although not so much relevant in geodesy, the case of the circle is on the same time so simple to understand and so complete from the theoretical point of view, that it has been worthwhile to devote Sect. 5.3 to it.

In Sect. 5.4 the same case is treated for the sphere, with the invariance group being that of rotations in R^3. The big theoretical advantage of this approach is that not only the estimation coefficients result as an application of the optimality principle, but also the definition of the covariance function springs out of it in a natural way.

In Sect. 5.4 it is also shown that the formalism set up in the previous paragraphs can be given a stochastic interpretation, to the effect that now T is considered as a random function, obtained by randomly rotating the true T. The formalism is then extended in Sect. 5.5 to the general case, in which we have whatever N observations, corresponding to admissible linear functionals, and we want to predict any other

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_5,

admissible linear functional of T. In particular, if we assume, invoking the Runge–Krarup theorem, that T is a function harmonic down to a Bjerhammar sphere, any rotated version of T will continue to be harmonic in the same domain, and the principle above devised, applies.

Since a function harmonic in the exterior of a sphere has a natural representation in terms of spherical harmonics, its coefficients will become random variables, when the field T they represent is random too. The properties of such $\{T_{nm}\}$ as well as their relation to the covariance function of T, are examined in Sect. 5.6. In Sect. 5.7 the item of a local modelling of the covariance function is analyzed and several examples are presented, including those most widely applied in practical computations.

The local computation of a (residual) quasi-geoid from (residual) gravity anomalies is then presented as an example of the so-called *least squares collocation theory*.

Finally, in Sect. 5.9 the optimal combination of a global model, for instance derived from satellite observations, and local data to produce the best local prediction of the geoid, is explicitly solved; a case this that is becoming increasingly important in these years.

5.2 An Introduction to the Problem

Following the developments of Chaps. 3 and 4 we could say that our anomalous gravity potential T has been approximated in the long wavelengths range by a global model T_M and in the very short wavelengths range by the residual terrain correction model T_{RC}, so that a residual anomalous potential

$$T_r = T - T_M - T_{RC} \tag{5.1}$$

has now to be estimated.

This has to be done by using the *residual* observations, which in linearized form are written as

$$\begin{aligned} y_i &= L_i(T_r) + v_i \\ &= L_i(T) + v_i - L_i(T_M) - L_i(T_{RC}) \\ &= Y_i - L_i(T_M) - L_i(T_{RC}), \end{aligned} \tag{5.2}$$

where Y_i are the original observations, y_i the observations reduced by the effects of T_M and T_{RC}, v_i is the observational error. Typical for (5.2), but not the only case considered in the book, is the observation of free air gravity anomalies, for which the relation holds

$$L_i(T) = \left(-\frac{\partial T}{\partial h} + \frac{\gamma'}{\gamma} T\right)\Big|_{P_i}. \tag{5.3}$$

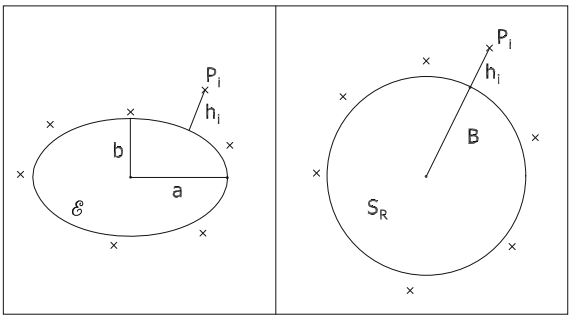


Fig. 5.1 The spherical approximation mapping of the interpolation problem: P_i measurement points, h_i heights over E and S_R

In order to avoid a heavy notation, while developing our methodological apparatus we shall simply put

$$u(P) = T_r(P). \tag{5.4}$$

So, due to all our reductions, $u(P)$ is a harmonic field which, in an ideal case, we expect to be harmonic down to the ellipsoid because the signal caused by the large and smooth density anomalies should be accounted for by T_M and the high frequency signal due to the residual terrain height should be subtracted by means of T_{RC}. Then we could reasonably think of our problem as the one of interpolating the observations (5.2) with a function harmonic down to $\mathcal{E}$.

Since we are approximating the last couple of meters in terms of height anomaly $\zeta = \gamma^{-1}u = \gamma^{-1}T$, we shall accept a spherical approximation set up, for the approximation procedure, in the sense that we map $\mathcal{E}$ to a mean sphere S_R of radius R and we reason with functions harmonic down to S_R (see Fig. 5.1).

Therefore our problem now is to find a function $\widehat{u}$ harmonic down to S_R, such that $y_i - L_i(\widehat{u})$ be small, in the sense of the order of magnitude of ν_i (i.e. of σ_{ν_i}), and as close as possible to u.

It is clear in fact that, as the number of observation points, N, can be very large, but in any event always finite, in principle we can always find many harmonic fields $\widehat{u}$ which in fact interpolate perfectly the data, $L_i(\widehat{u}) = y_i$, as shown very schematically in Fig. 5.2, where the observation points P_i are taken directly on S_R and $L_i(u) = u(P_i)$ is represented in terms of geoid, $\gamma^{-1}u(P_i)$.

Generally speaking, since in nature masses will tend to find a minimum energy configuration (compatibly with the endogenous forces generated by geological processes) and energy is in any way a quadratic positive functional of $u(P)$ which is smaller the smoother is the field, we would prefer an interpolator as smooth as possible, among those that reproduce the data. Even more, if a noise ν is part of our model, we would accept that $L_i(\widehat{u})$ will depart form y_i, with residuals of the order of σ_ν, and on the same time $\widehat{u}$ to be as smooth as possible.

If a smoothness index is taken in terms of a square norm, we are led to the Tikhonov principle which is illustrated and worked out in Part III, Chap. 12. Yet, as one can see in this chapter, the solution does depend quite essentially on the

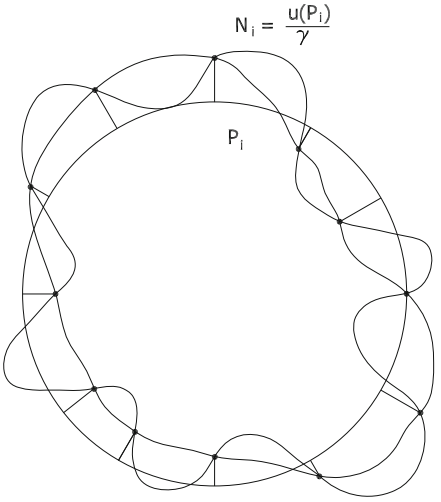


Fig. 5.2 Two different exact interpolations of $\frac{u(P_i)}{\gamma} = N_i$, by two different fields $\widehat{u}(P)$

specific norm chosen to measure the smoothness of $\widehat{u}$, when the norm is represented by a suitable reproducing kernel $K(P, Q)$.

In other words, we have a so-called *norm choice problem* which is absolutely unsolvable on a pure analytical ground. So we shall follow here a different approach which, as we will see, will lead basically to the same solution as that of Sect. 12.4 of Part III but with a precise choice for the reproducing kernel. This solution is based on the choice of an invariant estimator and minimum mean square prediction error, and on its stochastic interpretation.

Notice that in principle we expect $u(P)$ to be harmonic down to $\mathcal{E}$, then approximated by S_R. Yet such condition will never be precisely satisfied; nevertheless by choosing an interpolator $\widehat{u}$ which is authentically harmonic down to S_R we don't prevent ourselves to approximate as closely as we like the true $u(P)$, because of Runge–Krarup theorem (see Sect. 3.5).

In fact, as proved in Part III, Chap. 13, the restrictions of functions $\widehat{u}$ harmonic in $\Omega_R \equiv (r \geq R)$ to the set Ω_e of points exterior to the earth surface S_e, are dense in any reasonable Hilbert space to which we can think that $u(P)$ belongs, for instance in $HL^2(S_e)$, namely the functions harmonic in Ω_e and square integrable on S_e. So, from now on, we shall ignore the problem of the masses between S_e and S_R not perfectly modelled.

5.3 The Principle of Minimum Square Invariant Prediction Error by a Simple Example

In order to select a particular satisfactory solution to our interpolation problem, we have first to define an index expressing analytically our degree of satisfaction, or, if you like, of dissatisfaction, and then to maximize such an index in the former case,

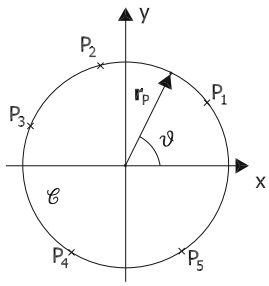


Fig. 5.3 The set up of the interpolation problem on the circle

or, on the contrary, to minimize it in the latter case. This is a problem of optimization theory, where the choice of the target function is always the first fundamental step (see for instance Vapnik 1982, Chap. 2). We choose to minimize a quadratic function of the prediction error, averaged in some suitable sense.

In order to set up our criterion we prefer to start with a simple example where our choice will become very transparent.

Example 1. Assume you have a field $u(P)$ where $P \in C$, a unit circle, so that P can be uniquely identified by a unit vector $\mathbf{r}_P$ or by the angle ϑ of $\mathbf{r}_P$ with respect to the x axis (see Fig. 5.3).

To make things easier we shall assume from the beginning that $u(P)$ has zero mean on C, i.e. that

$$\int_0^{2\pi} u(P)d\vartheta = \int_0^{2\pi} u(\vartheta)d\vartheta = 0. \tag{5.5}$$

Now assume you have observed the values of $u(P)$ at some points P_i

$$y_i = u(P_i), \quad i = 1, 2, \ldots, N \tag{5.6}$$

without any error, and you want to predict $u(P)$ at some other point P. As we see, we have a pure interpolation problem on C.

We note first of all that a *predictor* will be in general a function of the observations $\{y_i\}$ of the points $\{P_i\}$ where the observations are taken and of the prediction point P, in such a way that we are able to compute it when we know $\{y_i\}$ and we fix P;

$$\widehat{u}(P) = F(P, P_1, \ldots, P_N; y_1, \ldots, y_N). \tag{5.7}$$

Since reasoning in a general class of predictors $\{F\}$ is too complicated we shall restrict ourselves to the much simpler class of linear predictors, namely

$$\widehat{u}(P) = F(P, P_1, \ldots, P_N; y_1, \ldots, y_N) = \sum_{i=1}^{N} \lambda_i y_i = \sum_{i=1}^{N} \lambda_i u(P_i). \tag{5.8}$$

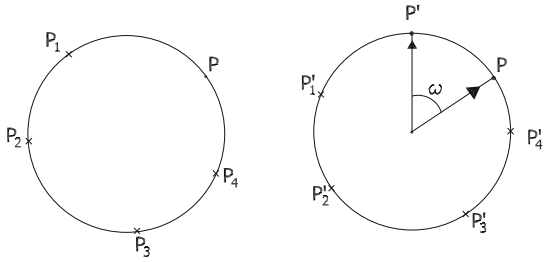


Fig. 5.4 A configuration (P, P_1, P_2, P_3, P_4) and its version $(P', P_1', P_2', P_3', P_4')$ rotated by ω

We observe that (5.8) is a homogeneous linear predictor, i.e. there is not a constant λ_0 in the formula; the reason is that, when we observe $y_1 = y_2 = \ldots y_N = 0$ we prefer the prediction of $u(P)$ to be zero too, i.e. its mean value on the circle, according to the hypothesis (5.5).

We notice also that, in (5.8), λ_i in general will be functions of $P, P_1 \ldots P_N$ but not of $\{y_i\}$, i.e.

$$\lambda_i = \lambda_i(P, P_1, \ldots, P_N) = \lambda_i(\vartheta, \vartheta_1, \ldots, \vartheta_N). \tag{5.9}$$

Whatever $\{\lambda_i\}$ we choose, the corresponding prediction error is

$$\begin{aligned} e(P, P_1, \ldots, P_N) &= u(P) - \widehat{u}(P) \\ &= u(P) - \sum_{i=1}^{N} \lambda_i u(P_i). \end{aligned} \tag{5.10}$$

If we don't have any particular further information on $u(P)$ (for instance that in some regions of C, $u(P)$ is smoother or rougher) it is reasonable to further restrict our class of predictors by requiring that λ_i be invariant under rotation. Namely, take two configuration, $\{P, P_1, \ldots, P_N\}$ and $\{P', P_1', \ldots, P_N'\}$ obtained one from the other by a rotation ω of the circle (see Fig. 5.4);

We claim that if in the first case we have decided that $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ are good coefficients for our prediction job, then the *same* coefficients should work for $\{P', P_1', \ldots, P_N'\}$ because if $(y_1, \ldots, y_N)$ are observed at $(P_1, \ldots, P_N)$ and $\widehat{u}(P)$ is our prediction, then in case we observe again $(y_1, \ldots, y_N)$ at $(P_1', \ldots, P_N')$ we want to make the *same* prediction at P'.

This is translated into analytical terms as follows: let R_ω be a *rotation operator* acting according to the law

$$\begin{aligned} &R_\omega F(P, P_1, \ldots, P_N) = R_\omega F(\vartheta, \vartheta_1, \ldots, \vartheta_N) \\ &= F(P', P_1', \ldots, P_N') = F(\vartheta', \vartheta_1', \ldots, \vartheta_N') \\ &= F(\vartheta + \omega, \vartheta_1 + \omega, \ldots, \vartheta_N + \omega) \end{aligned} \tag{5.11}$$

where F is any function of $(P, P_1, \ldots, P_N)$; then our invariance constraint is

$$\forall \omega, \quad F(\vartheta, \vartheta_1, \ldots, \vartheta_N) \equiv F(\vartheta + \omega, \vartheta_1 + \omega, \ldots, \vartheta_N + \omega) \tag{5.12}$$

A function F satisfying (5.12) must have a particular form, namely

$$F(\vartheta, \vartheta_1, \ldots, \vartheta_N) = G(\vartheta_1 - \vartheta, \vartheta_2 - \vartheta, \ldots, \vartheta_N - \vartheta); \tag{5.13}$$

this derives from (5.12) by choosing $\omega = -\theta$.

So we agree that our prediction coefficients must satisfy (5.11) and (5.13). Accordingly if we apply R_ω to e (cf. (5.10)), we get

$$\begin{aligned} R_\omega e(P, P_1, \ldots, P_N) &= R_\omega u(P) - \sum_{i=1}^{N} \lambda_i R_\omega u(P_i) \\ &= u(\vartheta + \omega) - \sum_{i=1}^{N} \lambda_i u(\vartheta_i + \omega), \end{aligned} \tag{5.14}$$

where λ_i are left unchanged by R_ω because of our invariance hypothesis.

Now observe that due to the very definition of R_ω the identity holds

$$R_\omega\{F^2(P, P_1, \ldots, P_N)\} \equiv \{R_\omega F(P, P_1 \ldots P_N)\}^2. \tag{5.15}$$

Next we define the *mean invariant quadratic prediction error*[1] as

$$\mathcal{E}^2(P, P_1, \ldots, P_N) \equiv \frac{1}{2\pi} \int_0^{2\pi} d\omega R_\omega\{e^2(P, P_1, \ldots, P_N)\}. \tag{5.16}$$

The adjective *invariant* is used for $\mathcal{E}^2$ because it is indeed a rotation invariant function of $(P, P_1, \ldots, P_N)$. In fact, (exploiting also (5.15)),

$$\begin{aligned} &\forall \eta, \quad R_\eta \mathcal{E}^2(P, P_1, \ldots, P_N) = \mathcal{E}^2(R_\eta P, R_\eta P_1, \ldots, R_\eta P_N) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\omega R_\omega\{e^2(R_\eta P, R_\eta P_1, \ldots, R_\eta P_N\} \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\omega R_\omega R_\eta\{e^2(P, P_1, \ldots, P_N\} \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\omega R_{\omega+\eta}\{e^2(P, P_1, \ldots, P_N\} \\ &= \mathcal{E}^2(P, P_1, \ldots, P_N\}, \end{aligned} \tag{5.17}$$

since integrating in $d\omega$ from 0 to 2π is one and the same thing as integrating from η to $\eta + 2\pi$.

[1] In this chapter we will use $\mathcal{E}^2$ for the mean quadratic prediction error; confusion should not be made with the same symbol $\mathcal{E}$ used elsewhere to denote the ellipsoid.

With the help of (5.14) and (5.15) we can indeed perform an explicit computation of $\mathcal{E}^2$, giving

$$\mathcal{E}^2 = \frac{1}{2\pi}\int_0^{2\pi} d\omega u^2(\vartheta+\omega) - 2\sum_{i=1}^{N}\lambda_i \frac{1}{2\pi}\int_0^{2\pi} d\omega u(\vartheta+\omega)u(\vartheta_i+\omega)$$
$$+ \sum_{i,k=1}^{N}\lambda_i\lambda_k \frac{1}{2\pi}\int_0^{2\pi} d\omega u(\vartheta_i+\omega)u(\vartheta_k+\omega) \tag{5.18}$$

It is noteworthy that by introducing the two points function

$$C(\vartheta,\vartheta') = \frac{1}{2\pi}\int_0^{2\pi} d\omega u(\vartheta+\omega)u(\vartheta'+\omega) \tag{5.19}$$

we come to express $\mathcal{E}^2$ in a concise form as

$$\mathcal{E}^2 = C(\vartheta,\vartheta) - 2\sum_{i=1}^{N}\lambda_i C(\vartheta,\vartheta_i) + \sum_{i,k=1}^{N}\lambda_i\lambda_k C(\vartheta_i,\vartheta_k). \tag{5.20}$$

A particularly important remark is that

$$C(\vartheta+\eta,\vartheta'+\eta) = \frac{1}{2\pi}\int_0^{2\pi} d\omega u(\vartheta+\eta+\omega)u(\vartheta'+\eta+\omega) = C(\vartheta,\vartheta')$$

for the same reason used in the proof of (5.17). Therefore $C(\vartheta,\vartheta')$ is also invariant under rotation, namely, with a small abuse of notation,

$$C(\vartheta,\vartheta') = C(\vartheta-\vartheta'). \tag{5.21}$$

The function $C(\vartheta-\vartheta')$ is called a rotation invariant covariance function. In particular it is called a covariance function because it has the typical properties of a covariance; it is symmetric and positive definite.

Such properties are immediately derived from (5.19), but we shall come back to the item at the end of the section.

Minimizing $\mathcal{E}^2$ with respect to $\{\lambda_i\}$ is straightforward and gives the following result: put

$$\begin{cases} \boldsymbol{\lambda} = \{\lambda_i\} \\ i = 1,\dots,N \end{cases} \quad \begin{cases} C = \{C(\vartheta_i-\vartheta_k)\} \\ i,k = 1,\dots,N \end{cases} \quad \begin{cases} \mathbf{C}_\vartheta = \{C(\vartheta-\vartheta_i)\} \\ i = 1,\dots,N \end{cases} \tag{5.22}$$

then

$$\boldsymbol{\lambda} = C^{-1}\mathbf{C}_\vartheta. \tag{5.23}$$

It is interesting to observe that since both the vector $\mathbf{C}_\vartheta$ and the matrix C are rotationally invariant, then so is $\boldsymbol{\lambda}$ too, as it was required form the beginning.

We make a fundamental remark on our solution. Remember that by definition a random field on C (see for instance Rozanov 1982) is a function $\{v(P,\omega)\}$, with $P \in C$ and $\omega \in \Omega$ and with a probability distribution on Ω, satisfying some measurability hypotheses, so that $\forall\{P_1, P_2, \ldots, P_N\}$ we know the probability distribution of the N-vector $\mathbf{v}^t(\omega) = [v(P_1,\omega), \ldots, v(P_N,\omega)]$. Remember also that mean and covariance of $\{v(P,\omega)\}$ are defined as

$$\mu(P) = E\{v(P,\omega)\} = \int_\Omega v(P,\omega)dP(\omega) \tag{5.24}$$

$$\begin{aligned} C(P,P') &= E\{[v(P,\omega)-\mu(P)][v(P',\omega)-\mu(P')\} \\ &= \int_\Omega v(P,\omega)v(P',\omega)dP(\omega) - \mu(P)\mu(P'). \end{aligned} \tag{5.25}$$

Here, as in the rest of the section, it occurs sometimes that the same symbol P is used to mean a point in space and a probability distribution, in which case it is always $P(\omega)$; moreover in this context Ω is an abstract set and not $\overline{B}^c$.

Now let us go back to our field $u(P) = u(\vartheta)$, with $u(\vartheta)$ a periodic function, and define a random field $\{v(\vartheta,\omega)\}$ as

$$v(\vartheta,\omega) = R_\omega u(P) = u(\vartheta+\omega) \tag{5.26}$$

with ω uniformly distributed on C, i.e.

$$\Omega = [0, 2\pi], \qquad dP(\omega) = \frac{d\omega}{2\pi}. \tag{5.27}$$

By applying (5.24) and (5.25) with (5.27), we see that $\mu(P) \equiv 0$ and that $C(P, P')$ is exactly the same covariance that we already defined in (5.19). Moreover if we construct a linear predictor of $v(P,\omega)$ by

$$\widehat{v}(P,\omega) = \sum_{i=1}^{N} \lambda_i v(P_i,\omega) \tag{5.28}$$

and we compute the prediction error

$$e(P,\omega) = v(P,\omega) - \widehat{v}(P,\omega),$$

we end up with the following expression for its variance

$$\begin{aligned} &\sigma^2[e(P,\omega)] \equiv E\{e^2(P,\omega)\} \\ &\equiv C(P,P) - 2\sum_{i=1}^{N} C(P,P_i)\lambda_i + \sum_{i,k=1}^{N} \lambda_i\lambda_k C(P_i,P_k) \\ &\equiv \mathcal{E}^2(P,P_1,\ldots,P_N). \end{aligned} \tag{5.29}$$

Indeed minimizing (5.9) with respect to $\{\lambda_i\}$ is the same problem as minimizing (5.20) and therefore it has the same solution.

This settles the first corner stone of a quite general theorem of equivalence of different approaches, all producing the same type of linear predictors, so that each approach contributes to the theoretical and practical understanding of the collocation theory developed in the next sections.

5.4 On Collocation Theory, or the Wiener-Kolmogorov Principle Applied in Physical Geodesy

We want to generalize the example of the previous section, switching from the circle C to the sphere S_R, from the rotation R_ω on C to a 3D rotation $R_{\boldsymbol{\omega}}$, where $\boldsymbol{\omega}$ now becomes a triple of angles (for instance Euler angles), so as to apply the minimization of a suitably defined invariant quadratic error, or equivalently a minimum prediction error variance principle, to our field $u(P) = T_r(P)$, harmonic outside S_R.

This discussion parallels a similar discussion, already dating back to 1940/1950, among scientists working in signal analysis and stochastic processes theory. In that framework N. Wiener was more stressing the point of view of the invariant estimators, while A. Kolmogorov was more in favour of the pure stochastic interpretation. It is for this reason that we like to label our application in physical geodesy of such a principle after the names of both great scientists.

The method, known in Geodesy as *collocation*, was developed in 1960–1970 by Moritz and Krarup (see Moritz 1980; Krarup 2006, Chap. 4), again one stressing the stochastic, the other the deterministic interpretation. Here we like to follow more the already mentioned point of view of proving the possibility of interpreting in different ways equivalent results, thus giving a clearer perspective to their practical implementation.

The first item we need to settle is to find an analogous of the uniform mean over rotated configurations of N points $\{P_1, \ldots, P_N\}$.

Without going into more difficult mathematical arguments on group theory, for which we refer to literature Moritz (1980) and Sansò and Venuti (2002a), we simply aim at giving a definition, proving that this provides a result with the required properties.

We start by defining the action of the rotation operator R_ω as

$$R_\omega F(P_1, \ldots, P_N) = F(R_\omega P_1, \ldots, R_\omega P_N) \tag{5.30}$$

and we ask ourselves how an invariant F should be made [2]

[2] Often in group theory the inverse rotation matrix R_ω^t is used; since this is irrelevant in the present text and this is not useful, we stick to definition (5.30).

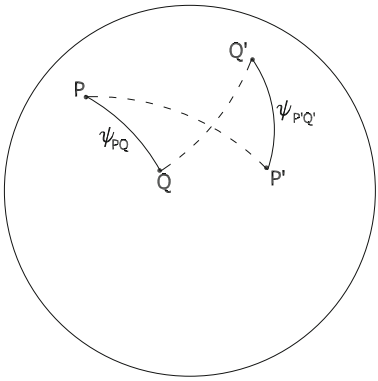


Fig. 5.5 The characterization of a rotation R_ω through the rigid motion of the arc $\overset{\frown}{PQ}$ over a sphere, $(P' = R_\omega P,\ Q' = R_\omega Q)$

Since under R_ω the polyhedron $\{P_1, \ldots, P_N\}$ is rigidly moved to another one $\{P'_1, \ldots, P'_N\}$, leaving the origin of R^3 fixed, we see that the following conditions are satisfied

$$r_{P'_i} = r_{P_i}; \quad \psi_{P'_i P'_j} = \psi_{P_i P_j}, \tag{5.31}$$

where we have denoted as usual with ψ_{PQ} the angle between $\mathbf{r}_P$ and $\mathbf{r}_Q$. It is easy to see that (5.31) is not only necessary but also sufficient for a rigid motion of $(P_1, \ldots, P_N)$ in the three-dimensional space, with the origin fixed in O. Therefore $F(P_1, \ldots, P_N)$ will be invariant under rotation if

$$F(P_1, \ldots, P_N) = F(\ldots r_{P_i} \ldots\ ;\ \ldots \psi_{P_i P_j} \ldots). \tag{5.32}$$

Next we note that in order to characterize a 3D rotation we need only to show how it acts on two points P, Q placed on a sphere.

Namely there is one and only one rotation sending PQ to $P'Q'$ on condition that $\psi_{P'Q'} = \psi_{PQ}$ (see Fig. 5.5) and $r_P = r_Q = r_{P'} = r_{Q'}$.

Since all what we shall really use in the sequel is the average of a two-points function, we concentrate on that, knowing that in any way the definition can be generalized to N points, in case of need. So let $F(P, Q)$ be any regular function of two points defined e.g. on the unit sphere; we put by definition

$$\begin{aligned} &E\{R_\omega[F(P, Q)]\} \\ &= \int dP(\omega) R_\omega F(P, Q) \\ &= A \int d\sigma_{P'} \int_{\psi_{P'Q'} = \psi_{PQ}} F(P', Q') d\alpha_{Q'}, \end{aligned} \tag{5.33}$$

where P' sweeps the whole unit sphere, while, for each fixed P', Q' runs on a circle of spherical radius ψ_{PQ}, occupying all the points of different azimuth α. The variable α ranges from 0 to 2π (see Fig. 5.6).

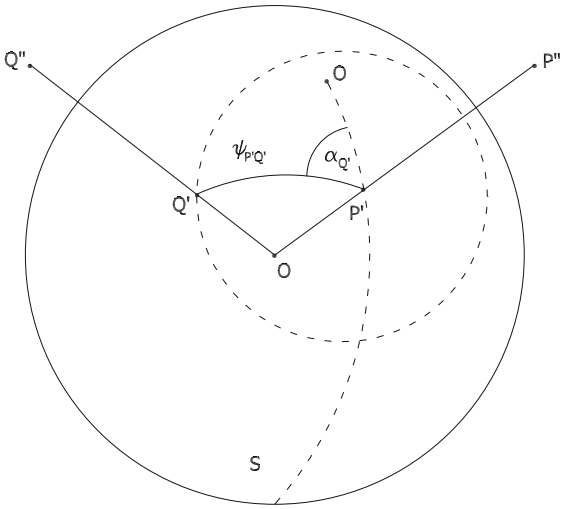


Fig. 5.6 Representation of the integration variable of (5.33): O is the center of S, P'', Q'' are in space, while P', Q' are their projection on S

As it obvious at the end the function (5.33) will depend on P, Q only through ψ_{PQ}, i.e. it will be invariant. Even if the points P, Q were outside the unit sphere, it is clear that (5.33) would depend in the end only on $r_{P''} = r_P, r_{Q''} = r_Q$ and ψ_{PQ} (see Fig. 5.6). So we can say that in general

$$E\{R_\omega[F(P,Q)]\} = C_F(r_P, r_Q, \psi_{PQ}), \tag{5.34}$$

i.e. it is a rotation invariant function. As for the normalization constant A appearing in (5.33), this is determined by considering that $dP(\omega)$ has to be a (uniform) probability distribution, so that one must have

$$E\{1\} = A \int d\sigma_{P'} \int_0^{2\pi} d\alpha_{Q'} = A \cdot 8\pi^2 \equiv 1,$$

implying

$$A = \frac{1}{8\pi^2}. \tag{5.35}$$

Now we can repeat the same reasoning as in Sect. 5.3. Namely if the observations y_i are just $u(P_i)$, $i = 1 \ldots N$, we define a *linear invariant* predictor

$$\widehat{u}(P) = \sum_{i=1}^{N} \lambda_i u(P_i), \tag{5.36}$$

with λ_i such that

$$R_\omega \lambda_i \equiv \lambda_i,$$

and an invariant quadratic prediction error

$$\begin{aligned}\mathcal{E}^2 &= E_\omega\{R_\omega[u(P) - \widehat{u}(P)]^2\} \\ &= C(P,P) - 2\sum_{i=1}^{N}\lambda_i C(P,P_i) + \sum_{i,k=1}^{N}\lambda_i\lambda_k C(P_i,P_k)\end{aligned} \tag{5.37}$$

where we have put

$$C(P,Q) = E\{R_\omega[u(P)u(Q)]\} = \frac{1}{8\pi^2}\int d\sigma_{P'} \int\limits_{\psi_{P'Q'}=\psi_{PQ}} d\alpha_{Q'}u(P')u(Q'), \tag{5.38}$$

also called the covariance function $u(P)$. Just as in (5.23), the minimum of (5.37) is achieved by

$$\lambda_j = \sum_{k=1}^{N} C_{jk}^{(-1)} C(P_k,P) \tag{5.39}$$

and the corresponding value of $\mathcal{E}^2$ is

$$\mathcal{E}^2_{\min} = C(P,P) - \sum_{i,j=1}^{N} C(P,P_i)C_{ij}^{(-1)}C(P_j,P). \tag{5.40}$$

In (5.39) and (5.40) we have used the short notation $C_{ik}^{(-1)}$, to mean the element (i,k) of the matrix C^{-1}, inverse of $C \equiv \{C(P_i,P_k)\}$.

Let us note that again the possibility of using a predictor like (5.39) depends on the availability of the covariance function of u, (5.38); for the moment we just assume it is known and we shall explain later how to estimate it from data.

As in Sect. 5.3 we observe that, if we define a random field v,

$$v(P,\omega) = R_\omega u(P) \tag{5.41}$$

and we postulate a uniform distribution of ω on the 3D rotation group, we receive a totally equivalent problem with the same analytical solution, on condition that

$$\begin{aligned}E_\omega\{v(P,\omega)\} &= \frac{1}{8\pi^2}\int d\sigma_P u(P)\int_0^{2\pi} d\alpha_Q \\ &= \frac{1}{4\pi}\int d\sigma_P u(P) = 0,\end{aligned} \tag{5.42}$$

what we assume to be true, because by hypothesis $u(P) \equiv T_r(P)$ and $T_r(P)$ certainly has a zero mean on any sphere centered at the origin. Note as well that

calling $C(P, Q)$, in (5.38), a covariance function, we are consistent with a standard terminology for random fields.

5.5 The General Collocation Problem

Based on the discussion of Sects. 5.3 and 5.4, from now on we accept the equivalence principle stating that we can proceed with our prediction algorithms either by minimizing the invariant quadratic error in a class of invariant linear estimators or by introducing the model of a random field, as in (5.41), and minimizing the mean square prediction error in a class of linear predictors. Invariant here means invariant with respect to the 3D rotation group, and expectation means averaging over a uniform distribution on the rotation group.

Let us first of all state our problem in the following form: we have observation equations

$$y_i = M_i(u) + v_i, \quad \imath = 1 \dots N \tag{5.43}$$

and we want to predict a functional of u, $L(u)$ by means of a linear homogenous predictor, i.e.

$$L(\widehat{u}) = \sum_{i=1}^{N} \lambda_i y_i; \tag{5.44}$$

to do that we want to apply the Wiener-Kolmogorov (W-K) principle.

To this aim we need to define clearly what is an admissible functional L applied to the random process $v(P, \omega)$.

In fact note that $v(P, \omega) = R_\omega u(P) = u(R_\omega P)$, is a function of two variables and that L will act only on the variable P, so that we expect

$$Y_0 = L_P\{v(P, \omega)\} \tag{5.45}$$

to be a (measurable) function of ω only, i.e. a random variable.

We note that, under suitable regularity conditions,

$$\begin{aligned} E_\omega\{Y_0\} &= E_\omega\{L_P[v(P, \omega)]\} \\ &= \int dP(\omega) L_P\{R_\omega u(P)\} \\ &= L_P\{\int dP(\omega) R_\omega u(P)\} \\ &= L_P\{E_\omega\{v(P, \omega)\}\} = 0, \end{aligned} \tag{5.46}$$

so we expect that all useful random variables of the type (5.45) have zero mean (with respect to ω).

Definition 1 (Admissible functionals). We state that a *functional* $L_P(\,)$ *is admissible*, if the corresponding random variable Y_0 has finite variance.

Namely we require that

$$\begin{aligned} E_\omega\{Y_0^2\} &= \int dP(\omega) L_P[u(R_\omega P)] L_Q[u(R_\omega Q)] \\ &= L_P\{L_Q\{\int dP(\omega) u(R_\omega P) u(R_\omega Q)\}\} \\ &= L_P\{L_Q C(P,Q)\} < +\infty. \end{aligned} \tag{5.47}$$

Covariance propagation. The above computation can be repeated when we need to compute the covariance

$$\begin{aligned} &E\{L_P[v(P,\omega)] M_Q[v(Q,\omega)]\} \\ &\quad = L_P\{M_Q\{E[v(P,\omega)v(Q,\omega)]\}\} \\ &\quad = L_P\{M_Q C(P,Q)\}. \end{aligned} \tag{5.48}$$

Formula (5.48) is in fact the covariance propagation formula for random fields.

To simplify formulas, from now on we shall use the short-hand notation (see Krarup 2006, Chap. 15)

$$\begin{cases} L_P C(P,Q) = C(L,Q) \\ L_P\{M_Q C(P,Q)\} = C(L,M). \end{cases} \tag{5.49}$$

Moreover we note that if we take a vector of functionals

$$\mathbf{L} = \begin{vmatrix} L_1(\cdot) \\ L_2(\cdot) \\ \vdots \\ L_N(\cdot) \end{vmatrix} \tag{5.50}$$

and we put

$$\mathbf{Y} = \mathbf{L}\{v(P,\omega)\}, \tag{5.51}$$

then indeed $\mathbf{Y}$ has zero mean,

$$E\{\mathbf{Y}\} = 0,$$

and a covariance matrix $C_{\mathbf{YY}}$ given by

$$\{C_{Y_i Y_k}\} = \{C(L_i, L_k)\} \tag{5.52}$$

which we write in vector form as

$$C_{\mathbf{YY}} = C(\mathbf{L}, \mathbf{L}^t). \tag{5.53}$$

Naturally $C(\mathbf{L}, \mathbf{L}^t)$ is symmetric and positive definite. Similarly the cross-covariance between the vector $\mathbf{Y}$ of (5.51) and $\mathbf{Z} = \mathbf{M}\{v(P, \omega)\}$ is just the matrix

$$C_{\mathbf{YZ}} = E\{\mathbf{Y}\mathbf{Z}^t\} = \{C(L_i, M_k)\} = C(\mathbf{L}, \mathbf{M}^t). \tag{5.54}$$

Now the last thing we need in order to perform our prediction is just to observe that in our models we have two stochastic quantities, the random field $v(P, \omega)$ and the noise vector $\boldsymbol{\nu}$. So we need first of all to represent the stochastic interaction between the two and then we need to warn the reader that when we shall use the expectation symbol $E\{\ \}$, without any particular index, we will mean averaging with respect to all random variables, while we shall use $E_\omega\{\ \}$ or $E_\nu\{\ \}$ when we want to perform an average with respect to a specific random variable.

To complete the hypotheses on the covariance structure of the problem we summarize them as follows:

$$E\{v(P, \omega)\} \equiv 0,\ E\{v(P, \omega)v(Q, \omega)\} = C(P, Q), \tag{5.55}$$

with $C(P, Q)$ a given invariant covariance function and with the propagation rule (5.48) for the covariances of linear functionals of v;

$$E\{\boldsymbol{\nu}\} = 0,\ \ E\{\boldsymbol{\nu}\boldsymbol{\nu}^t\} = C_{\nu\nu}; \tag{5.56}$$

furthermore we shall assume that the noise $\boldsymbol{\nu}$ and the random field v are linearly independent, i.e.

$$E\{v(P, \omega)\nu_i\} = 0,\ \ \forall P,\ \forall i, \tag{5.57}$$

implying also that for any admissible functional L,

$$E\{L_P[v(P, \omega)]\nu_i\} = 0. \tag{5.58}$$

With all these rules of calculus we proceed to establish the W-K principle, namely we start to compute the variance of the prediction error.

Remember that the observation equations and the linear predictor $\widehat{L(v)}$ were defined in (5.43) and (5.44), which we can write in vector form as

$$\mathbf{Y} = \mathbf{M}\{v\} + \boldsymbol{\nu} \tag{5.59}$$

$$L_P[\widehat{v(P}, \omega)] = \boldsymbol{\lambda}^t\mathbf{Y}. \tag{5.60}$$

If $\widehat{L(v)}$ is our predictor, the prediction error is

$$\begin{aligned} e(\omega) &= L(v) - \widehat{L(v)} \\ &= L(v) - \boldsymbol{\lambda}^t\mathbf{Y} \end{aligned} \tag{5.61}$$

and its variance can be computed by

$$\begin{aligned}\mathcal{E}^2 &= E\{e^2(\omega)\} = E\{L(v)^2\} + \qquad (5.62)\\ &\quad - 2E\{\boldsymbol{\lambda}^t \mathbf{Y} L(v)\} + E\{(\boldsymbol{\lambda}^t \mathbf{Y})^2\}\\ &= C(L, L) - 2\boldsymbol{\lambda}^t E\{\mathbf{Y} L(v)\} + \boldsymbol{\lambda}^t C_{\mathbf{YY}} \boldsymbol{\lambda}. \qquad (5.63)\end{aligned}$$

On the other hand

$$\begin{aligned}E\{\mathbf{Y} L(v)\} &= E\{\mathbf{M}(v) L(v)\} + E\{\boldsymbol{\nu} L(v)\} \qquad (5.64)\\ &= C(\mathbf{M}, L);\\ C_{\mathbf{YY}} &= E\{\mathbf{Y}\mathbf{Y}^t\} = E\{[\mathbf{M}(v) + \boldsymbol{\nu}][\mathbf{M}(v) + \boldsymbol{\nu}]^t\}\\ &= E\{\mathbf{M}(v)\mathbf{M}^t(v)\} + E\{\boldsymbol{\nu}\boldsymbol{\nu}^t\} = C(\mathbf{M}, \mathbf{M}^t) + C_{\boldsymbol{\nu}\boldsymbol{\nu}}. \qquad (5.65)\end{aligned}$$

Substituting in (5.62) we can then invoke the W-K principle claiming that the optimal predictor is the one that minimizes $\mathcal{E}^2$, namely the solution of the *normal equation system*

$$C_{\mathbf{YY}} \boldsymbol{\lambda} = C(\mathbf{M}, L) \qquad (5.66)$$

or

$$\boldsymbol{\lambda} = C_{\mathbf{YY}}^{-1} C(\mathbf{M}, L) \qquad (5.67)$$

with $C_{\mathbf{YY}}$ given by (5.65).

Going back to (5.60) we find the W-K predictor

$$\widehat{L(v)} = C(L, \mathbf{M}^t) C_{\mathbf{YY}}^{-1} \mathbf{Y} \qquad (5.68)$$

and substituting into (5.62) we get its squared prediction error as

$$\mathcal{E}^2 = C(L, L) - C(L, \mathbf{M}^t) C_{\mathbf{YY}}^{-1} C(\mathbf{M}, L). \qquad (5.69)$$

Formulas (5.68) and (5.69) are so important that it is worth representing them explicitly in components, namely

$$\widehat{L(v)} = \sum_{k,i=1}^{N} L_P\{M_{P_k} C(P, P_k)\} C_{Y_k Y_i}^{(-1)} Y_i \qquad (5.70)$$

with $C_{Y_k Y_i}^{(-1)}$ the element (k, i) of the inverse of the matrix $C_{\mathbf{YY}}$, i.e.

$$C_{Y_k Y_i} = M_{P_k}\{M_{P_i} C(P_k, P_i)\} + C_{\nu_k \nu_i}; \qquad (5.71)$$

moreover

$$\mathcal{E}^2 = L_P\{L_Q C(P,Q)\} + \tag{5.72}$$

$$-\sum_{k,i=1}^{N} L_P\{M_{P_k} C(P,P_k)\} C^{(-1)}_{Y_k Y_i} L_P\{M_{P_i} C(P_i,P)\}. \tag{5.73}$$

We note that in most cases $C_{v_k v_i}$ is diagonal and, when $M_k(\ )$ are functionals representing the same type of measurement, many times we put $C_{\nu\nu} = \sigma_v^2 I$, although this is not really necessary in our formulas that represent the most general case.

Example 2. We want already here to specify how formulas (5.70), (5.72) work for the most prominent case of this book, namely the prediction of the anomalous potential $T(P)$ (loosely speaking one could say the geoid prediction) from observed pointwise gravity anomalies $\Delta g(P_i)$, $i = 1 \dots N$.

Let us remember that here $T(P)$ and $\Delta g(P)$ mean the residual anomalous potential and the residual gravity anomaly. We mention that in this case $L(\ )$, the functional to be predicted, is just the evaluation of T at the point P,

$$L(T) = ev_P(T) = T(P).$$

As for the gravity anomaly at P, we can usefully reason as follows; first we define a *gravity anomaly operator* A which actually transforms the function $T(P)$ into another function $\Delta g(P)$

$$\Delta g(P) = A(T) \equiv -\frac{\partial T}{\partial h}(P) + \frac{\gamma'}{\gamma} T(P), \tag{5.74}$$

then we evaluate the field $\Delta g(P)$ at a specific measurement point P_k,

$$\begin{aligned} M_k(T) &= ev_{P_k}\{A(T)\} \\ &= \Delta g(P_k). \end{aligned} \tag{5.75}$$

Put in this way we understand that to compute the covariance of M_k, M_i or that of M_k, L one can proceed in two steps. First we define a *covariance function* of $\Delta g(P)$ according to

$$\begin{aligned} C_{\Delta g \Delta g}(P,Q) &= E\{\Delta g(P)\Delta g(Q)\} \\ &= E\{A_P[v(P,\omega)]A_Q[v(Q,\omega)\} \\ &= A_P\{A_Q C(P,Q)\} \end{aligned} \tag{5.76}$$

where

$$v(P,\omega) = R_\omega T(P)\ ; \tag{5.77}$$

then we apply the evaluation at specific measurement points, namely

$$\begin{aligned} C(M_k, M_i) &= ev_{P_k}\{ev_{P_i} C_{\Delta g \Delta g}(P_k, P_i)\} \\ &= C_{\Delta g \Delta g}(P_k, P_i). \end{aligned} \tag{5.78}$$

Accordingly we define the *cross covariance* between the two fields $T(P)$ and $\Delta g(P)$ as

$$C_{T\Delta g}(P, Q) = E\{v(P, \omega) A_Q[v(Q, \omega)]\} = A_Q C(P, Q),$$

with v given by (5.77) and then we evaluate T at a particular point P and Δg at a particular point P_k, thus obtaining

$$\begin{aligned} C(P, M_k) &= ev_P\{ev_{P_k} C_{T\Delta g}(P, P_k)\} \\ &= C_{T\Delta g}(P, P_k). \end{aligned} \tag{5.79}$$

With the above specified rules, the best linear predictor, or collocation predictor of $T(P)$ is (see (5.70))

$$\widehat{T}(P) = \sum_{k,i} C_{T\Delta g}(P, P_k)\{C_{\Delta g \Delta g}(P_k, P_i) + \sigma^2_{\Delta g}\delta_{ik}\}^{(-1)} \Delta g_{\text{obs}}(P_i), \tag{5.80}$$

in (5.80) we have assumed that $C_{v_i v_k} = \sigma^2_{\Delta g}\delta_{ik}$ and we have written $\Delta g_{\text{obs}}(P_i)$ for Y_i.

The corresponding prediction error then becomes (see (5.72)).

$$\begin{aligned} \mathcal{E}^2 = C(P, P) + \\ - \sum_{k,i} C_{T\Delta g}(P, P_k)\{C_{\Delta g \Delta g}(P_k, P_i) + \sigma^2_{\Delta g}\delta_{ik}\}^{(-1)} C_{\Delta g T}(P_i, P) \end{aligned} \tag{5.81}$$

Remark 1. Recalling the definition of covariance of a function $T(P)$ (see (5.38)) namely

$$\begin{aligned} C(P, Q) &= E\{R_\omega T(P) R_\omega T(Q)\} \\ &= \int dP(\omega) T(R_\omega P) T(R_\omega Q) \end{aligned} \tag{5.82}$$

we see that, when T is a regular harmonic function,

$$\Delta_P C(P, Q) = \int dP(\omega) \Delta_P T(R_\omega P) T(R_\omega Q) \equiv 0, \tag{5.83}$$

in fact it is known that the Laplace operator is invariant under rotation, so that if $T(x, y, z)$ is harmonic as function of (x, y, z) and R_ω sends (x, y, z) into (x', y', z') then (see Exercise 1 in Sect. 5.12)

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)T(x',y',z')$$
$$\equiv\left(\frac{\partial^2}{\partial x'^2}+\frac{\partial^2}{\partial y'^2}+\frac{\partial^2}{\partial z'^2}\right)T(x',y',z')=0.$$

Naturally (5.83) implies $\Delta_Q C(P,Q)=0$ as well, because $C(P,Q)$ is a symmetric function of P and Q.

Now take a general collocation formula with $L_P = ev_P$ and $\{M_k\}$ whatever; similarly to (5.80), if we put

$$\xi_k = \Sigma_i\{C(M_k,M_i)+C_{v_k v_i}\}^{(-1)}Y_i \tag{5.84}$$

we see that the collocation predictor of $T(P)$ can be written as

$$\widehat{T(P)}=\sum_{k=1}^{N}C(P,M_k)\xi_k. \tag{5.85}$$

If we let P free to vary over $\Omega_R\equiv\{r_P\geq R\}$, we can interpret (5.85) more as an approximation of the whole function $T(P)$ than as a pointwise prediction. As such we see that our approximate solution $\widehat{T(P)}$ is automatically harmonic, namely

$$\Delta_P\widehat{T}(P)=\sum_{k=1}^{N}\Delta_P C(P,M_k)\xi_k\equiv 0. \tag{5.86}$$

This is indeed a nice property of our approximation theory.

5.6 Covariance and Spectral Harmonic Calculus

The functions $\widehat{T}(P)$ by which we do approximate the residual potential $T_r(P)$ are all harmonic in Ω_R, as stated in the previous section (Remark 1).

Therefore these functions can be represented by the convergent series

$$T(P)=\sum_{n,m=2}^{+\infty}\sum_{m=-n}^{n}T_{nm}S_{nm}(r_P,\vartheta_P,\lambda_P) \tag{5.87}$$

$$S_{nm}(r_P,\vartheta_P,\lambda_P)=\left(\frac{R}{r_P}\right)^{n+1}Y_{nm}(\vartheta_P,\lambda_P).$$

If we apply to $T(P)$, given by (5.87), the rotation operator we get, with $P'=R_\omega P$,

$$R_\omega T(P)=T(P')=\sum_{n=2}^{+\infty}\sum_{m=-n}^{n}T_{nm}\left(\frac{R}{r_P}\right)^{n+1}Y_{nm}(\vartheta_{P'},\lambda_{P'}); \tag{5.88}$$

on the same time we can state that

$$T(P') = \sum_{n=2}^{+\infty} \sum_{m=-n}^{n} T_{nm}(\omega) \left(\frac{R}{r_P}\right)^{n+1} Y_{nm}(\vartheta_P, \lambda_P), \tag{5.89}$$

because indeed $T(P')$ is also harmonic as a function of P. Naturally the harmonic coefficients of $T(P')$ as function of P, are not the same T_{nm} which appear in (5.87) and in particular they will depend on the relation between P' and P, namely on the specific rotation R_ω applied; this is why we have denoted them $T_{nm}(\omega)$.

We want to study the property of the functionals of T,

$$(P \in S_R), \qquad T_{nm}(\omega) = \frac{1}{4\pi} \int Y_{nm}(\vartheta_P, \lambda_P) T(R_\omega P) d\sigma_P \tag{5.90}$$

and their relation to the original T_{nm}. First of all we notice that, as for all admissible functionals, $E\{T_{nm}(\omega)\} = 0$ and

$$E\{T_{nm}(\omega) T_{jk}(\omega)\} = \frac{1}{(4\pi)^2} \int d\sigma_P \int d\sigma_Q Y_{nm}(\vartheta_P, \lambda_P) Y_{jk}(\vartheta_Q, \lambda_Q) \cdot \\ \cdot E\{T(R_\omega P) T(R_\omega Q)\}. \tag{5.91}$$

On the other hand since the covariance of T is spherically invariant ($P' = R_\omega P$, $Q' = R_\omega Q$),

$$E\{T(R_\omega P) T(R_\omega Q)\} = C(\psi_{P'Q'}) = C(\psi_{PQ}). \tag{5.92}$$

As a function of ψ, $C(\psi)$ is also a function of $\cos\psi$ so that we can write

$$t = \cos\psi; \qquad \overline{C}(t) \equiv C(\psi) = \sum_{n=0}^{+\infty} c_n P_n(t) \\ = \sum_{n=0}^{+\infty} c_n P_n(\cos\psi) \tag{5.93}$$

with (see (3.46))

$$c_n = \frac{2n+1}{2} \int_{-1}^{1} \overline{C}(t) P_n(t) dt \\ = \frac{2n+1}{2} \int_0^{\pi} C(\psi) P_n(\cos\psi) \sin\psi d\psi \tag{5.94}$$

Therefore, recalling the summation rule (3.54), we can substitute in (5.91) and (5.92)

$$C(\psi_{PQ}) = \sum_{p,q=0}^{+\infty} c_p (2p+1)^{-1} Y_{pq}(\vartheta_P, \lambda_P) Y_{pq}(\vartheta_Q, \lambda_Q) \tag{5.95}$$

so that by virtue of the orthogonality of $\{Y_{nm}(\vartheta, \lambda)\}$ we find

$$E\{T_{nm}(\omega) T_{jk}(\omega)\} = \Sigma_{p,q} \frac{c_p}{2p+1} \delta_{pn} \delta_{qm} \delta_{pj} \delta_{qk} = \frac{c_n}{2n+1} \delta_{nj} \delta_{mk}. \tag{5.96}$$

Hence $T_{nm}(\omega)$ are uncorrelated to one another and their variances are the same for all orders in degree n,

$$\sigma^2(T_{nm}) = \sigma_n^2 = \frac{c_n}{2n+1} \tag{5.97}$$

We will call σ_n^2 the degree variances of individual coefficients and c_n the full power degree variances. Although this name has already been used in (3.173) we shall soon see that we are justified in using it here because we will prove that c_n is identical with $\overline{\sigma}_n^2$ given in (3.173).

In fact the following remarkable result holds (see also Moritz 1980).

Lemma 1. *The distribution of* $\mathbf{T}_n \equiv \{T_{nm}\}$ *in* R^{2n+1} *(remember that we have* $2n+1$ *orders for each degree* n*) is singular, its support is the sphere with squared radius*

$$|\mathbf{T}_n(\omega)|^2 = \sum_{m=-n}^{n} T_{nm}^2(\omega) = c_n \tag{5.98}$$

and in fact $\mathbf{T}_n(\omega)$ *is uniformly distributed on this sphere.*

There are two consequences of this lemma: the first is that if we know even approximate values for T_{nm}, we can directly estimate $C(\psi_{PQ})$, given by (5.93), with $c_n = \Sigma_m T^2{}_{nm}$.

Namely the harmonic coefficients of one particular function given on S_R, provide us the degree variances of the process generated by randomly rotating this function.

We notice here as well that the formula $c_n = \Sigma_m T^2{}_{nm}$ justifies the name given to c_n of full power degree variances, in fact we can verify now that $c_n = \overline{\sigma}_n^2$ according to the previous definition on (3.173).

The other consequence is that the Lemma gives an answer to a guess popping up from times to times in geodesy, that the distribution of T, and then for instance of $\{T_{nm}\}$ too, could be normal (cf. Jekeli 1991). Indeed this is not possible in a strict sense, as observed long ago by Lauritzen (see Lauritzen 1973), because then $\{T_{nm}\}$ for fixed n would all be independent with zero mean and variance σ_n^2, what would imply that

$$\Sigma_m T^2{}_{nm}(\omega) \sim \sigma_n^2 \chi_{2n+1}^2, \tag{5.99}$$

i.e. it cannot be a constant with respect to ω. Yet (5.99) shows that this variable has a variance tending to zero. In fact (5.99) implies

$$\sigma^2[\Sigma_m T_{nm}^2(\omega)] = \sigma_n^4 2 \cdot (2n+1)$$

which must tend to zero since

$$\Sigma_n \sigma_n^2 (2n+1) = \Sigma_n \Sigma_m T_{nm}^2 < +\infty \tag{5.100}$$

by hypothesis.

Indeed (5.100) implies $\sigma_n^2(2n+1) \to 0$ and, a fortiori, $\sigma_n^4(2n+1) \to 0$. So from the practical point of view the field T, at least above a certain degree, could still be approximately normal.

The use of (5.97) simplifies the calculation of various covariances and cross-covariances for fields which have an easy spectral representation, as we show in the next example.

Example 3. As we have seen in (5.80), to apply the present theory to the determination of a gravimetric quasi-geoid we need $C_{T\Delta g}(P, Q)$ and $C_{\Delta g \Delta g}(P, Q)$. If we apply the spherical approximation formula (cf. (5.101))

$$\Delta g = -\frac{\partial T}{\partial r} - \frac{2}{r} T$$

that, in terms of harmonic coefficients translates into

$$\Delta g_{nm} = \frac{n-1}{R} T_{nm}, \tag{5.101}$$

we get straightforwardly

$$E\{T_{nm} \Delta g_{jk}\} = \delta_{nj} \delta_{mk} \frac{n-1}{R} \sigma_n^2(T) \tag{5.102}$$

and

$$E\{\Delta g_{nm} \Delta g_{jk}\} = \delta_{nj} \delta_{mk} \frac{(n-1)^2}{R^2} \sigma_n^2(T). \tag{5.103}$$

implying

$$c_n(\Delta g) = \frac{(n-1)^2}{R^2} c_n(T). \tag{5.104}$$

With these rules we can put

$$C_{T\Delta g}(P,Q) = \Sigma_{n,m}\frac{(n-1)}{R}\sigma_n^2(T)S_{nm}(r_P,\vartheta_P,\lambda_P)S_{nm}(r_Q,\vartheta_Q,\lambda_Q)$$
$$= \Sigma_n\frac{(n-1)}{R}\sigma_n^2(T)\left(\frac{R^2}{r_P r_Q}\right)^{n+1}(2n+1)P_n(\cos\psi_{PQ}) \tag{5.105}$$

and

$$C_{\Delta g\Delta g}(P,Q) = \Sigma_{n,m}\frac{(n-1)^2}{R^2}\sigma_n^2(T)S_{nm}(r_P,\vartheta_Q,\lambda_P)S_{nm}(r_Q,\vartheta_Q,\lambda_Q)$$
$$= \Sigma_n\frac{(n-1)^2}{R^2}\sigma_n^2(T)\left(\frac{R^2}{r_P r_Q}\right)^{n+1}(2n+1)P_n(\cos\psi_{PQ}). \tag{5.106}$$

Let us note that in particular (5.106) coincides, in spherical approximation, with (5.76).

It is useful to observe that not all the fields that can be derived from T possess a spherical invariant covariance, although the spectral calculus, when applicable, facilitates the calculations as the next example shows.

Example 4. We want to compute the covariance of $T_\lambda = \frac{\partial T}{\partial\lambda}$. Note that this quantity is just the eastern deflection of the vertical η multiplied by $r\sin\vartheta$. To this aim let us observe that, according to our definition of $Y_{nm}(\vartheta,\lambda)$ (cf. (3.50) and (3.51)) we have

$$\frac{\partial}{\partial\lambda}Y_{nm}(\vartheta,\lambda) = -mY_{n,-m}(\vartheta,\lambda). \tag{5.107}$$

But then

$$T_\lambda(P) = \Sigma_{n,m}(-m)T_{nm}S_{n,-m}$$
$$= \Sigma_{n,m}mT_{n,-m}S_{n,m}$$

or

$$(T_\lambda)_{nm} = mT_{n,-m}. \tag{5.108}$$

The last relation implies

$$E\{(T_\lambda)_{nm}^2\} = m^2\sigma_n^2(T) \tag{5.109}$$

so that T_λ has not *degree variances*, i.e. the variances of $(T_\lambda)_{nm}$ are not the same for all orders m.

It is useful here to observe that the covariance of T_λ can also be derived directly from $C(P, Q)$ with the following formula

$$C_{T_\lambda T_\lambda}(P, Q) = E\{T_\lambda(P)T_\lambda(Q)\} = \frac{\partial^2}{\partial\lambda_P \partial\lambda_Q} C(P, Q). \tag{5.110}$$

If we put

$$C(P, Q) = C(r_P, r_Q, \psi_{PQ}) = \overline{C}(r_P, r_Q, \cos\psi_{PQ}) \tag{5.111}$$

and we note that

$$\cos\psi_{PQ} = \sin\vartheta_P \sin\vartheta_Q \cos(\lambda_P - \lambda_Q) + \cos\vartheta_P \cos\vartheta_Q$$

so that

$$\frac{\partial}{\partial\lambda_Q} \cos\psi_{PQ} = \sin\vartheta_P \sin\vartheta_Q \sin(\lambda_P - \lambda_Q)$$

and

$$\frac{\partial^2}{\partial\lambda_P \partial\lambda_Q} \cos\psi_{PQ} = \sin\vartheta_P \sin\vartheta_Q \cos(\lambda_P - \lambda_Q),$$

we can compute (5.110).

Put

$$\overline{C}' = \frac{\partial}{\partial t}\overline{C}(r_P, r_Q, t)$$
$$\overline{C}'' = \frac{\partial^2}{\partial t^2}\overline{C}(r_P, r_Q, t);$$

then you find

$$C_{T_\lambda T_\lambda} = \overline{C}'(r_P, r_Q, \cos\psi_{PQ}) \sin\vartheta_P \sin\vartheta_Q \cos(\lambda_P - \lambda_Q) + \\ - \overline{C}''(r_P, r_Q, \cos\psi_{PQ})(\sin\vartheta_P \sin\vartheta_Q \sin(\lambda_P - \lambda_Q))^2,$$

which is not a function of ψ_{PQ} only, i.e. it is not a rotation invariant function.

Remark 2. In order to perform the covariance calculus of horizontal derivatives, a simple approach is, after fixing the two point P and Q, to compute the full covariance of the derivatives along the great circle connecting P and Q and orthogonal to it. The result can then be rotated to produce covariances of derivatives in any direction (Tscherning and Rapp 1974).

To get acquainted with the covariance spectral calculus we propose to the reader Exercise 2 at the end of the chapter.

5.7 The Estimate of Global Covariance Functions

The whole building of collocation theory rests on the assumption that there is a covariance function of the unknown $T(P), C_{TT}(P, Q)$, and that this function be known in some way. Since there is no theoretical a priori model for it we can only rely on data themselves to obtain an estimate of $C_{TT}(P, Q)$. Naturally the best theoretical framework to do that, would be a unified *estimation theory* where both $T(P)$ and $C_{TT}(P, Q)$ are optimally estimated together from data.

At this point indeed the problem becomes highly non-linear and, although some theoretical work has been done in this direction, no numerical experiments have been performed for the moment (Sansò and Venuti 2002a). So in practice we have to live with a two-steps procedure in which we first *estimate* $C_{TT}(P, Q)$, with an admissible model, and then we use it to apply the rest of collocation theory. This parallels very much what we are doing in the ordinary least squares theory (Koch, 1987) where we have to estimate both the vector of the parameters and the covariance matrix of the observable variables. In least squares theory however this practice is justified because we can prove that a variation of such covariance matrix induces a second order variation into the estimator of the parameters. Fortunately here we have again a similar situation as it has been proved in Sansò et al. (2000). So there is a reasonable argument to accept the two-step procedure. Yet the question is open on how to estimate practically $C_{TT}(P, Q)$ from data (see also Part II, Chap. 7).

We have two formulas relating the covariance function to observable quantities: one is its definition (5.92) that writes more explicitly as

$$P, Q \in S_Q,\ C_{TT}(\psi) = \frac{1}{4\pi} \int d\sigma_P T(P) \int_{\psi_{PQ}=\psi} d\alpha_Q T(Q); \qquad (5.112)$$

the other one is

$$C_{TT}(\psi_{PQ}) = \sum_{n=2}^{+\infty} c_n P_n(\cos \psi_{PQ}) \qquad (5.113)$$

with

$$c_n = \sum_{m=-n}^{n} T_{nm}^2. \qquad (5.114)$$

Both formulas require the knowledge of T on S_R (directly in (5.112) and through T_{nm} in (5.114)); both express $C_{TT}(P, Q)$ when $P, Q \in S_R$ and then can be harmonically continued in $P, Q \in \Omega_R$ by

$$C_{TT}(P, Q) = \sum_{n=1}^{+\infty} \left(\frac{R^2}{r_P r_Q} \right)^{n+1} c_n P_n(\cos \psi_{PQ}). \qquad (5.115)$$

Yet, since the quantity related to T that we know best at present, at the level of the ellipsoid, here approximated by S_R, is Δg, averaged in blocks, as explained in Chap. 3, the model (5.106) has been rather used, namely

$$C_{\Delta g \Delta g}(P, Q) = \sum_{n=2}^{+\infty} \left(\frac{R^2}{r_P r_Q} \right)^{n+1} c_n(\Delta g) P_n(\cos \psi_{PQ}). \tag{5.116}$$

where

$$c_n(\Delta g) = \frac{(n-1)^2}{R^2} c_n(T) = \frac{(n-1)^2}{R^2} \sum_{m=-n}^{n} T_{nm}^2. \tag{5.117}$$

Naturally with our finite data set we can only estimate $c_n(\Delta g)$ up to some maximum degree $N_{\max}$. It is by interpolating the empirical spectrum of Δg, i.e. (5.117), and then extrapolating it above $N_{\max}$ that we can have some model extending to all degrees up to infinity. The idea is similar to what we presented in Sect. 3.8, but with much more refined models which, beyond giving a better interpolation of empirical data, have also the advantage that the series (5.115) and (5.117) can be added providing us with closed analytical forms, more manageable from the numerical point of view. The argument and the relative models will be taken up in more details in the next section. What is interesting at this point is to underline two facts. The first is that all models include in both $c_n(T)$ and $c_n(\Delta g)$ an exponential factor which can therefore interpreted as $\left(\frac{R_B}{R}\right)^{2(n+1)}$, meaning that our kernel $C_{TT}(P, Q)$ will be harmonic down to a smaller sphere than S_R, in fact down to the Bjerhammar radius R_B, which in the most famous of such models (cf. Tscherning and Rapp 1974), has a value $R_B \cong 6{,}370\,\text{km}$. Note that R_B is different from the mean earth radius $R \cong 6{,}371\,\text{km}$, by 1 km only. The second is that, despite its usefulness, the degree variances of this global covariance function above $N_{\max}$ cannot well represent the local physical reality of our gravity field. In fact at the scales of 100 km down to 1 km the actual gravity field displays features so diverse from one part of the globe to the other that putting them all together into a unique covariance function prevents us from the construction of a very fine approximation of T, and then of the geoid, as required nowadays.

This argument calls for another step in our approximation road, where the local features of T or Δg are accounted for. We could say another step zooming into a smaller data area A and applying some kind of multi-resolution analysis concept. This will be achieved by means of the so-called local covariance functions.

We conclude the section with still another Example that will become useful in the sequel. This answers in the affirmative to the question: is it possible to have isotropic covariances on the bounding sphere that have a finite support, i.e. a $C(\psi)$ and a fixed arc $\Delta < \pi$ such that $C(\psi) = 0$ for $\forall \psi \geq \Delta$? In the example, we will construct one of such covariances, $M_\Delta(\psi)$, so that, recalling that the product of two covariance functions is again a covariance function, we can then construct for every $C(\psi)$ a finite support counterpart just by taking $C_\Delta(\psi) = M_\Delta(\psi) \cdot C(\psi)$.

Example 5. Let us recall that if we take at the north pole a function equal to 1 just when the colatitude ϑ is such that $\vartheta \leq \Delta$, and is equal to zero outside,

$$\chi_\Delta(\vartheta) = \begin{cases} 1 & \vartheta \leq \Delta \\ 0 & \vartheta > \Delta, \end{cases}$$

one can write

$$\chi_\Delta(\vartheta) = \sum_{n=0}^{+\infty} \beta_n P_n(\cos\vartheta)$$

where the so-called *Meissel's coefficients* β_n are given explicitly by (see also Sect. 3, A.4)

$$(t = \cos\vartheta) \qquad \beta_n = \frac{2n+1}{2} \int_{\cos\Delta}^{1} P_n(t)dt$$
$$= \frac{1}{2}[P_{n-1}(\cos\Delta) - P_{n+1}(\cos\Delta)].$$

Note that the relation between β_n and the coefficients of the moving average operator, defined in Sect. 3, A.4 is

$$\beta_n = \frac{1}{4}(2n+1)(1-\cos\Delta)M_n(\Delta).$$

Recalling that $Y_{n0} = \sqrt{2n+1}\,P_n(\cos\vartheta)$ we can write also

$$\chi_\Delta(\vartheta) = \sum_{n=0}^{+\infty} \frac{\beta_n}{\sqrt{2n+1}} Y_{n,0}(\vartheta)$$

If we consider this function as a potential on the sphere and we compute its covariance in spectral from (cf. (5.113) and (5.114)) we find

$$M_\Delta(\psi) = \sum_{n=0}^{+\infty} \frac{\beta_n^2}{2n+1} P_n(\cos\psi)$$

On the other hand if we compute the same covariance by (5.112) we see that we must fix P in the cap $D(O,\Delta)$ of radius Δ around the north pole O, we must fix a radius ψ and then take the product of $\chi_\Delta(\vartheta_P)$ by the average of $\chi_\Delta(\vartheta_Q)$ on the circle of radius ψ around P; finally we integrate in P over $D(P,\Delta)$. Note that when P is outside $D(O,\Delta)$, the integrand in (5.112) is automatically zero.

Now if P is in $D(O,\Delta)$ and on the same time $\psi > 2\Delta$, the circle of radius ψ and centre P, will not intercept anymore $D(O,\Delta)$ and, as result, we will have

$$M_\Delta(\psi) = 0, \ \forall \psi > 2\Delta.$$

The situation is illustrated in Fig. 5.7.

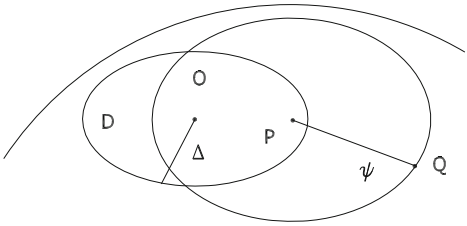


Fig. 5.7 Domains of integration used in the computation of the covariance of $T = \chi_\Delta(\vartheta_P)$

Let us observe explicitly that although we can construct covariances of finite support on the spherical boundary, as soon as we go to an external sphere, $r > R$, $C(P, Q)$ cannot be anymore zero on any part of the sphere of positive measure, otherwise as a harmonic function it should be zero everywhere (see Sacerdote and Sansò 1991).

5.8 The Estimate of Local Covariance Functions

As defined in (5.38), with the further specification of definition (5.76) we can say that the covariance function of the gravity anomaly field $\Delta g(P)$, at the level of the mean earth sphere, S_{R_e}, is given by

$$\begin{aligned} &P, Q \in S_{R_e}, \\ &C_{\Delta g \Delta g}(P, Q) = E\{\Delta g(P)\Delta g(Q)\} \\ &\quad = \frac{1}{8\pi^2} \int d\sigma_{P'} \int_{\psi_{P'Q'}=\psi_{PQ}} d\alpha_{Q'} \Delta g(P')\Delta g(Q') = C_{\Delta g \Delta g}(\psi_{PQ}); \end{aligned} \tag{5.118}$$

analogous formulas hold for $C_{TT}(P, Q)$ and $C_{T\Delta g}(P, Q)$ which are the main ingredients needed to derive the estimates (5.80) and (5.81).

The relation between the three functions is given by (5.76) and (5.79) in the ordinary geometric space and by (5.105) and (5.106) in the spectral domain. Although we derived them for the residual potential, represented by the random field $v(P, \omega) = T_r(R_\omega P)$, they basically hold for any random field similarly defined by means of its values on the sphere S_{R_e}, with the help of a uniform distribution on the rotation group, and harmonically continued in $\Omega_{R_e} \equiv \{r \geq R_e\}$. So in order to be close to the applications considered in this book we shall reason in this section on the covariance of Δg_r, with the understanding that the same arguments apply to any random field having an isotropic covariance function.

Moreover, such a remark will be used in next sections.

From (5.118) and a set of observed values

$$Y_i = \Delta g(P_i) + v_i, \quad i = 1, 2 \ldots N \tag{5.119}$$

with v_i independent noises of equal variance σ_v^2, we can reasonably build an estimator of the covariance in a very similar way of what is done with random processes, with respect to a time variable.

In fact, consider the following expression

$$\widehat{C}_{\Delta g\Delta g}(\overline{\psi}) = \frac{1}{N(\overline{\psi}, \Delta)} \Sigma_{\{i,k\}} Y_i Y_k, \tag{5.120}$$

where the summation is extended only to the pair of points $\{i, k\}$ such that

$$\overline{\psi} - \Delta < \psi_{P_i P_k} \leq \overline{\psi} + \Delta \tag{5.121}$$

and $N(\overline{\psi}, \Delta)$ is the number of such pairs.

Observe that, recalling also (5.55), (5.56) and (5.57),

$$\begin{aligned} E_{\omega,v}\{Y_i Y_k\} &= E_\omega\{\Delta g(P_i)\Delta g(P_k)\} + \sigma_v^2 \delta_{ik} \\ &= C_{\Delta g\Delta g}(\psi_{P_i P_k}) + \sigma_v^2 \delta_{ik}. \end{aligned} \tag{5.122}$$

As far as $\overline{\psi} - \Delta \geq 0$, i.e. $\psi_{P_i P_k} > 0$, we always have $\delta_{ik} = 0$ in (5.122), so that from (5.120) we find again, denoting $\{i, k\}$ the set of pairs satisfying (5.121),

$$E_{\omega,v}\{\widehat{C}_{\Delta g\Delta g}(\overline{\psi})\} = \frac{1}{N(\overline{\psi}, \Delta)} \Sigma_{\{i,k\}} C_{\Delta g\Delta g}(\psi_{P_i P_k}). \tag{5.123}$$

Now, if we assume that the observation points $\{P_i\}$ are well distributed, so that $\psi_{P_i P_k}$ sweeps in a fairly homogeneous way the interval $[\overline{\psi} - \Delta, \overline{\psi} + \Delta]$ and if we further agree that Δ is such that $N(\overline{\psi}, \Delta)$ is large enough e.g. at least larger than 10, and on the same time small enough, to allow $C_{\Delta g\Delta g}(\psi)$ to be almost linear in the interval $[\overline{\psi} - \Delta, \overline{\psi} + \Delta]$, we deduce from (5.123)

$$E_{\omega,v}\{\widehat{C}_{\Delta g\Delta g}(\overline{\psi})\} \approx C_{\Delta g\Delta g}(\overline{\psi}), \tag{5.124}$$

namely $\widehat{C}_{\Delta g\Delta g}(\overline{\psi})$ is a quasi-unbiased estimator of $C_{\Delta g\Delta g}(\overline{\psi})$.

Furthermore we note that (5.120) can be considered as well as a discretization of formula (5.112) or its analogous for Δg.

Accordingly, once the value of Δ has been fixed, what is in fact one of the very issues for the data analyzer, we can derive estimates $\widehat{C}_{\Delta g\Delta g}(\overline{\psi})$ for

$$\overline{\psi} = \Delta, 3\Delta, 5\Delta \ldots (\ell m + 1)\Delta. \tag{5.125}$$

Furthermore, by taking $i = k$ in (5.122), we derive

$$E\left\{\frac{1}{N}\sum_{i=1}^{N} Y_i^2\right\} = C_{\Delta g\Delta g}(0) + \sigma_v^2, \tag{5.126}$$

i.e.

$$S_y^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 \tag{5.127}$$

is an unbiased estimator of $C_{\Delta g \Delta g}(0) + \sigma_\nu^2$. All together the values

$$S_y^2;\ \widehat{C}_{\Delta g \Delta g}(\Delta);\ \widehat{C}_{\Delta g \Delta g}(3\Delta) \ldots \widehat{C}_{\Delta g \Delta g}((2m+1)\Delta) \tag{5.128}$$

constitute what is called the *empirical covariance function*; when Δg is the residual gravity anomaly Δg_r and the points $\{P_i\}$ are taken from a *local* area A only, we have a local empirical covariance function.

Note that, in order that such empirical covariance function could be further used in the prediction process, some conditions have to be fulfilled at least approximately. We already said about the choice of Δ, but we also have to assume that when data come from a local area A, $(2m+1)\Delta$ (see (5.118)) be significantly smaller than the size of A, identified with its diameter when A is a cap or with its side if A is a squared geographic block; at the same time $C_{\Delta g \Delta g}((2m+1)\Delta)$ and the other tail values of $C_{\Delta g \Delta g}$ beyond $(2m+1)\Delta$, should be small enough to make the correlation with observations beyond this distance negligible; moreover the size of A should be big enough to let the field Δg to have a zero average on it, i.e.

$$\frac{1}{N}\sum_{i=1}^{N} \Delta g(P_i) \approx 0, \tag{5.129}$$

as otherwise we could not write a covariance estimator in the form (5.120).

In reality, having an empirical average significantly different from zero on A would mean just that there is an important correlation of Δg_r in A with Δg_r outside A, so that we cannot hope to derive a good local estimate of T in A because we are lacking essential information.

One further concern is that the height of the points P_i should not have too strong a variation in A; in fact we see (cf. (5.116)) that if all points have the same height h, then the degree variances of Δg are just modified by a factor $\left(\frac{R_e}{R_e+h}\right)^{2n+4}$, that can be accounted for in modelling the covariance, while if $r_i = R_e + h_i$ is quite variable, then the covariance of the signal coming from h_i will enter into the empirical values $\widehat{C}_{\Delta g \Delta g}(\overline{\psi})$.

Finally we remind that our estimate (5.120) is relevant only if the residual $\Delta g_r(P)$ has a behaviour statistically homogeneous and isotropic in A; in other words there should not be in $\Delta g_r(P)$ features that make one part of A to look statistically very different from another one. This is typically achieved if the remove step for the model and for the residual terrain correction components has been correctly performed and the area A is suitably selected by the analyzer.

We get hold of an empirical covariance function that we need to transform into a model covariance function, namely into a function possessing the correct properties of symmetry and positive definiteness, without which the collocation prediction formulae loose any significance. This is the case if we impose to the model covariance to satisfy the relation (5.116), namely

$$C_{\Delta g\Delta g}(P,Q)=\sum_{n=2}^{+\infty}c_n(\Delta g)\left(\frac{R^2}{r_P r_Q}\right)^{n+2}P_n(\cos\psi_{PQ}),\tag{5.130}$$

with positive full power degree variances $c_n(\Delta g)$.

Now the point is how to model $c_n(\Delta g)$, taking also into account that we are talking about Δg_r, so that we expect $c_n(\Delta g)$ to have a different meaning when $n \le M$ (M being the maximum degree of our global model $T_M(P)$) than when $n > M$.

In fact if we write for the coefficients $T_{nm}^{(M)}$ of the global model the relation

$$T_{nm}^{(M)}=T_{nm}+\tau_{nm}\tag{5.131}$$

with τ_{nm} the estimation error for the coefficient T_{nm}, we see that in the low frequency band (cf. (5.101)),

$$(n\le M),\quad \Delta g_{r,nm}=\frac{n-1}{R}\tau_{nm}\tag{5.132}$$

so that

$$(n\le M),\quad c_n(\Delta g)=\frac{(n-1)^2}{R^2}\sum_{m=-n}^{n}\tau_{nm}^2,\tag{5.133}$$

according to (5.104).

Now (5.133) expresses the full power degree variances of the estimation errors $\{\tau_{nm}\}$, when the average is taken over the full rotation group. If we further average (5.133) with respect to the random variables τ_{nm}, which represent the propagation of the observation (and model) errors from original data to the estimates $T_{nm}^{(M)}$, we can define what are called *error degree variances*, namely

$$(n\le M),\quad \varepsilon_n(\Delta g_r)=E_\tau\{c_n(\Delta g_r)\}=\sum_{m=-n}^{n}\sigma^2(\tau_{nm}).\tag{5.134}$$

The variances $\sigma^2(\tau_{nm})$ are available from least squares estimates up to degrees of a few hundreds, or are derived by noise propagation through quadrature formulas (see Rapp 1997a; Pavlis et al. 2008), so we can claim that ε_n are known at least up to the specific degree M, which is useful in the present context (see Remark 3 below).

As for higher degrees, $n > M$, the full power degree variances are usually modelled by means of some parametric form. Typical are formulas of the type

$$c_n(\Delta g)=C_0h^{n+2}\frac{A(n)}{B(n)}\tag{5.135}$$

where

$$0<h<1\tag{5.136}$$

and $A(n)$, $B(n)$ are polynomials in n such that $B(n)$ has no zeroes for integer values larger than 1. The big advantage of the form (5.135) is that in many cases it becomes possible to add the series (5.130) obtaining an explicit analytic expression which is then quite comfortable to be used in further computations (see Sect. 5.9).

Remark 3. Let us put

$$h = \frac{R_B^2}{R^2},\ (R_B < R) \tag{5.137}$$

in (5.135) and substitute it back into (5.130); we find then

$$C_{\Delta g \Delta g}(P, Q) = \sum_{n=2}^{+\infty} \frac{A(n)}{B(n)} \left(\frac{R_B^2}{r_P r_Q} \right)^{n+2} P_n(\cos \psi_{PQ}). \tag{5.138}$$

Since $|P_n(\cos \psi)| \leq 1$, it is clear that (5.138) is converging in $r_P, r_Q > R_B$, whatever be the polynomials A and B; therefore any collocation solution that uses this covariance will be harmonic down to a sphere with radius R_B. As already mentioned at the end of Sect. 5.6, the constant R_B is called a *Bjerhammar radius* after the work of A. Bjerhammar (see for instance Bjerhammar 1987); whence the index B.

Summarizing the previous general discussion, we arrive at a model of local covariance function that can be expressed as

$$C^{\text{Mod}}_{\Delta g \Delta g}(P, Q) = a \sum_{n=2}^{M} \varepsilon_n \frac{(n-1)^2}{R^2} \left(\frac{R^2}{r_P r_Q} \right)^{n+2} P_n(\cos \psi_{PQ}) + C_r(P, Q) \tag{5.139}$$

$$C_r(P, Q) = \sum_{n=M+1}^{+\infty} c_n(\Delta g) \left(\frac{R^2}{r_P r_Q} \right)^{n+2} P_n(\cos \psi_{PQ}) \tag{5.140}$$

$$c_n(\Delta g) = C_0 h^{n+2} \frac{A(n)}{B(n)}. \tag{5.141}$$

Parameters of the representation (5.139), (5.140) and (5.141) are: the calibration constant a, the degree M used in the specific remove-restore procedure, the constant C_0, the Bjerhammar radius R_B, i.e. the value of h, the coefficients of the polynomials $A(n)$, $B(n)$ which however can be normalized to have the zero degree coefficients equal to 1, namely $a_0 = b_0 = 1$.

By using all these parameters one can interpolate the empirical covariance function, using only the values outside the origin $\widehat{C}_{\Delta g \Delta g}(\Delta), \ldots, \widehat{C}_{\Delta g \Delta g}((2m+1)\Delta)$.

In this covariance modelling process it is important to use M as a parameter because the experience shows that many times the use of R_B only does not allow to reach the right shape of the covariance in the first (and most important) part of $C_{\Delta g \Delta g}(\psi)$, typically decreasing from the value $C_{\Delta g \Delta g}(0)$.

The value S_y^2 (cf. (5.127)) is then used to estimate σ_v^2,

$$\widehat{\sigma}_v^2 = S_y^2 - \widehat{C}_{\Delta g \Delta g}(0). \tag{5.142}$$

As it is obvious one must have

$$\widehat{\sigma}_v^2 \geq 0 \tag{5.143}$$

for this estimate to be acceptable; therefore (5.143) acts as a constraint for the model

$$C_{\Delta g \Delta g}^{\text{Mod}}(0) \leq S_y^2. \tag{5.144}$$

All in all, this estimation procedure casts so to say into a theoretically acceptable form the statistical behaviour of Δg_r in the specific area A, captured by the empirical estimates (5.120). Therefore, despite its global appearance, $C_{\Delta g \Delta g}^{\text{Mod}}$ represents in fact the physical correlation of Δg_r in the area A and in general it should not be used for another area. This reflects, to some extent, the multi-resolution character of the solution we are elaborating, step after step.

Example 6. It is important to understand that the transition from Δg to Δg_r removes power from $C_{\Delta g \Delta g}$, namely it damps its value at the origin and at the same time it reduces the correlation length, i.e. the smallest value ψ_c for which the relation

$$C_{\Delta g \Delta g}(\psi) = \frac{1}{2} C_{\Delta g \Delta g}(0) \tag{5.145}$$

is satisfied. More properly one could say that the transition from Δg to Δg_r reduces the index $\dfrac{C_{\Delta g \Delta g}(0)}{\psi_c}$, that could be taken as an indicator of the smoothness of the covariance. In this respect, it is interesting to observe the sequence of the covariance functions for the full signal of free air Δg over the area $6^\circ \leq \lambda \leq 20$, $36^\circ \leq \varphi \leq 47^\circ$ corresponding to a domain A covering the Italian region (Fig. 5.8), and the covariance function of the reduced Δg_r over the same region (Fig. 5.9). Finally in Fig. 5.9 we show as well the covariance from the Tschering–Rapp family (see formula (7.16) in Part II, Chap. 7) that interpolates $\widehat{C}_{\Delta g_r \Delta g_r}$.

Notice that in the chosen land area the gravity signal is quite variable, due to the complex geological structure of the region. So the covariance of the global gravity field, reflecting a mean behaviour for the whole earth, suggests a behaviour smoother than that implied by the local covariance in Fig. 5.8. On the other hand the covariance of Δg_r is both less powerful and smoother than that of the free air anomalies.

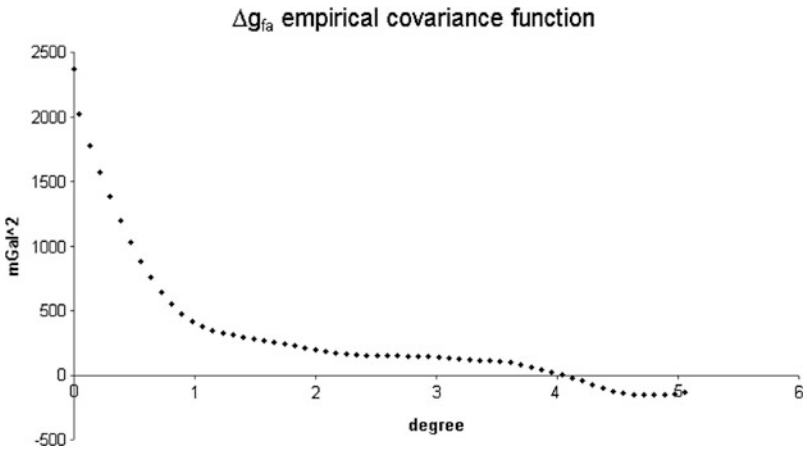


Fig. 5.8 The free air gravity anomaly empirical covariance over the Italian area

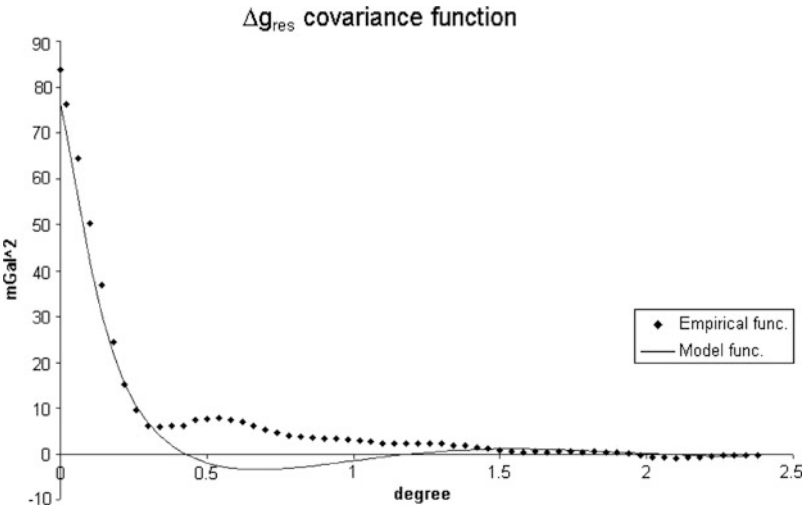


Fig. 5.9 The empirical covariance of the reduced gravity anomaly over the Italian area and the best fitting Tscherning–Rapp model

5.9 Covariance Parametric Models

As we have seen in the two previous sections, an estimation procedure for the covariance function of T or Δg passes through the adaptation of a parametric model to suitable empirical covariance values.

For this purpose let us note that if we accept the model (5.135) and we put

$$s = \frac{R_B^2}{r_P r_Q},\ t = \cos\psi \tag{5.146}$$

our target is to sum a series of the form

$$C_{\Delta g \Delta g}(s,t) = \sum_{n=2}^{+\infty} \rho_{\Delta g}(n) s^{n+2} P_n(t) \tag{5.147}$$

with $\rho(n)$ a rational function of n.

Since it is convenient in the present context, we shall however start from the covariance of T, that in this case, with the notation (5.146), can be written

$$C_{TT}(s,t) = \sum_{n=2}^{+\infty} \rho_T(n) s^{n+1} P_n(t). \tag{5.148}$$

In performing our calculus we shall need a few relations that we list for the comfort of the reader. We start by recalling (see (3.16) and (3.17)) the definition of generating function

$$G(s,t) = \sum_{n=0}^{+\infty} s^n P_n(t) = \frac{1}{\sqrt{1+s^2-2st}} \tag{5.149}$$

and the obvious relation

$$\sum_{n=2}^{+\infty} s^n P_n(t) = G(s,t) - 1 - st. \tag{5.150}$$

Then we have

$$\begin{cases} \frac{\partial}{\partial s} G^{-1}(s,t) = (s-t) G(s,t) \\ \frac{\partial}{\partial s} G(s,t) = -(s-t) G^3(s,t). \end{cases} \tag{5.151}$$

Furthermore, as one can verify by direct differentiation, one has

$$\int_0^s G(\sigma,t)\, d\sigma = \log \frac{s - t + G^{-1}(s,t)}{1-t}; \tag{5.152}$$

note that when $s \to 0$ both members tend to zero.

Moreover we observe that, for any $F(s,t)$,

$$-\frac{\partial}{\partial r_P} F(s,t) = \frac{s}{r_P} \frac{\partial}{\partial s} F(s,t) \tag{5.153}$$

and similarly for $-\frac{\partial}{\partial r_Q} F(s,t)$.

With such tools a number of intermediate results are derived in the exercises at the end of the chapter, that the reader is invited to make.

We continue the section by concentrating on one of the covariance models that are most widely used in modelling gravity covariances. Before doing so we underline again that such a model can be used for both, global and local covariance modelling. In fact any global model of which we know the sum in analytical form, namely

$$C(s,t) = \sum_{n=0}^{+\infty} c_n s^{n+2} P_n(t) \tag{5.154}$$

can be turned into a truncated form of the type

$$\begin{aligned} C_M(s,t) &= \sum_{n=M+1}^{+\infty} c_n s^{n+2} P_n(t) \\ &= C(s,t) - \sum_{n=0}^{M} c_n s^{n+2} P_n(t), \end{aligned} \tag{5.155}$$

which is easily computed because $C(s,t)$ has a closed form and the second term in (5.155) is just a finite sum up to a few hundred terms.

The Tscherning–Rapp model. This model (see Tscherning and Rapp 1974) has, in its classical formulation, the general form (5.130) and (5.135), parameterizing the gravity full power degree variances as

$$c_n(\Delta g) = A \left(\frac{R_B^2}{R^2} \right)^{n+2} \cdot \frac{n-1}{(n-2)(n+B)}, \quad n \geq 3, \tag{5.156}$$

or, what amounts to the same, the form (5.138) with

$$\frac{A(n)}{B(n)} = \frac{A(n-1)}{(n-2)(n+B)}, \quad n \geq 3. \tag{5.157}$$

For reasons that are explained in Appendix A.2, the parameter B is restricted to integer values.

The computation of $C_{\Delta g \Delta g}(s,t)$ corresponding to the choices (5.157) is fully worked out in Appendix A.2. The result can be cast into the form

$$C_{\Delta g \Delta g}(s,t) = A \left\{ \frac{B+1}{B+2} K_B(s,t) + \frac{1}{B+2} K_{-2}(s,t) \right\} \tag{5.158}$$

and the algorithms to compute $K_B(s,t)$ and $K_{-2}(s,t)$ have to be found in Appendix A.2.

With similar arguments one can compute as well the covariance function of T and the cross-covariance of T and Δg which are essential to perform the prediction of T from Δg and compute the corresponding prediction error.

We have

$$C_{TT}(s,t) = AR^2 \left\{ \frac{1}{(B+2)} \frac{1}{s} K_{-2}(s,t) + \right. \tag{5.159}$$
$$+ \frac{1}{(B+1)(B+2)} \frac{1}{s} K_B(s,t) - \frac{1}{B+1} [s - s^2 t - sG^{-1}(s,t) +$$
$$\left. - s^2 + \log \frac{1 - st + G^{-1}(s,t)}{2} - s^3 P_2(t) \right] \Bigg\}$$

and

$$C_{T\Delta g}(s,t) = A \frac{R^2}{r_P (B+2)} \left\{ \frac{1}{s} K_{-2}(s,t) - \frac{1}{s} K_B(s,t) \right\} . \tag{5.160}$$

Note that in (5.160) Δg is evaluated at P while T is evaluated at Q and we have here $s = \frac{R_B}{r_P r_Q}$, $t = \cos \psi_{PQ}$.

5.10 The Least Squares Collocation (l.s.c.) Solution

By *solution* we mean here computing the predictor (5.68) with its prediction error variance (5.69), when the problem at hand is fully general. When we have to predict T from Δg, we have to utilize formulas (5.80) and (5.81). When we apply the latter formulas to a local data set, $\{P_i\} \in A$, of residual gravity anomalies, $\Delta g_r^{\text{obs}}(P_i)$, then we can predict *local* values of the residual anomalous potential $\widehat{T}_r(P)$.

A l.s.c. solution is exactly one such solution when a local covariance function is used in formula (5.80) and (5.81).

We notice here that there seems to be a certain degree of contradiction in applying the W-K principle of Sect. 5.4 to the present local context. In fact, by definition the covariance function of Sect. 5.4 is obtained by averaging on the full sphere, or better on the full rotation group; on the contrary the local covariance function used in a l.s.c. solution is derived only for the area A where we have data and it would be different for the true earth in another area.

Since the formula for the isotropic covariance function, (5.38), was in fact obtained from the minimum quadratic invariant error principle (5.37), it seems interesting to ask whether there is an analogous minimum quadratic error principle, valid for the data in the area A only, leading us to the use of a *local* covariance function. A rigorous answer to this question would be in the negative sense. However it is feasible to build a local theory implying a definition of a local

covariance function that is only approximately isotropic and is close to what is suggested by the estimation formula (5.118).

Yet this goes beyond the scope of this presentation and here we limit ourselves to some more elementary considerations.

Basically our solution would be justified at least in a mean square sense, if the field T_r we want to estimate had, outside the area A and over all the rest of the sphere, the same statistical behaviour. If we impose such a hypothesis by definition, we will have a prediction which is optimal for this *virtual* field and on the same time it agrees with ours, at least in terms of observations, in the area A.

So the question is not whether the local covariance is good for the whole sphere (which is not) but rather what is the region in space where our *local* approximation procedure gives valid answers.

Fortunately collocation theory helps by giving us the tool to compute the prediction error (see (5.72) and (5.81)) and we can decide to go with the prediction point as far as possible till the prediction error reaches a predefined threshold. In this sense it is useful to observe that sometimes it is convenient to fix a threshold for the relative prediction error, namely, if $T(P)$ is the predicted functional,

$$\mathcal{E}_r(P) = \left\{ \frac{\mathcal{E}(P)^2}{C_{TT}(P,P)} \right\}^{(1/2)} \tag{5.161}$$

$$= \left\{ 1 - \frac{\sum_{i,k=1}^{N} C_{T,T}(P,M_i) C^{(-1)}_{Y_i Y_k} C_{TT}(P,M_k)}{C_{TT}(P,P)} \right\}^{(1/2)} .$$

This expresses the ratio of the prediction error to the signal we want to predict and can be fixed to levels like 1%, and 5% or others. For instance, one can decide to estimate a residual geoid of 1 m, r.m.s., with an error of 1 cm.

A warning has to be done at this point: when formulas like (5.81) or (5.161) are used in an extrapolation mode, i.e. for points P outside the area A, they give us always optimistic values because outside A the actual residual gravity field might not be well-represented, as for its statistical behaviour, by the same local covariance that has been estimated form values in A only. As a matter of fact this is of no great concern because numerical experience shows that already inside A, close to its boundary, $\mathcal{E}^2(P)$ and $\mathcal{E}_r^2(P)$ increase to unacceptable values and the prediction has to stop.

Remark 4. The above phenomenon can be understood qualitatively on the basis of the following reasoning. Remember that the local covariance function is estimated from empirical values and we have agreed that those have to become small at angular distance $\psi > \Delta$ for some Δ much smaller than the size of A. Accordingly, exploiting the possibility illustrated in the Example 5, we can model the theoretical

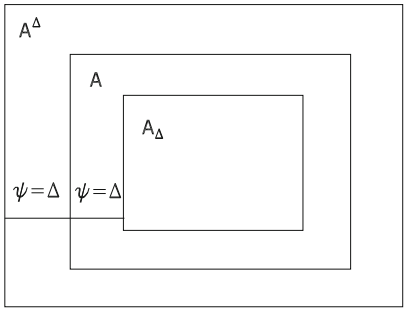


Fig. 5.10 A the set covered by data; A^{Δ} the set where data give some information; A_{Δ} the set where a good prediction can be performed

local covariance to have a finite support, i.e. to go strictly to zero on the sphere, when $\psi > \Delta$.

So, assume one has to perform a prediction at P, on the sphere, from observed values $T(P_i)$.

We see that outside the set $A^{\Delta} \equiv \{P \; ; \; \psi_{PQ} \leq \Delta \text{ for some } Q \in A\}$ (see Fig. 5.10) the l.s.c. predictor of $\widehat{T}(P)$ is $\widehat{T}(P) = 0$. In fact if the observation points P_i are all in A and P is outside A^{Δ}, $\psi_{PP_i} > 0, \forall i$ and then $\widehat{T}(P)$, written in the form

$$\widehat{T}(P) = \sum_{i=1}^{N} \xi_i C(\psi_{PP_i}) \tag{5.162}$$

is indeed zero. On the contrary, if we are well inside A, depending on the density of data and on the signal to noise ratio, we can have a good prediction of T. Let's reason now on a belt for instance of width Δ in A, i.e. in $A \backslash A_{\Delta}$, with $A_{\Delta} \equiv \{P \in A \; ; \; \psi_{PQ} < \Delta \Rightarrow Q \in A\}$. We expect that important information for the prediction of $T(P)$ is lost when $P \in A \backslash A_{\Delta}$ and correspondingly the prediction error becomes higher (see Fig. 5.10).

The above reasoning, though not rigorous, gives an idea of what happens in reality. A few exercises at the end of the section will be useful to the reader to enter into the subject.

Now that we have roughly agreed how to settle the prediction domain in a horizontal direction, we have to address the problem of the vertical dimension of this domain. The following trivial example can help in grasping the problem.

Example 7. Assume that $T(P)$ has covariance function

$$C(P, Q) = \sum_{n=m}^{+\infty} c_n \left(\frac{R^2}{r_P r_Q} \right)^{n+1} P_n(\cos \psi_{PQ}) \; ;$$

assume that at Q, with $r_Q = R$, we have observed $T(Q)$ without noise and we want to predict $T(P)$ along the radius passing through Q. By applying (5.70) with

evaluation functionals and one observation only we get (note that $\psi_{PQ} = 0$ and $P_n(1) = 1$)

$$\begin{aligned}\widehat{T}(P) &= C(P,Q)C^{-1}(Q,Q)T(Q)\\ &= \frac{\sum_{n=m}^{+\infty} c_n \left(\frac{R}{r_P}\right)^{n+1}}{\sum_{n=m}^{+\infty} c_n} T(Q).\end{aligned} \tag{5.163}$$

and the corresponding relative error from (5.161) is

$$\mathcal{E}_r^2(P) = 1 - \frac{\left[\sum_{n=m}^{+\infty} c_n \left(\frac{R}{r_P}\right)^{n+1}\right]^2}{\sum_{n=m}^{+\infty} c_n \cdot \sum_{n=m}^{+\infty} c_n \left(\frac{R}{r_P}\right)^{2n+2}} \tag{5.164}$$

If we take the limit for $r_P \to \infty$ of (5.164), we receive

$$\lim_{r_P \to \infty} \mathcal{E}_r^2(P) = 1 - \frac{c_m}{\sum_{n=m}^{+\infty} c_n}. \tag{5.165}$$

Then we expect $\mathcal{E}_r(P)$ to be close to 1 when r_P increases, i.e. P moves to the zenith of Q. For instance take for c_n the simple model

$$c_n = h^n$$

with h close to 1, then we see from (5.165) that

$$\mathcal{E}_r^2(P) \to h$$

i.e. the relative error becomes almost 100%. So if we fix a threshold for $\mathcal{E}_r$ then we will find an upper limit for the height where our solution is acceptable.

The phenomenon, highlighted in the Example 7, has general character and is basically related to the fact that if Q_i are observation points with $r_{Q_i} = R$ and P is taken on a higher sphere, $r_P > R$, then $C(P, Q_i)$ is modified by multiplying c_n by the factor $\left(\frac{R}{r_P}\right)^{n+1}$; this corresponds to giving more weight to the low frequencies and to damp the high frequencies so that the shape of the covariances is flattened. In turn this implies that we need more measurements distant from the prediction area, to perform a good prediction job. Accordingly we understand that data on a larger

area are needed to make a prediction with fixed relative error. Or, equivalently, when we rise in height the area of valid predictions has to be reduced.

Remark 5. Another way to approach the "localization" of the approximation to T is to push even further our simplification of reference model to arrive to the so-called *planar approximation*, where the reference gravity vector is always pointing to a parallel direction. Also in this case the collocation concept can be applied with the advantage of having available the Fourier transform machine (see Chap. 10 of this Part II). An interesting connection can then be established between planar and spherical covariance functions (see Forsberg 1987).

5.11 On the Optimal Combination of Global Coefficients and Local Observations

The procedure of removing from the anomalous potential, and all the corresponding observables, a global model T_M and then applying to the residual part T_r the collocation prediction, based on data in a local area A only, as explained in Sects. 5.8 and 5.10, is not strictly rigorous. As a matter of fact one should apply the W.K. principle to a full combination of the available information, namely the local data *and* the global model coefficients. Beyond the rigor, one of the advantages of proceeding along this line is that we can overcome the request that $T_r(P)$ be of zero average on A; such a request in fact is sometimes restrictive, specially if we have to predict the potential with high accuracy in a small area.

So we assume we have performed only a smoothing for the high frequency residual terrain correction and we call again $T(P)$ the remaining unknown potential. Then we consider as given information a set of *local* observations

$$\begin{aligned} Y_i &= M_i(T) + \nu_i, \\ 1 &= 1, 2 \ldots J \end{aligned} \tag{5.166}$$

with T a random field with a global covariance

$$C_{TT}(P, Q) = C(P, Q) = \sum_{n=2}^{+\infty} c_n(T) \left(\frac{R^2}{r_P r_Q} \right)^{n+1} P_n(\cos \psi_{PQ}) \tag{5.167}$$

which for the moment we consider as known. As usual ν_i are observation noises with zero mean and a known covariance matrix C_ν, moreover ν_i are independent of T. In vector from we write (5.166) as

$$\mathbf{Y} = \mathbf{M}(T) + \boldsymbol{\nu} \tag{5.168}$$

with first moments specified as usual by

$$E\{\mathbf{Y}\} = 0,\ C_{YY} = C(\mathbf{M}, \mathbf{M}^t) + C_{vv}. \tag{5.169}$$

In addition we shall assume to know the harmonic coefficients of T to some degree N, namely

$$\begin{aligned} T_{nm}^M &= T_{nm} + \tau_{nm} \\ -n \le m &\le n\ ;\ n = 2, \ldots, N. \end{aligned} \tag{5.170}$$

In (5.170) T_{nm} are the *true* harmonic coefficients of T, that we write as linear functionals

$$T_{nm} = \frac{1}{4\pi} \int T(R, \vartheta, \lambda) Y_{nm}(\vartheta, \lambda) d\sigma = H_{nm}(T) \tag{5.171}$$

and τ_{nm} are the *errors* of the known coefficients on the nature of which we shall comment later on. We find it convenient to vectorize (5.170) as $N - 1$ vector equations, namely

$$\mathbf{T}_n^M = \mathbf{T}_n + \boldsymbol{\tau}_n = \mathbf{H}_n(T) + \boldsymbol{\tau}_n. \tag{5.172}$$

The error vectors $\boldsymbol{\tau}_n$ are assumed to be of zero average and to have covariance matrices

$$G_n = E\{\boldsymbol{\tau}_n \boldsymbol{\tau}_n^t\}\ ; \tag{5.173}$$

moreover, though not essential, we shall assume that

$$E\{\boldsymbol{\tau}_n \boldsymbol{\tau}_\ell^t\} = \delta_{n\ell} G_n, \tag{5.174}$$

i.e. $\boldsymbol{\tau}_n$ and $\boldsymbol{\tau}_\ell$ referring to different degrees are uncorrelated.

Furthermore we assume that all $\boldsymbol{\tau}_n$ are not correlated with the random field $T(P)$, $E\{T(P)\boldsymbol{\tau}_n\} = 0$.

In addition, although it is possible that the same observations $\mathbf{Y}$ have been used too in the estimate of $\mathbf{T}_n^M$, since in this case they are mixed with a much larger data set coming from everywhere on the earth, outside A, we shall assume that the correlation of $\boldsymbol{\tau}_n$ and $\mathbf{Y}$ is zero, namely

$$E\{\mathbf{Y}\boldsymbol{\tau}_n^t\} = 0. \tag{5.175}$$

In principle predicting by collocation any functional $L(T)$ of T is nothing new, however the specific form of the functionals $\mathbf{H}_n$ and their covariance and cross-covariances with $\mathbf{Y}$ are such as to provide the solution in a very suggestive form.

So deciding to limit ourselves to $L_P(T) = T(P)$ and so to search the predictors in the form

$$\widehat{T}(P) = \boldsymbol{\lambda}^t \mathbf{Y} + \sum_{n=2}^{N} \boldsymbol{\alpha}_n^t \mathbf{T}_n^M \tag{5.176}$$

we can construct directly the normal system for the unknowns $\boldsymbol{\lambda}$ and $\{\boldsymbol{\alpha}_n\}$. To do so it is convenient first to compute some cross-covariances. For the sake of convenience, to follow the vectorized notation (5.172) we can put

$$T(P) = \sum_{n=2}^{+\infty} \sum_{m=-n}^{n} T_{nm} S_{nm}(r\vartheta, \lambda) = \sum_{n=2}^{+\infty} \mathbf{T}_n^t \mathbf{S}_n(P), \tag{5.177}$$

implicitly defining $\mathbf{S}_n$.

Then we have, recalling that $\sigma_n^2 = \frac{c_n(T)}{2n+1}$,

$$E\{\left(\mathbf{T}_n^M\right)\left(\mathbf{T}_\ell^M\right)^t\} = C(\mathbf{H}_n, \mathbf{H}_\ell^t) + G_n \delta_{n\ell} \tag{5.178}$$

$$= E\{\mathbf{T}_n \mathbf{T}_\ell^t\} + G_n \delta_{n\ell} = (\sigma_n^2 I + G_n)\delta_{n\ell},$$

$$E\{\mathbf{Y}\left(\mathbf{T}_n^M\right)^t\} = C(\mathbf{M}, \mathbf{H}_n^t) = E\{\mathbf{M}(T)\mathbf{T}_n^t\} \tag{5.179}$$

$$= E\left\{\sum_{\ell=2}^{+\infty} \mathbf{M}(\mathbf{S}_\ell^t)\mathbf{T}_\ell \mathbf{T}_n^t\right\} = \sigma_n^2 \mathbf{M}(\mathbf{S}_n^t),$$

$$E\{\mathbf{Y}T(P)\} = C(\mathbf{M}, P), \tag{5.180}$$

$$E\{\mathbf{T}_n^M T(P)\} = C(\mathbf{H}_n, P) = \sigma_n^2 \mathbf{S}_n(P). \tag{5.181}$$

Since the normal equation system has general form

$$\begin{cases} C_{YY}\boldsymbol{\lambda} + \sum_{\ell=2}^{N} C_{YT_\ell^M}\boldsymbol{\alpha}_\ell = C_{YT} \\ C_{T_n^M Y}\boldsymbol{\lambda} + \sum_{\ell=2}^{N} C_{T_n^M T_\ell^M}\boldsymbol{\alpha}_\ell = C_{T_n^M T} \end{cases} \tag{5.182}$$

$$(n = 2, \ldots, N),$$

by using the specifications (5.178) through (5.181) we find

$$C_{YY}\boldsymbol{\lambda} + \sum_{\ell=2}^{N} \sigma_\ell^2 \mathbf{M}(\mathbf{S}_\ell^t)\boldsymbol{\alpha}_\ell = C(\mathbf{M}, P) \tag{5.183}$$

$$\sigma_n^2[\mathbf{M}(\mathbf{S}_n^t)]^t\boldsymbol{\lambda} + (\sigma_n^2 I_n + G_n)\boldsymbol{\alpha}_n = \sigma_n^2 \mathbf{S}_n(P). \tag{5.184}$$

The partitioned form of this system suggests to solve (5.184) with respect to $\boldsymbol{\alpha}_n$ and then substitute back into (5.183). In this way, posing

$$\Gamma_n = \sigma_n^2(\sigma_n^2 I_n + G_n)^{-1}, \tag{5.185}$$

$$\boldsymbol{\alpha}_n = \Gamma_n \mathbf{S}_n(P) - \Gamma_n[\mathbf{M}(\mathbf{S}_n^t)]^t \boldsymbol{\lambda}, \tag{5.186}$$

we find

$$(C_{YY} - \sum_{\ell=2}^{n} \sigma_\ell^2 \{[\mathbf{M}(\mathbf{S}_\ell^t)]\Gamma_\ell[\mathbf{M}(\mathbf{S}_\ell^t)]^t\})\boldsymbol{\lambda} \tag{5.187}$$

$$= C(\mathbf{M}, P) - \sum_{\ell=2}^{N} \sigma_\ell^2 \mathbf{M}(\mathbf{S}_\ell^t)\Gamma_\ell \mathbf{S}_\ell(P).$$

As we see, we have now a unique equation in $\boldsymbol{\lambda}$, i.e. (5.187). In order to better understand its meaning we set in clear the components of the relevant matrices and vectors. We have

$$\{[\mathbf{M}(\mathbf{S}_\ell^t)]\Gamma_\ell[\mathbf{M}(\mathbf{S}_\ell^t)]^t\}_{ij} \tag{5.188}$$

$$= \sum_{k,h=-\ell}^{\ell} M_i\{S_{\ell k}(P_i)\}\Gamma_{\ell,kh} M_j[S_{\ell h}(P_j)]$$

$$= M_i \left\{ M_j \left\{ \sum_{k,h=-\ell}^{\ell} \Gamma_{\ell,kh} S_{\ell k}(P_i) S_{\ell h}(P_j) \right\} \right\}.$$

So, if we call

$$C_\Gamma(P, Q) = \sum_{\ell=2}^{N} \sigma_\ell^2 \mathbf{S}_\ell^t(P)\Gamma_\ell \mathbf{S}_\ell(Q) \tag{5.189}$$

we can state that

$$\sum_{\ell=2}^{N} \sigma_\ell^2 [\mathbf{M}(\mathbf{S}_\ell^t)\Gamma_\ell[\mathbf{M}(\mathbf{S}_\ell^t)]^t = C_\Gamma(\mathbf{M}, \mathbf{M}^t). \tag{5.190}$$

Similarly

$$\sum_{\ell=2}^{N} \sigma_\ell^2 \mathbf{M}(\mathbf{S}_\ell^t)\Gamma_\ell \mathbf{S}_\ell(P) = C_\Gamma(\mathbf{M}, P), \tag{5.191}$$

so that (5.187) becomes

$$[C_{yy} - C_\Gamma(\mathbf{M}, \mathbf{M}^t)]\boldsymbol{\lambda} = C(\mathbf{M}, P) - C_\Gamma(\mathbf{M}, P). \tag{5.192}$$

To further elaborate on (5.192) we find

$$C_{YY} - C_\Gamma(\mathbf{M}, \mathbf{M}^t) = C(\mathbf{M}, \mathbf{M}^t) - C_\Gamma(\mathbf{M}, \mathbf{M}^t) + C_v, \tag{5.193}$$

The (5.193) suggests the introduction of the *reduced* covariance

$$\begin{aligned} C(P, Q) - C_\Gamma(P, Q) &= \sum_{\ell=2}^{+\infty} \sigma_\ell^2 \mathbf{S}_\ell^t(P)\mathbf{S}_\ell(Q) - \sum_{\ell=2}^{+\infty} \sigma_\ell^2 \mathbf{S}_\ell^t(Q)\Gamma_\ell \mathbf{S}_\ell(Q) \\ &= \sum_{\ell=2}^{+\infty} \mathbf{S}_\ell^t(P)\sigma_\ell^2(I - \Gamma_\ell)\mathbf{S}_\ell(Q) = \overline{C}(P, Q), \end{aligned} \tag{5.194}$$

where (5.194) we have implicitly introduced the convention that

$$\Gamma_\ell \equiv 0, \ell > N \tag{5.195}$$

so as to extend directly the summation to infinity.

Another remark on (5.194) is that $\overline{C}(P, Q)$ is a true covariance function because the matrices $\sigma_\ell^2(I - \Gamma_\ell)$ are positive definite.

In fact, recalling (5.185),

$$\begin{aligned} \sigma_\ell^2(I - \Gamma_\ell) &= \sigma_\ell^2[(\sigma_\ell^2 I_\ell + G_\ell)^{-1}(\sigma_\ell^2 I + G_\ell) - \sigma_\ell^2(\sigma_\ell^2 I + G_\ell)^{-1}] \\ &= \sigma_\ell^2(\sigma_\ell^2 I + G_\ell)^{-1} G_\ell = \Gamma_\ell G_\ell. \end{aligned} \tag{5.196}$$

Since $I - \Gamma_\ell$ is symmetric and Γ_ℓ, G_ℓ too, one has that $\Gamma_\ell G_\ell = G_\ell \Gamma_\ell$ implying that (5.195) can be written as

$$\sigma_\ell^2(I - \Gamma_\ell) = G_\ell^{(1/2)} \Gamma_\ell G_\ell^{(1/2)}; \tag{5.197}$$

thus showing the positive definiteness of $I - \Gamma_\ell$.

With the help of (5.196) and (5.194) gets the form

$$\overline{C}(P, Q) = \sum_{\ell_2}^{N} \mathbf{S}_\ell^t(P) G_\ell \Gamma_\ell \mathbf{S}_\ell(Q) + \sum_{\ell=N+1}^{+\infty} \sigma_\ell^2 \mathbf{S}_\ell^t(P)\mathbf{S}_\ell(Q) \tag{5.198}$$

Remark 6. Let us assume that the errors of the model coefficients, $\boldsymbol{\tau}_\ell$, have further covariances that are proportional to the identity, i.e. these errors have the same variance per degree and are independent, then one can put

$$G_\ell = \sigma_{\tau\ell}^2 I = \frac{\varepsilon_\ell}{2\ell+1} I \ , \ \sigma_\ell^2 = \frac{c_\ell}{2\ell+1}, \ \Gamma_\ell = \rho_\ell I, \ \rho_\ell = \frac{\sigma_\ell^2}{\sigma_\ell^2 + \sigma_{\tau\ell}^2}$$

and one finds

$$\overline{C}(P,Q) = \sum_{\ell=2}^{N} \rho_\ell \varepsilon_\ell \left(\frac{R^2}{r_P r_Q}\right)^{\ell+1} P_\ell(\cos\psi_{PQ}) + \sum_{\ell=N+1}^{+\infty} c_\ell(T) \left(\frac{R^2}{r_P r_Q}\right)^{\ell+1} P_\ell(\cos\psi_{PQ}). \tag{5.199}$$

This is an almost perfect counterpart of (5.139) and (5.140) with the difference that here we are using the reduced covariance of T, there the local covariance of Δg.

The most remarkable difference between (5.199) and (5.139) is in the factors $\rho_\ell = \frac{\sigma_\ell^2}{\sigma_\ell^2+\sigma_{\tau\ell}^2}$ multiplying the error degree variances.

On account of the identity

$$\rho_\ell \varepsilon_\ell = \frac{\sigma_\ell^2 \sigma_{\tau\ell}^2 (2\ell+1)}{\sigma_\ell^2 + \sigma_{\tau\ell}^2} = \frac{\sigma_{\tau\ell}^2}{\sigma_\ell^2 + \sigma_{\tau\ell}^2} c_\ell = \chi_\ell c_\ell$$

we see that (5.199) can be written as well as

$$\overline{C}(P,Q) = \sum_{\ell=2}^{+\infty} \chi_\ell c_\ell(\tau) \left(\frac{R^2}{r_P r_Q}\right)^{\ell+1} P_\ell(\cos\psi_{PQ}) \tag{5.200}$$

if we agree that $\chi_\ell \equiv 1$ when $\ell > N$. The form (5.200) shows clearly that the role of the error $\tau_{\ell m}$ is to turn down the degree variances of T when the ratio signal to noise is high while it leaves c_ℓ unaltered for the high degrees of the model where $\sigma_{\tau\ell}^2$ becomes larger. Note however that if we stop the model at N such that $\sigma_{\tau\ell}^2 = \sigma_\ell^2$, when $\ell = N$, then we have $\chi_N = \frac{1}{2}$.

Another remark is that the degrees above N in (5.199) can be modelled on the basis of local data as described in Sect. 5.8 of this chapter.

In terms of $\overline{C}$ our reduced normal system (5.192) becomes

$$(\overline{C}(M,M) + C_v)\boldsymbol{\lambda} = \overline{C}(\mathbf{M},P), \tag{5.201}$$

implying the solution of a classical collocation normal system with covariance $\overline{C}(P,Q)$. Once $\boldsymbol{\lambda}$ is found form (5.201), we can go back to (5.186) and we can write

$$\boldsymbol{\alpha}_n^t = \mathbf{S}_n^t(P)\Gamma_n - \boldsymbol{\lambda}^t \mathbf{M}(\mathbf{S}_n^t)\Gamma_n. \tag{5.202}$$

Therefore (5.176) gives

$$\widehat{T}(P) = \boldsymbol{\lambda}^t \mathbf{Y} + \sum_{n=2}^{N} \mathbf{S}_n^t(P) \Gamma_n \mathbf{T}_n^M + \\ -\boldsymbol{\lambda}^t \mathbf{M} \left(\sum_{n=2}^{N} \mathbf{S}_n^t \Gamma_n \mathbf{T}_n^M \right). \tag{5.203}$$

This suggests to introduce a modified model

$$T_\Gamma(P) = \sum_{n=2}^{+\infty} \mathbf{S}_n^t(P) \Gamma_n \mathbf{T}_n^M \tag{5.204}$$

so that (5.203) writes

$$\begin{aligned} \widehat{T}(P) &= \boldsymbol{\lambda}^t \mathbf{Y} - \boldsymbol{\lambda}^t \mathbf{M}(T_\Gamma) + T_\Gamma(P) \\ &= \boldsymbol{\lambda}^t \{\mathbf{M}(T - T_\Gamma) + \boldsymbol{\nu}\} + T_\Gamma(P). \end{aligned} \tag{5.205}$$

So our optimal solution is in fact the result of a remove-restore procedure, where the optimal model to be used however is not simply

$T^M(P) = \sum_{n=2}^{N} \left(\mathbf{T}_n^M\right)^t \mathbf{S}_n(P)$, but rather $T_\Gamma(P)$.

It is noteworthy that in accordance with this interpretation, the normal equation for $\boldsymbol{\lambda}$, (5.201), can be viewed as an ordinary collocation equation if we observe that $\overline{C}(P, Q)$ is in reality the covariance function of $T - T_\Gamma = v(P)$. In fact

$$v(P) = T - T_\Gamma = \sum_{n=2}^{N} \mathbf{S}_n^t(P)(I - \Gamma_n)\mathbf{T}_n + \\ -\sum_{n=2}^{N} \mathbf{S}_n^t(P) \Gamma_n \boldsymbol{\tau}_n + \sum_{n=N+1}^{+\infty} \mathbf{S}_n^t(P) \mathbf{T}_n \tag{5.206}$$

so that, by covariance propagation

$$C_{vv}(P, Q) = \sum_{n=2}^{N} \mathbf{S}_n^t(P) \sigma_n^2 (I - \Gamma_n)^2 \mathbf{S}_n(Q) \\ + \sum_{n=2}^{N} \mathbf{S}_n^t(P) \Gamma_n G_n \Gamma_n \mathbf{S}_n(Q) + \sum_{n=N+1}^{+\infty} \sigma_n^2 \mathbf{S}_n^t(P) \mathbf{S}_n(Q). \tag{5.207}$$

With the help of (5.196), it is not difficult to prove that

$$\sigma_n^2(I-\Gamma_n)^2+\Gamma_nG_n\Gamma_n=\Gamma_nG_n(I-\Gamma_n)+\Gamma_nG_n\Gamma_n$$
$$=\Gamma_nG_n=G_n\Gamma_n, \tag{5.208}$$

so that (5.207) is identical with (5.198). Let us observe that the covariance $\overline{C}(P,Q)$ in general is not isotropic unless the conditions $G_\ell=\sigma_{\tau\ell}^2 I$, studied in Remark 6, are satisfied.

Therefore $\overline{C}(P,Q)$, in the low degrees components, should not be empirically estimated in the usual way if the mentioned conditions are not fulfilled. In fact, if we do so, we loose information on the stochastic structure of $\boldsymbol{\tau}_\ell$.

Although the ideas presented in this section have been formulated since some years their implementation in numerical tests is relatively recent (Pail et al. 2010). These however have given good results in both cases, the estimation of global enhanced models or the prediction of very local geoid models. In this respect we have confirmed the guess that the hypothesis of zero local mean value for Δg_r is not required in the present situation.

A final point is worth mentioning, on the interpretation of $\boldsymbol{\tau}_\ell$, i.e. errors in the model coefficients. These errors have been usually interpreted as the propagation to $\mathbf{T}_\ell^M$ of the noise present in the observations used in their estimation. This certainly accounts for the difference of $\mathbf{T}_\ell^M$ with respect to the true T_ℓ. This point of view has been taken up in Sect. 5.8.

However when we model a local covariance function and we compare the statistical behaviour of the low degrees coefficients between their global definition and their local appearance in the area A, we might find a considerable difference between the two, specially on account of the dimension of A. In this respect, consider that an area of $10^\circ\times 10^\circ$ is just $\frac{1}{648}$ times the area of the whole sphere. Although there are in literature examples of attempts to model even globally non homogenous covariances (Rummel and Schwarz 1977) we feel that the subject is far from being settled. So we just state here that, the way in which this kind of variability, that is reflected into a *localization error* for $\mathbf{T}_\ell$, could be included and accounted for into our data analysis, will be object of future research.

5.12 Exercises

Exercise 1. Let $(\mathbf{r})=(x_1,x_2,x_3)$ be a Cartesian coordinate system and $(\mathbf{r}')=(x_1',x_2',x_3')$ another Cartesian system rotated with respect to the first. Assume that $T(\mathbf{r})=T(x_1,x_2,x_3)$ is a harmonic function in an open set Ω, that the rotation transforms into the open Ω'. Put

$$v(x_1',x_2',x_3')=T[x_1(\mathbf{r}'),x_2(\mathbf{r}'),x_3(\mathbf{r}')]\,;$$

prove that $v(x_1',x_2',x_3')$ is harmonic in Ω'.

(**Hint**: note that

$$v[\mathbf{r}'(\mathbf{r})] \equiv T(\mathbf{r})$$

and observe that

$$x_i' = \Sigma_k R_{ik} x_k,$$

where $R \equiv [R_{ik}]$ is the rotation matrix between $(\mathbf{r})$ and $(\mathbf{r}')$. Recall that $R^t R = I$. Compute $\Sigma_i \frac{\partial^2}{\partial x_i^2} T$ by using the chain rule and prove that

$$\Sigma_i \frac{\partial^2 T}{\partial x_i^2} = \Sigma_k \frac{\partial^2 v}{\partial x_k'^2})$$

Exercise 2. Compute in spectral form and in spherical approximation the following covariances and cross-covariances

$$C_{\delta g \delta g}(P,Q),\ C_{\delta g \Delta g}(P,Q),\ C_{T \delta g}(P,Q),$$
$$C_{T_{rr} T_{rr}}(P,Q),\ C_{T_{rr} \Delta g}(P,Q).$$

Furthermore, put $T_\vartheta = \frac{\partial}{\partial \vartheta} T(P)$ and compute $C_{T_\vartheta T_\vartheta}(P,Q)$, following the last calculation of Example 4.

Exercise 3. Recalling the definition (5.148), assume that

$$\rho_T(n) = \frac{C_T}{n+1}\ ; \tag{5.209}$$

show that the corresponding degree variances of T and Δg are

$$c_n(T) = \frac{C_T}{(n+1)} \left(\frac{R_B^2}{R^2}\right)^{n+1}$$
$$c_n(\Delta g) = \frac{C_T}{R_B^2} \frac{(n-1)^2}{(n+1)} \cdot \left(\frac{R_B^2}{R^2}\right)^{n+2}.$$

(**Hint:** compare (5.130), (5.135) and (5.138) with (5.147) and recall the relation (5.104)).

Exercise 4. Consider the covariance function of T when (5.209) holds; prove that

$$C_{TT}(s,t) = C_T \left[\log \frac{s - t + G^{-1}(s,t)}{1-t} - s - \frac{1}{2} t s^2\right]. \tag{5.210}$$

(**Hint:** note that $\frac{s^{n+1}}{n+1} = \int_0^s \sigma^n d\sigma$; use this in (5.148), exchange integration and summation and use (5.149) and (5.152)).

Exercise 5. Prove that, with the covariance (5.210),

$$C_{T\delta g}(P,Q) = -\frac{\partial}{\partial r_P} C_{TT}(s,t) \tag{5.211}$$
$$= C_T \frac{s}{r_P}[G(s,t) - 1 - ts]$$
$$C_{\delta g \delta g}(P,Q) = -\frac{\partial}{\partial r_Q} C_{T\delta g}(s,t) \tag{5.212}$$
$$= \frac{C_T}{R_B^2} s^2 [(1-ts)G^3(s,t) - 1 - 2ts] \; ;$$

then find the corresponding crosscovariances and covariances $C_{T\Delta g}, C_{\Delta g \Delta g}$, by propagation through the linear relation

$$\Delta g(P) = \delta g(P) - \frac{2}{r_P} T(P). \tag{5.213}$$

Exercise 6. Put into (5.210) $r_P = r_Q = R = 6,371$ and $R_B = 6,361$; moreover, compute the covariance at the origin, i.e. $\psi = 0 \Rightarrow t = 1$, and impose that

$$C_{TT}(s,1) = \sigma^2(T) = \gamma^2 \sigma^2(N)$$
$$= 978^2\,\mathrm{Gal}^2 \cdot 1^2\,\mathrm{m}^2 \cong 0.956 \cdot 10^6\,\mathrm{Gal}^2\,\mathrm{m}^2$$

show that in this case

$$C_T = 0.224 \cdot 10^6\,\mathrm{Gal}^2\,\mathrm{m}^2.$$

By using this value in (5.212) show that

$$C_{\delta g \delta g}(s,1) \cong 559 \cdot 10^{-6}\,\mathrm{Gal}^2$$

i.e.

$$\sigma(\delta g) \cong 23.6\,\mathrm{mGal}.$$

In other words a mean square geoid of 1m with the spectrum implied by (5.209) corresponds to a mean square gravity disturbance of 23.5 mGal.

The reader is warned that these numbers do not refer to the true gravity field but they are just realistic.

(**Hint:** note that if one puts $t = 1$ in (5.210) one gets the indefinite form $\frac{0}{0}$. Therefore the limit for $t \to 1$ has to be computed by the de l'Hopital rule.)

Exercise 7. Assume that two values of geoid N_{-1}, N_1 are observed without noise at -1 km and 1 km from the origin respectively (see Fig. 5.11).

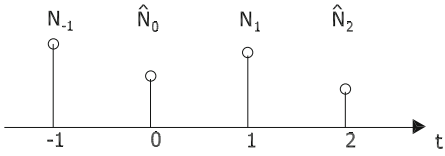


Fig. 5.11 Observed and predicted values according to the exercise

Assume that the covariance of N along the axis t (cf. Fig. 5.11) is given by

$$C(t_1, t_2) = q^{|\tau|} = e^{-\alpha|\tau|}$$
$$\tau = t_1 - t_2, \quad \alpha = \log \frac{1}{q},\ q < 1.$$

Prove that the optimal prediction $\widehat{N}(t)$ at $t = 0$ and $t = 2$ is given by

$$\widehat{N}(0) = \frac{9}{1 - q^4}(N_{-1} + N_1)$$
$$\widehat{N}(2) = qN_1$$

and the corresponding quadratic prediction errors are

$$\mathcal{E}^2(0) = 1 - \frac{2q^2}{1 + q^2}$$
$$\mathcal{E}^2(2) = 1 - q^2.$$

Note that $\mathcal{E}^2(2) > \mathcal{E}^2(0)$ because the extrapolation error is larger than the interpolation error. For instance, with $q^2 = \frac{1}{2}$ one has $\mathcal{E}^2(0) = \frac{1}{3}$, $\mathcal{E}^2(2) = \frac{1}{2}$.

Exercise 8. Assume that the geoid $N(t)$ along a section (line) has covariance

$$C(t_1, t_2) = e^{-\alpha\tau^2}$$
$$\tau = t_1 - t_2.$$

Assume that one has observed at $t = 0$ both the geoid $N_0 = N(0)$ and its derivative $\varepsilon_0 = \frac{dN}{dt}(0)$, i.e. basically the deflection of the vertical changed of sign. The observation noises have respectively standard deviations σ_N and σ_ε.

Compute the prediction $N(t)$ and the corresponding prediction error for every t and verify that

$$\widehat{N}(t) = e^{-\alpha\tau^2}\left[\frac{1}{1 + \sigma_N^2} N_0 + \frac{2\alpha}{2\alpha + \sigma_\varepsilon^2}\tau\varepsilon_0\right]$$
$$\mathcal{E}^2(t) = 1 - e^{-2\alpha\tau^2}\left[\frac{1}{1 + \sigma_N^2} + \frac{4\alpha^2}{2\alpha + \sigma_\varepsilon^2}\right].$$

(**Hint**: first compute for any t_1, t_2 the functions $C(t_1, t_2)$, $\frac{\partial}{\partial t_2} C(t_1, t_2)$, $\frac{\partial^2}{\partial t_1 \partial t_2} C(t_1, t_2)$ and then put $t_1 = t_2 = 0$.

Note that in this way $C(N(0), \varepsilon(0)) = \frac{\partial}{\partial t_2} C(t_1, t_2)\Big|_{t_1=t_2=0} = 0$.).

Exercise 9. We use the same symbols and the same covariances of Exercise 5. Assume one has measured without noise δg at a point Q, put $t = \cos \psi_{PQ}, r_P = R$, and predict $\widehat{N}(P)$ for every P. In particular prove that, choosing $P = Q$ (i.e. $t = 1$) one has

$$\widehat{N}(Q) = \frac{\widehat{T}(Q)}{\gamma} = \frac{R_B}{\gamma} \frac{(1-s)}{[1-(1-s)^2(1+2s)]} \delta g(Q)$$

$$\mathcal{E}^2(Q) = C_T \left\{ \log \frac{1}{1-s} - s - \frac{1}{2}s^2 - \frac{s^6}{[1-(1-s)^2(1+2s)]} \right\}$$

Appendix

A.1

We want to prove the relation (5.98), sending the interested reader to the literature Moritz (1980) and Sansò (1986) for the distribution of the vector $\mathbf{T}(\omega)$.

We have

$$\begin{aligned} \sum_{m=-n}^{n} T_{nm}^2(\omega) &= \frac{1}{(4\pi)^2} \int d\sigma_P T(R_\omega P) \int d\sigma_Q T(R_\omega Q) \\ &\quad \cdot \sum_{m=-n}^{n} Y_{nm}(\vartheta_P, \lambda_P) Y_{nm}(\vartheta_Q, \lambda_Q) \\ &= \frac{2n+1}{(4\pi)^2} \int d\sigma_P \int d\sigma_Q T(R_\omega P) T(R_\omega Q) P_n(\cos \psi_{PQ}) \\ &= \frac{2n+1}{(4\pi)^2} \int d\sigma_{P'} \int d\sigma_{Q'} T(P') T(Q') P_n(\cos \psi_{P'Q'}) \, ; \end{aligned} \tag{5.214}$$

the last equality is justified because $\psi_{PQ} = \psi_{P'Q'}$ and the double integral over the sphere can be performed with any angular coordinates giving always the same result.

Now we organize the double integral in (5.214) as follows; first fix P' and let Q' circulate around P' at a distance $\psi_{P'Q'} = \psi$; then integrate in $d\sigma_{P'}$; then we finally let ψ to vary from 0 to π. We get, putting $d\sigma_{Q'} = \sin \psi d\psi d\alpha$ into (5.214), recalling also the definition (5.38) and using (5.94),

$$\sum_{m=-n}^{n} T_{nm}^2(\omega) = \frac{(2n+1)}{2}\int_0^\pi d\psi \sin\psi P_n(\cos\psi)\cdot \tag{5.215}$$
$$\cdot\frac{1}{8\pi^2}\int d\sigma_{P'}\int_{\psi_{P'Q'}=\psi} T(P')T(Q')d\alpha_{Q'}$$
$$= \frac{2n+1}{2}\int_0^\pi d\psi \sin\psi P_n(\cos\psi)C(\psi) = c_n,$$

as it was to be proved.

A.2

We want to prove formula (5.156), providing the explicit form of $K_B(s,t)$ and $K_{-2}(s,t)$. We first expand (5.157) into the sum of fractions, with the identity

$$\frac{n-1}{(n-2)(n+B)} \equiv \frac{B+1}{B+2}\frac{1}{n+B} + \frac{1}{B+2}\frac{1}{n-2}$$

so that we can write

$$C_{\Delta g\Delta g}(s,t) = A\left\{\frac{B+1}{B+2}\sum_{n=3}^{+\infty}\frac{s^{n+2}}{n+B}P_n(t) + \frac{1}{B+2}\sum_{n=3}^{+\infty}\frac{s^{n+2}}{n-2}P_n(t)\right\}$$
$$= A\left\{\frac{B+1}{B+2}K_B(s,t) + \frac{1}{B+2}K_{-2}(s,t)\right\} \tag{5.216}$$

We compute at first the last term:

$$K_{-2}(s,t) = s^4\sum_{n=3}^{+\infty}\frac{s^{n-2}}{n-2}P_n(t) \tag{5.217}$$
$$= s^4\int_0^s\sum_{n=3}^{+\infty}\sigma^{n-3}P_n(t)d\sigma$$
$$= s^4\int_0^s\frac{1}{\sigma^3}\left\{\sum_{n=0}^{+\infty}\sigma^n P_n(t) - 1 - \sigma t - \sigma^2 P_2(t)\right\}d\sigma$$
$$= s^4\int_0^s\frac{1}{\sigma^3}\left\{G(\sigma,t) - 1 - \sigma t - \sigma^2 P_2(t)\right\}d\sigma$$
$$= \frac{s^2}{2}[1 + 2ts - (3ts+1)G^{-1}(s,t)] - s^4 P_2(t)\log\frac{1 - st + G^{-1}(s,t)}{2}$$
$$+s^4\frac{7t^2-1}{4}.$$

The last integral is calculated with the help of mathematical tables adjusting the integration constant in such a way that both members of (5.217), multiplied by s^{-4}, tend to 0 when s tends to 0. As for the first term one writes, assuming $B > 0$,

$$\begin{aligned} K_B(s,t) &= s^{2-B}\sum_{n=3}^{+\infty}\frac{s^{n+B}}{n+B}P_n(t) \\ &= s^{2-B}\int_0^s\sum_{n=3}^{+\infty}\sigma^{n+B-1}P_n(t)d\sigma \\ &= s^{2-B}\int_0^s\sigma^{B-1}\left\{\sum_{n=0}^{+\infty}\sigma^n P_n(t) - 1 - \sigma t - \sigma^2 P_2(t)\right\}d\sigma \\ &= s^{2-B}\int_0^s\sigma^{B-1}G(\sigma,t)d\sigma - \frac{s^2}{B} - \frac{s^3}{B+1}t - \frac{s^4}{B+2}P_2(t). \end{aligned} \tag{5.218}$$

Now the integrals

$$I_B = \int_0^s \sigma^{B-1}G(\sigma,t)d\sigma \tag{5.219}$$

can be computed, for integer values of B, by exploiting a recursive relation, namely

$$I_{k+1} = \frac{s^{k-1}}{k}G^{-1}(s,t) + \frac{(2k-1)}{k}tI_k - \frac{k-1}{k}I_{k-1} \tag{5.220}$$

which is derived from the identity

$$\frac{\partial}{\partial s}\left[s^{k-1}G^{-1}(s,t)\right] = \left[ks^k - (2k-1)ts^{k-1} + (k-1)s^{k-2}\right]G(s,t), \tag{5.221}$$

integrating both members from 0 to s and re-arranging. In order to trigger (5.220) we need two initial values of I_k, for instance I_1, I_2. But I_1 has already been given in (5.152) and I_2 is easy to compute since, recalling (5.151),

$$\begin{aligned} I_2 &= \int_0^s \sigma G(\sigma,t)d\sigma = \int_0^s(\sigma - t)G(\sigma,t)d\sigma + t\int_0^s G(\sigma,t)d\sigma \\ &= G^{-1}(s,t) - 1 + tI_1. \end{aligned} \tag{5.222}$$

The relations (5.216), (5.217), (5.220), (5.152) and (5.222) all together give the explicit form of the covariance function of Δg for every integer B. For a global use of this covariance the model (3.181) coming from the best fit of EGM08 degree variances between degrees 180 and 1,800, can be used, with the only warning that

in (3.181) one has $\overline{\sigma}_\ell^2 = c_\ell\left(\frac{T}{\gamma}\right)$, whereas we treat here $c_n(\Delta g)$ related to the former by the relation

$$c_n(\Delta g) = \frac{(\ell-1)^2}{\overline{R}^2} c_n(T) = \frac{(\ell-1)^2}{\overline{R}^2}\left(\frac{GM}{\overline{R}^2}\right)^2 \overline{\sigma}_\ell^2. \tag{5.223}$$

We notice by the way that also the improved model (3.178) transformed according to (5.223) can be added by applying exactly the same methods presented in the Appendix and the decomposition

$$\frac{\ell-1}{(\ell-2)(\ell+4)(\ell+17)} = \frac{1}{114}\frac{1}{\ell-2} + \frac{5}{78}\frac{1}{\ell+4} - \frac{18}{247}\frac{1}{\ell+17}. \tag{5.224}$$

Part II
Methods and Applications

Chapter 6
Global Gravitational Models

Nikolaos K. Pavlis

6.1 Outline of the Chapter

This chapter discusses the development and use of *Global Gravitational Models* (GGMs), specifically those GGMs that are represented in the form of spherical (and/or ellipsoidal) harmonic coefficients. With the mathematical details having been presented in Chap. 3 of Part I of this book, the focus here is on the main concepts and considerations involved in the design and in the choice of alternative techniques and strategies that can be used to develop GGMs. Recent advances in geodetic techniques, in particular the availability of dedicated geopotential mapping missions on one hand and the availability of very high resolution GGMs on the other, provide the natural setting for the discussion that follows. Section 6.2 provides an introductory overview of the main concepts and distinguishes between *Global* and *Regional (or Local)* models, the latter being discussed in subsequent chapters within this part of the book. Section 6.3 discusses the aspects involved with the representation of GGMs and the characteristics of the data that are used to create the GGMs. Section 6.4 discusses the new satellite missions that are dedicated to the mapping of the gravitational field from space, and the advances and challenges that these missions introduce to GGM developments. Section 6.5 discusses the combination of the gravitational information obtained from satellites with the information obtained from surface data, which permit the development of very high resolution GGMs like EGM2008. Sections 6.2–6.5 provide the main concepts underlying the development of GGMs, omitting intentionally the mathematical and numerical details. In contrast, Sect. 6.6 discusses in some detail the specific mathematical and numerical procedures that may be used for the development of GGMs. For this purpose, two models are used as representative examples in Sect. 6.6 – EGM96, which represents the state-of-the-art *before* the availability of data from CHAMP and GRACE, and EGM2008, which represents currently the global model with the highest accuracy (developed *prior* to the availability of data from GOCE) and also the highest resolution. Section 6.7 discusses briefly the data requirements and the availability of the data necessary to develop GGMs. Section 6.8 deals with several

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_6,

aspects related to the use of a GGM and its by-products. The focus here is on the computation of the geoid, especially with regards to the treatment of permanent tide effects and the computation of height anomalies and geoid undulations referring to some specified ellipsoid of revolution and its normal gravity potential. Section 6.9 briefly discusses temporal (non-tidal) variations of the gravitational potential arising from the redistribution of mass within the Earth system, and the very recent advances in the monitoring and mapping of these variations from space, which resulted from the analysis of data from the GRACE satellite mission. Finally, Sect. 6.10 provides some outlook. This entire chapter is written in a way that focuses mostly on the concepts associated with global gravitational modeling, and the evolution of the art and science of the development of GGMs during the last 25 years or so. A rather extensive list of references is provided, so that the reader will be able to locate specific documents that provide the mathematical details associated with various aspects of GGM developments.

6.2 Introduction

A *Global Gravitational Model* (GGM) is a mathematical approximation to the external gravitational potential of an attracting body. We will focus here on the case where the attracting body is the Earth, although many of the concepts that we discuss apply equally well to other planets and celestial bodies. A GGM consists of a set of numerical values for certain parameters, the statistics of the errors associated with these values (as expressed, e.g., in their error covariance matrix), and a collection of mathematical expressions, numerical values, and algorithms that allow a user to perform:

1. **Synthesis**, i.e., computation of the numerical values of quantities related to the gravitational potential (*functionals* of the field), given the position of the evaluation point.
2. **Error Propagation**, i.e., computation of the expected errors of the computed functionals, as implied by the propagation of the errors of the parameters defining the GGM.

A GGM must be able to support such computations at arbitrary points, located on or above the Earth's surface, in a fashion that is both rigorous and efficient. In addition, a GGM should fulfill certain conditions stemming from the underlying physics. Namely, it should represent a scalar function of position that is harmonic outside the attracting masses and vanishes at infinity as the reciprocal of the distance between attracted point and attracting mass element. Moreover, the GGM should permit the computation of any functional of the field in a way that guarantees *self-consistency*. This means that the model should preserve the relationships (differential or integral) between the various functionals. A GGM has numerous uses, both operational and scientific (see also Tscherning 1983), including:

1. Orbit determination applications necessary for space surveillance (the detection, tracking, and orbit prediction of Earth-orbiting objects).

2. Inertial navigation applications for trajectory determination of airplanes and missiles.
3. Geoid undulation computations necessary to transform a geometric height, to an elevation referenced to an equipotential surface. This application has attracted great interest in recent years, because GPS positioning and gravimetrically determined geoid heights offer the possibility of determining orthometric heights and height differences without the need for the expensive and laborious spirit leveling (Schwarz et al. 1987).
4. Oceanographic applications that require the estimation of the Dynamic Ocean Topography (DOT) and its slopes, quantities that are directly related to ocean circulation. This application puts very stringent accuracy and resolution requirements on GGMs (Ganachaud et al. 1997).
5. A unique, accurate high resolution GGM may be used to provide the reference surface for the realization of a *Global Vertical Datum* (Rapp and Balasubramania 1992).
6. Geophysical prospecting applications where, in combination with other information (e.g., seismic data), a GGM may provide important constraints that aid the determination of underlying density distributions.

These and other applications represent integral parts of various civilian and military activities. Each of these applications has (in general) different accuracy and resolution requirements, as far as the supporting GGM is concerned. For example, due to the attenuation of the gravitational field with increasing altitude, a relatively low resolution GGM (e.g., a spherical harmonic expansion to maximum degree 70 or 90) is currently adequate for the precise orbit determination of most Earth-orbiting satellites. In contrast, accurate determination of the slopes of the equipotential surface (deflections of the vertical) demands a GGM of much higher resolution.

Geodesists have at various times developed "special purpose" models that optimize performance for a particular application (e.g., orbit determination of a particular satellite, or geoid undulation computation over a specific geographic region). Although such "tailored" models have found some uses in the past, the ultimate goal has always been the development of a *unique*, general purpose, GGM that addresses the different and diverse applications in an optimal manner, without over-performing in one application at the expense of its performance in others.

The development of a high-resolution GGM is a task that involves the optimal combination of a variety of data (satellite, land, marine, airborne). This is because a single data type with both global coverage and with uniformly high accuracy and high spectral sensitivity does not (yet) exist. The aforementioned data are of complementary character (in terms of spectral sensitivity and/or of geographic coverage), so that their optimal combination enables a GGM to satisfy the variety of applications described before. "Class" solutions of this type (e.g., EGM96) may include not only parameters that describe the gravitational potential, but also parameters that describe the Dynamic Ocean Topography, tides, Earth orientation and tracking station position parameters, as well as a plethora of "nuisance" parameters necessary to model completely the content of certain data types (e.g., biases and

delays associated with certain satellite tracking data). The result of a successful GGM development effort is a model that can be used as a *standard* for numerous applications, over a substantial period of time.

6.2.1 Local and Regional Gravimetric Models

The accuracy and the resolving power of the data that were used in its development dictate the accuracy and resolution of a GGM. Geopolitical and/or proprietary issues many times prevent the individual or the team developing a GGM from having access to *all* the existing data. However, over some regions, data of higher accuracy and/or resolving power (geographically dense sets of gravity and elevation data) may be available to some individual(s) or may become available *after* the reference GGM has been developed. These data may be used in combination with the existing GGM to improve the accuracy and/or resolution of the determination of *one or more specific functionals of the field*, over the region where the detailed data became available. This local or regional "densification" can produce a specific local or regional gravimetric product or model.

Such densification has been among the favorite geodetic activities over many decades now, and represents the geodesist's way of creating a *multi-resolution* gravitational model resembling a "quilt": i.e., patches of fine detail (the *Local Gravimetric Models* – LGMs) are sewn on top of a more or less homogeneous piece of fabric (the reference GGM). *Geodesists do not necessarily have to re-evaluate the reference GGM every time a new set of data (a new patch) becomes available locally*. Such re-evaluation is mostly warranted if new and improved satellite data become available, spanning a sufficiently long time period, and/or if new terrestrial data (of higher accuracy and/or resolution) become available over areas with substantial geographic extent.

6.2.2 Global Versus Local Gravimetric Models: Similarities and Differences

It is useful to consider some important points related to the development and the nature of global and local gravimetric models.

- The most time-consuming, expensive, and laborious task in the development of both global and local gravimetric models is the data collection, validation, and pre-processing. In comparison, the time and effort required for the model estimation is almost negligible.
- Existing global gravitational models, developed using spherical harmonics as the representational basis, allow the computation of *any* functional of the field (geoid undulations, gravity anomalies, deflections of the vertical, second order gradients

of the potential) *anywhere* outside the attracting masses. These computed values are, of course, subject to commission and omission errors (see Sect. 6.6.2.4 for the definition of these terms). In contrast, currently available local or regional gravimetric models consist usually of geographic grids containing the estimated values of one or more *specific* functionals of the field (e.g., geoid undulations, deflections of the vertical), but *cannot* support the computation of arbitrary field functionals at arbitrary locations.

- Global gravitational models are accompanied by increasingly more complete and reliable error estimation. In contrast, existing local or regional gravimetric models are seldom accompanied by error statistics computed rigorously from the error estimates of the input data.
- Determination of a global gravitational model is *not* an interpolation problem. The gravimetric geoid surface is *not* directly observable. Multi-resolution representations that have been used in some studies to decompose and depict various levels of detail within a *given* geoid surface, address (at best) interpolation problems but fall short of addressing the much more challenging and important gravimetric estimation problems. The geodesist's main problem is how to determine the geoid (and other field functionals) from heterogeneous and noisy data – not how to interpolate it, once it has been determined.
- Determination of a local/regional model of a given gravimetric functional *may* reduce to an interpolation/extrapolation problem, if the functional of interest (e.g., gravity anomaly) is also observed *directly* within the area of interest.

The argument that the use of spherical (or ellipsoidal) harmonics as the representational basis for a GGM has the disadvantage that local data updates necessitate global updates of the model (re-computation of the GGM), would have been true if geodesists relied *only* on the GGM for *all* gravimetric applications. At present they do not. LGMs can be used to address efficiently local and regional data updates and applications. Furthermore, even if the re-computation of a GGM is required, due to some specific update of regional surface gravity data, such a re-computation can be done very efficiently and expeditiously at present, as long as the underlying satellite-only model does not have to be re-computed.

6.3 Signal Representation and Data Characteristics

Although geodesists have variously considered and studied the representation of the gravitational potential using point masses (Sünkel 1981b, 1983), finite element methods (Meissl 1981; Baker 1988) and splines (Sünkel 1984; Jekeli 2005), these approaches have seen only limited application in the representation of (especially) the "static" (i.e., time-averaged) gravitational field of the Earth. Spherical harmonics have prevailed as the standard form used for the representation of the gravitational potential globally, from the very early days of global determinations, to the present. Indeed, the set of coefficients of a spherical harmonic expansion of the gravitational

potential has become pretty much synonymous to a GGM. Rapp (1998) provides a review of the geopotential modeling developments of the twentieth century, which includes an extensive list of references.

The Earth's external gravitational potential, V, at a point P defined by its geocentric distance (r_P), geocentric co-latitude (θ_P) and longitude (λ_P), can be expressed as:

$$V(r_P, \theta_P, \lambda_P) = \frac{GM}{r_P}\left[1 + \sum_{n=2}^{\infty}\left(\frac{a}{r_P}\right)^n \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta_P, \lambda_P)\right]. \tag{6.1}$$

GM is the geocentric gravitational constant (the product of the universal gravitational constant, G, times the mass of the Earth including its atmosphere, M) and a is a scaling factor associated with the fully-normalized, unitless, spherical harmonic coefficients C_{nm} (a is usually numerically equal to the equatorial radius of an adopted mean-Earth ellipsoid). The surface spherical harmonic functions are defined as (Heiskanen and Moritz 1967, Sect. 1-14):

$$Y_{nm}(\theta_P, \lambda_P) = \overline{P}_{n|m|}(\cos\theta_P) \cdot \begin{Bmatrix} \cos m\lambda_P \\ \sin |m|\,\lambda_P \end{Bmatrix} \begin{matrix} \text{if } m \geq 0 \\ \text{if } m < 0. \end{matrix} \tag{6.2}$$

$\overline{P}_{n|m|}(\cos\theta_P)$ is the fully-normalized associated Legendre function of the first kind, of degree n and order $|m|$. In practice, the degree summation is truncated to some finite degree N, which depends on the resolving power of the available data. In turn, N defines (approximately) the resolution of the GGM. The goal of global high-resolution gravitational modeling is to estimate, as accurately as possible, the coefficients C_{nm}, through the optimal combination of gravitational information from a variety of data sources. Of equal importance is the estimation of reliable error estimates for the C_{nm} values. The estimated C_{nm} values can then be used to compute functionals of the field (e.g., geoid undulations, gravity anomalies, etc.) while their associated errors (and error correlation when available) can be propagated to yield the errors of the derived functional(s). *Before* the dawn of the new millennium and the availability of data from the satellite missions CHAMP and GRACE, four kinds of gravitational information were commonly available for the development of high-degree combination gravitational models like EGM96 (Lemoine et al. 1998):

1. Information obtained from the analysis of satellite orbit perturbations that are deduced from tracking data. This is of critical importance for the accurate determination of the low degree part of the model. *Satellite-only* models have progressed from solutions to degree 4 in the early 1960s, to models complete to degree 70 or 90 available at present. These advances were made through the availability of ever more accurate tracking data acquired over a continuously expanding constellation of Earth orbiters. Tracking data from approximately 40 satellites have been used in the development of the satellite-only solution supporting EGM96 (denoted EGM96S) (Lemoine et al. 1998). These data include

optical, radio Doppler and radio interferometric observations, Satellite Laser Ranging (SLR), Doppler Orbit determination and Radiopositioning Integrated on Satellite (DORIS), and Satellite-to-Satellite Tracking (SST) data from the Global Positioning System (GPS) and Tracking and Data Relay Satellite System (TDRSS) constellations to lower Earth orbiters. Despite these advances, these tracking data types are incapable of resolving the fine structure of the field, due to the attenuation of the gravitational signal with altitude. Moreover, the available satellites do not sample uniformly the range of inclinations and altitudes, which is a necessary condition for the de-correlation of the harmonic coefficients estimated from satellite tracking data only. This causes strong correlation especially among coefficients of higher degrees and necessitates the use of a priori constraints in the development of satellite-only models (Lerch et al. 1979).

2. Surface (land, marine, and airborne) gravimetric data that are in principle capable of resolving both long and short wavelength features of the gravity field. This however requires uniform global coverage with dense gravity data of uniformly high accuracy. The best available data sets circa 1996 (Kenyon and Pavlis 1996) represent information derived from over 4,000 sources of detail gravity data collected over several decades. The accuracy and density of point data vary substantially with geographic region, with extended regions (e.g., Antarctica) being practically void of gravity measurements. Gravity anomaly data are susceptible to various systematic errors (Heck 1990). These errors, in conjunction with the non-uniformity of coverage, degrade the long wavelength integrity of the gravitational information that can be extracted from surface gravimetry. Nevertheless, surface and airborne gravimetry presently provide the only data that can resolve short wavelength gravity features, especially over land areas. In addition, ship borne gravity measurements aid the separation of the geoid from the DOT signal when used in combination with satellite altimetry.
3. Satellite altimeter data have enabled an unsurpassed mapping of the field over the oceans, both in terms of accuracy and in terms of resolution. TOPEX/Poseidon (T/P) (Fu et al. 1994) (as well as its follow-on missions Jason-1 and Jason-2) routinely provides estimates of the Sea Surface Height (SSH) which, for the first time, are not significantly contaminated by radial orbit error (RMS radial orbit error at the ± 2 cm level). However, altimetric measurements are confined over the ocean areas bounded by the satellite's inclination, and furthermore provide a mapping of the sum of the geoid undulation plus the DOT. These aspects weaken somewhat the contribution of altimeter data in the determination of the long wavelength gravitational field and necessitate the appropriate modeling and estimation of the DOT when altimeter SSH data are used in combination solutions. There is however another way of incorporating altimeter data into a high-degree GGM, which is discussed next.
4. The combination of altimeter data from multiple missions, some of which have produced very closely spaced ground tracks, has provided a dense sampling of most of the ocean's surface. These data, in the form of SSH and/or SSH slopes, can be used to estimate ocean-wide sets of gravity anomalies, at very fine resolution (e.g., $2' \times 2'$ and $1' \times 1'$), as it is discussed in detail in Chap. 9.

Areal averages of these values can be merged with corresponding land and airborne gravity anomalies and gravity anomalies inferred from models of the topographic-isostatic potential (Pavlis and Rapp 1990), to produce a complete global equi-angular grid of gravity anomalies. The geometry of such grids allows the applicability of very efficient harmonic analysis (and synthesis) methods (Rizos 1979; Colombo 1981a), which have revolutionized the development and use of very high-degree spherical harmonic expansions. These approaches allow also efficient combination with *satellite-only* information, as was done, e.g., by Rapp and Pavlis (1990). However, incorporation of altimeter data into a GGM in this fashion requires some a priori knowledge of the DOT (or some other iterative approach – see Sect. 6.6.2.2), so that the altimetry-derived gravity anomalies are estimated from appropriately corrected SSH.

Satellite tracking, altimetric, and surface gravimetric data are of complimentary character both in a spectral as well as in a geographic sense. Their combination enables the determination of the gravitational field, over a wider band of its spectrum, with improved accuracy, than can be obtained by using any of the three data types alone. The particular means of combining these data, in order to develop a high-degree GGM, constitutes a solution strategy. A critical consideration in the design of a solution strategy is the treatment of altimeter data (Rapp 1993), i.e., if these data will be incorporated as in (3) or as in (4) above. OSU91A (Rapp et al. 1991) and EGM96 (Lemoine et al. 1998) represent the result of implementing a particular solution strategy, whereby altimeter data were used as in (3) for the determination of the low degree part of these models (maximum degree 50 and 70 respectively), and as in (4) for the higher degree part. The main disadvantages of this strategy are that: (a) the high degree GGM is obtained in a "piece-wise" fashion and, (b) a complete error covariance matrix exists only for the low degree portion of the model. N.K. Pavlis in (Lemoine et al. 1998, Chap. 8) discussed specific reasons for the selection of that particular estimation strategy. Certain characteristics of the above data types that are particularly important for their effective combination are discussed next.

- ***Information Content.*** The observables within the above four categories contain information not only about the gravitational field, but also about numerous other effects. Some of these effects are of interest in their own right (e.g., the DOT information contained within altimetric SSH), while others represent, at least as far as gravitational modeling is concerned, (more or less) systematic noise (e.g., the non-conservative forces acting on a satellite). In either case, effective incorporation of a particular data type into the combination solution requires precise modeling and optimal estimation of all the effects and signals contained within the observable. Otherwise, the estimated gravitational model can be severely corrupted by the mis-modeled (or un-modeled) systematic effects.
- ***Spectral Sensitivity Overlap.*** The development of a GGM through a least-squares adjustment combining different data types is meaningful, provided that the data used in the adjustment share some common degree of sensitivity to the gravitational signal over a certain portion of its spectrum (a range of

harmonic degrees). Otherwise, there is little "adjustment" being performed to data representing disjoint spectral bands. For example, existing satellite-only models have a narrow spectral sensitivity overlap with models recovered from surface gravity data alone. This complicates considerably the problem of optimal combination of these two data types. On the other hand, this also means that setting up and inverting extremely large linear systems corresponding to very high degree models may not be necessary, if a single data type (e.g., a complete global equi-angular grid of gravity anomalies) uniquely determines the higher degree portion of such a GGM.

- ***Relative Weighting.*** The optimal estimation of a GGM depends critically on the optimality of the relative weights assigned to the different data types. Considering the numerous sources of data that are involved, this is a very large *component of variance* estimation problem, complicated further by the fact that the extraction of gravitational information from satellites' orbit observations is a strongly non-linear problem. Although approximate solutions to this relative weight estimation problem have been used with considerable success (Lerch 1991), many times the experience and intuition of the model developer(s) guide the selection of appropriate data weights more than anything else.

6.4 The New Satellite Missions

The satellite data used for the development of all GGMs published by the end of the twentieth century represent tracking of "targets of opportunity", i.e., of spacecraft designed and equipped with instrumentation for applications other than the mapping of the gravitational field from space. As a result of three satellite missions, this situation has changed dramatically during the last few years. These three missions are CHAMP (Rapp et al. 1996), GRACE (Grace 1998), and GOCE (ESA SP-1233 1999). Table 6.1 summarizes the main characteristics of these missions.

A nice discussion regarding the concepts involved in these three mission scenarios can be found in (ESA SP-1233 1999, Sect. 2.3). Mapping of the gravitational field from space requires missions that adhere as much as possible to the following fundamental design constraints:

- Uninterrupted tracking in three spatial dimensions.
- Measurement or compensation of the effects of non-gravitational forces.
- Orbital altitude as low as possible, to enhance sensitivity to the gravitational signal, and inclination as high as possible, to permit (near) global coverage.
- Counteraction of the field's attenuation at altitude through the measurement of derivatives of the potential.

All three missions above have in common the high-low Satellite-to-Satellite Tracking component (SST-*hl*) from the GPS (and GLONASS in the case of GOCE) constellation, and the measurement of non-gravitational forces by the on-board accelerometers. These data permit highly accurate orbit determination for all three

Table 6.1 Main characteristics of three satellite missions

Mission	Status	Orbit	Mission objective	Instrumentation, tracking, and comments
CHAMP	Launched on 7/15/2000 Active	Alt. = 450 km $e \approx 0.004$ $i = 87°$	Gravity and Magnetic fields Atmospheric Limb Sounding Ionosphere Sounding	3-axis STAR accelerometer GPS and SLR Altitude will decay from 450 km (BOL) to 300 km (EOL)
GRACE	Launched on 3/17/2002 Active	Alt. = 485 km $e \approx 0.001$ $i = 89°$	Gravity field and its temporal variation	3-axis accelerometers (1 per s/c) GPS and SLR K-band inter-satellite ranging between the 2 s/c
GOCE	Launched on 3/17/2009 Active	Alt $\approx$ 250 km $i = 96.7°$ Sun-Synchronous	Gravity field (Especially static)	Six 3-axis accelerometers forming the gradiometer GPS/GLONASS and SLR

missions, and in addition may enhance the gravitational field determination at very long wavelengths (very low degrees). In addition to that, GRACE involves the continuous measurement of the range between two identical satellites that "chase each other", which constitutes a low-low SST formation (SST-*ll*). GOCE's accelerometer array on the other hand provides the measurements necessary to determine the gravitational tensor (i.e., the 3×3 matrix of second order spatial derivatives of the gravitational potential) at altitude. GOCE is unique in the sense that it will provide boundary data at altitude covering the entire Earth, *except* for two polar caps of $\sim 6.7°$ radius (due to the satellite's inclination). The data from each of these three missions result in different sensitivities to the gravitational spectrum. Simulation studies examining the geopotential recovery attainable from these (and other) mission scenarios were reported e.g., by Sneeuw and Ilk (1997). Figure 6.1 depicts the degree amplitude spectra (square root of the degree variance) of the geoid undulation signal and its error as predicted from EGM96S and from CHAMP, GRACE, and GOCE mission simulations. In the same figure an estimate of the degree amplitude spectrum of the DOT and of geoid undulation effects predicted from a postulated model of vertical datum inconsistencies are shown. The latter is just one of several systematic error sources possibly affecting terrestrial gravity anomaly data (Heck 1990), but not necessarily the dominant source of error, as an analysis by Pavlis (1988) indicated.

Two main questions arise when considering the data from these satellite missions:

1. What is the optimal way of analyzing the data from these missions?
2. What is the optimal way of combining their data with existing data, e.g., from surface gravimetry and from satellite altimetry, in order to develop high-degree combination gravitational models?

1. ***Data Analysis.*** In the case of CHAMP the gravitational information is extracted from the analysis of the perturbations of a low Earth orbiter, in a fashion similar to other existing satellite missions. However, CHAMP's low orbit, in conjunction

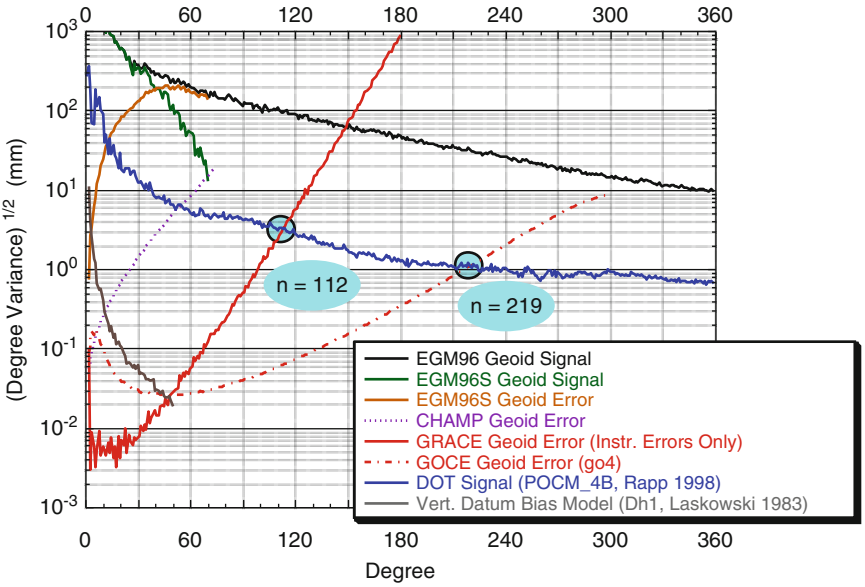


Fig. 6.1 Degree amplitude spectra

with the accelerometer data and with the availability of nearly global tracking data coverage, enabled *for the first time* the determination of an accurate long wavelength global gravitational model from a *single* satellite's data. Indicative of this "new state of affairs" is that a very preliminary solution (complete to degree and order 91) was already developed based on a *single* month's worth of CHAMP data *only* and was presented during the 2001 meeting of the International Association of Geodesy (IAG) by Reigber et al. (2001). Although significantly better models that include GRACE data have by now surpassed considerably this preliminary solution, it served as a good example of the improvements that were to follow.

Compared to CHAMP, GRACE added the SST-*ll* component, which permitted higher resolution gravitational information to be extracted from the analysis of the orbital perturbation *differences* along the line-of-sight of the two low orbiting satellites. One can use traditional orbit perturbation analysis methods to process the GRACE data and derive a GGM, e.g., in spherical harmonics. GGM01S (Tapley et al. 2004) and GGM02S (Tapley et al. 2005) were estimated following such a procedure. This analysis scenario, albeit costly, is within current computational capabilities for models extending to degree and order 180 or so. Geodesists have also considered alternative analysis methods for GRACE-type missions (e.g., Wolff 1969; Colombo 1981b; Jekeli 1999a; Rowland et al. 2002). Such methods provide higher computational efficiency at the cost of committing certain approximations. Luthcke et al. (2006) reported monthly gravitational solutions determined from GRACE inter-satellite range-rate data *alone*, which

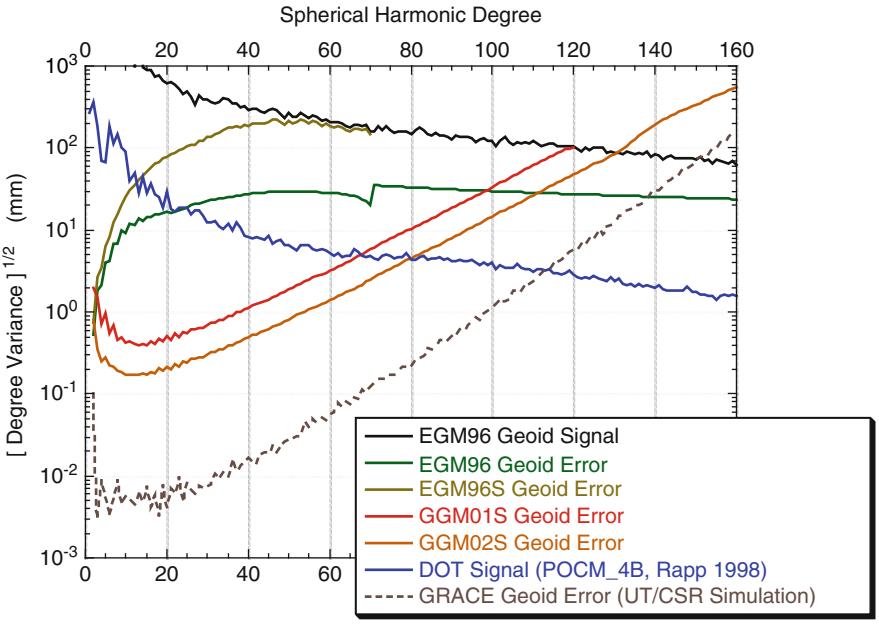


Fig. 6.2 Degree amplitude spectra comparing actual (GGM01S, GGM02S) versus simulated GRACE performance

are significantly less affected by certain systematic errors ("stripping") than those solutions that incorporate simultaneously the GPS data (SST-*hl*) in their development. Setting aside for a moment the details of optimal GRACE data processing, it is important to recognize here the quantum leap that has been accomplished with the GRACE mission. Approximately 14 months of GRACE data *alone* have been used to develop GGM02S, whose cumulative geoid undulation error to degree 70 is less than ± 1 cm (Tapley et al. 2005, Fig. 6.2). By comparison, the cumulative geoid undulation error to degree 70 for EGM96, which required the combination of data from tens of satellites, along with surface gravity and satellite altimetry, was ± 19 cm (Lemoine et al. 1998, Table 10.3.2-1).

As far as GOCE is concerned, numerous investigations of the various aspects of its data analysis and of the development of a GGM from them can be found in ESA (2000). One particular issue that receives increased attention relates to the polar gaps and their impact on analysis schemes that exploit regularity and completeness in the data coverage (e.g., block-diagonal normal equation formation schemes).

2. ***Data Combination.*** While the existing satellite gravity-mapping missions (CHAMP and especially GRACE) have already delivered (or promise to deliver as in the case of GOCE) quantum leaps in the accuracy and resolution of the *satellite-only* gravitational models (see Fig. 6.1), there is still a need to combine that information with terrestrial gravity and satellite altimetry data in an

optimal fashion. This is required so that a seamless extension of the gravitational spectrum can be achieved, taking advantage of the rich high frequency content of the surface and altimetric data. The higher resolution of GRACE- and GOCE-based *satellite-only* models will significantly increase the spectral overlap with surface gravity and altimetry. This will enable for the first time the estimation of high-resolution models of the DOT and of the significant reduction of systematic errors present in surface gravity data. Whether these estimates will be obtained within *comprehensive* combination solutions (i.e., solutions where gravitational potential coefficients are estimated *simultaneously* with several other parameter sets), or using some other approach is (to some extent) still under investigation. Comprehensive solutions have the advantage that they provide a complete error covariance matrix associated with *all* estimated parameters. They are however computationally demanding, as we elaborate in the next paragraph.

Let us focus for a moment on the parameter sets corresponding to potential coefficients, DOT, and systematic surface gravity errors. Let us also assume for the sake of this argument that all three sets will be represented using spherical harmonic coefficient sets complete to degrees M (potential), L (DOT), and K (surf. gravity errors), respectively. One can deduce approximate values for M and L (e.g., from Fig. 6.1), but not for K. The spectral (and geographic) behavior of the surface gravity systematic errors and their maximum resolvable degree will have to be estimated *directly*, e.g., from a preliminary comparison of surface gravimetry with the information implied by GRACE and GOCE. In contrast to the DOT, for which spectral and geographic estimates exist e.g., from the analysis of oceanographic models (see Fig. 6.1), and can be used to guide the selection of L, only very limited information exists regarding the spectral and geographic behavior of surface gravity errors (Pavlis 2000). There are good reasons to try to estimate these surface gravity systematic errors optimally, and also to try to explain their origin. Such estimates may reveal not only problems associated with the gravity anomaly (and/or elevation) data themselves, but also possibly problems related to the (pre-) processing and modeling of gravity anomaly data (e.g., analytical continuation). Furthermore, direct estimation of these effects may help resolve some outstanding questions related to the weighting of surface gravity relative to the satellite information, and should shed new light on the spectral distribution of surface gravity errors (at least to the degree resolvable by the satellite information). To illustrate the challenges implied by the new satellite missions, we consider here the development of a (*hypothetical*) GGM, following a procedure similar to that used for the development of the degree 70 part of EGM96 (see Lemoine et al. 1998, Chap. 7). With spherical harmonics as the representational basis for the three parameter sets discussed above, each of the new satellite missions implies certain values for M, L, and K. Using Fig. 6.1 as our guide (at least for M and L), we conclude that:

- This combination scenario can be readily implemented in the case of CHAMP.
- In the case of GRACE, such a scenario implies $M \approx 140$ (potential coefficients), $L \approx 110$ (DOT coefficients), and a K value that will probably not exceed $\sim$90

(surf. gravity error coefficients). The total number of parameters to be estimated in this case will be $O(40\,\mathrm{k})$. Despite its size, this problem is well within our current computational capabilities, as Pavlis and Kenyon (2003) have already demonstrated. One should also notice that using the error curves of Fig. 6.1 to estimate M, L, and K may result in values that are larger than those implied by the *actual* (as opposed to the *simulated*) performance of these missions. For the case of GRACE, Fig. 6.2 allows a comparison between *actual* (GGM01S, GGM02S) and *simulated* performance.

- Treating the GOCE case however in the same fashion could result in parameter sets that will be approximately four times as large as the set for GRACE. Combination of GOCE data with surface gravity and altimetry, in a comprehensive least-squares adjustment fashion, may require the development of new innovative approaches that economize on number of parameters and/or result in patterned normal equation systems that adhere to efficient formation and inversion algorithms. Whatever these techniques might be, they should allow also the efficient computation of the error variance and covariance of the various recovered fields (and of their functionals), as a function of geographic location, as well as in a spectral form.

6.5 Beyond the Sensitivity of Satellite Data

There is obviously a limit to the gravitational information that can be extracted from space-borne sensors, which implies a limit to the resolution (maximum degree) of satellite-only models. Observations made on the Earth's surface (or on airplanes flying at low altitudes) can extend the resolution of gravitational models considerably. Surface and airborne data like gravity anomalies, gravity disturbances, etc. are therefore capable of supporting the development of much higher resolution gravitational models, than those developed based on satellite data only. One way of developing such high degree models (the way used to develop the degree 71–360 part of EGM96) involves first the formation of a regular grid of areally averaged values of some functional of the field that covers completely the globe. These values represent data derived on the basis of other primary observables, using techniques like Least Squares Collocation (LSC). Surface gravity anomalies are a suitable choice for such a functional, both from a spectral sensitivity and from an availability viewpoint. Very high degree and order spherical harmonic models are then developed from the analysis of such global anomaly grids, using efficient harmonic analysis techniques like those put forward by Colombo (1981a). Wenzel (1998, 1999) reported such expansions complete to degree and order 1,800. Pavlis et al. (2005) reported the development and evaluation of the PGM2004A preliminary gravitational model, extending to maximum degree and order 2,160, and identified some aspects related to its development that required further improvement. Computational efficiency was *not* one of them. An expansion to degree 2,160, given a complete $5' \times 5'$ grid of area-mean gravity anomaly values required

approximately 30 min to execute on a Sun Fire v480 workstation with four 6.2 GHz Ultra SPARC III processors, and this time includes the computation of the values of the integrals of Associated Legendre functions. Expansions of this kind can augment models developed via comprehensive combination adjustments, thereby defining "composite" high degree GGMs (like EGM96). Certain aspects of such "composite" models may be criticized. These aspects are discussed next.

1. ***Piece-wise nature.*** The fact that "composite" models (as EGM96) are not developed via a single combination solution least-squares adjustment is considered a drawback of this approach. But is it really necessary to strive for such a single step adjustment? One should recognize that beyond the maximum degree resolvable from satellite sensor data, the gravitational information is *uniquely* determined from surface data (including altimetry-derived gravity anomalies). This implies (as mentioned before) that no "adjustment" is really being performed over this high degree spectral band; therefore a single step approach may be more of a nicety rather than a necessity. What is necessary though is that the transition from the low to the high degree spectral band is seamless in terms of both signal and error. In essence, this piece-wise approach, with spherical (or ellipsoidal) harmonics as the representational basis, may be viewed as a (limiting) case of a "remove-compute-restore" gravimetric approximation that is performed globally and in the spectral domain. In this case, the "remove" step corresponds to the low-pass filtering of surface gravity and satellite altimetry data using a preliminary high-degree expansion, which is done to minimize aliasing effects (see Pavlis 1988, Sects. 5.2.4, 5.2.5); the "compute" step refers to the (relatively) low degree part of the field that is developed through the comprehensive combination solution; and finally the "restore" step corresponds to the augmentation of the low degree combination solution with the high-degree expansion coefficients.
2. ***Lack of complete error covariance matrix.*** This is a critical shortcoming of the currently available high resolution GGMs. Using spherical harmonics as the representational basis implies that in order to obtain propagated error estimates, *with geographic specificity*, for derived functionals of the field, one has to form and propagate complete error covariance matrices corresponding to the maximum degree of the model. Clearly this is a very computationally demanding proposition for existing models (to degree and order 360), let alone ultra high-degree expansions like EGM2008 (Pavlis et al. 2008) that will be discussed in more detail in Sect. 6.6.2. A model complete to degree and order 2,160 involves ∼4.7 million coefficients. This would be the dimension of the (symmetric) error covariance matrix that needs to be formed, to allow conventional error propagation. A much more efficient solution to this problem has been developed and presented initially by Pavlis (2005). This technique recognizes again that beyond a certain degree M (corresponding to the resolution of the *satellite-only* model), the GGM is uniquely determined from surface gravity (terrestrial, airborne, and altimetry-derived) data. This implies that complete error covariance matrix propagation may *only* be necessary for the portion of the model up to degree M. Using band-limited kernels (beyond degree M and up to the

maximum degree of the model) within integral formulas, one is able to compute the error contribution of the harmonics beyond M in a much more efficient manner, through global convolutions. The propagated error components for the different spectral bands are subsequently added (in a quadratic sense), assuming no data error correlation across the different spectral bands.

3. ***Data pre-processing requirements.*** The development of high degree GGMs using the procedures discussed previously does require a considerable effort for the pre-processing of available surface (land, marine, and airborne) data, as well as for the prediction of altimetry-derived gravity anomalies. The harmonic analysis approaches developed by Colombo (1981a) are best suited for the analysis of complete grids of a *single* data type, referring to a surface of revolution (e.g., a rotational ellipsoid). Furthermore, the efficiency of such estimators is based in part on rather strong assumptions concerning the signal and error covariance functions of the data (homogeneity and isotropy). Since the actual measurements do not comply with such configurations in general, several pre-processing steps are required to transform the primary observables to quantities that adhere to the requirements of the estimator (at least to a certain degree of approximation). LSC prediction of area-mean values of gravity anomalies from (the combination of) measurements acquired on land, sea, and air, aims to produce a *single* data type out of the several types of (possibly overlapping) measurements that may be available over a given area. The same technique is used to derive gravity anomalies from dense sets of altimetric SSH data. Analytical continuation aims to artificially reduce these gravity anomalies to quantities that refer to a surface of revolution (e.g., the reference ellipsoid). These artificial quantities, when analytically continued in the opposite direction, are supposed to reproduce the input gravity anomalies. The prediction of area-mean gravity anomalies on a regular grid, and their analytical continuation, produce derived data that adhere to the *geometric* requirements of the estimator. LSC, which could be used to derive a GGM without much need for pre-processing of the *original* data, requires the formation and inversion of a matrix whose size equals the number of observations. This is an impossible task for the foreseeable future. *Efficient* techniques that can make use of the *original* data with minimal pre-processing requirements are (still) desirable.

The treatment of the *stochastic* properties of the data is even more complex than the treatment of *geometric* requirements. Availability of error variance estimates is pretty much the best that an analyst can hope for, and most times these estimates reflect data precision rather than data accuracy. The majority of gravimetric approximation studies (both global and local) either neglects completely any error correlation between the data, or attempt to account for it in empirical (many times not well justified) ways. It seems reasonable (at least to this author) to consider the error of currently available area-mean surface gravity anomaly data as composed of two main components: (a) a long wavelength component originating from systematic errors e.g., in the base network, the "ties" to it, long wavelength errors in elevations, etc., and, (b) a short wavelength component reflecting things

like the accuracy and density of local data, the local roughness of the field and of the topography, etc. The former component is expected to have relatively low standard deviation (order of ± 2 mGal) but very long correlation lengths (continental or even global scale); the latter may have standard deviations that in certain regions exceed ± 30 mGal, but its correlation lengths are expected to be short (a few tens of km). The best way to account for the long wavelength component of these errors is probably by direct estimation of systematic errors in combination solutions with *satellite-only* models from missions like GRACE and GOCE. How to treat the short wavelength component with rigor and efficiency, both in the estimation of a GGM and in the subsequent error propagation, remains still an open question.

6.6 State-of-the-Art Global Gravitational Modeling

In this section we discuss the main aspects of the development of two global gravitational models, representative of the state-of-the-art at the respective time of their development: EGM96 (Lemoine et al. 1998), which represents the state-of-the-art before the availability of data from CHAMP and GRACE, and EGM2008 (Pavlis et al. 2008; 2012), which represents currently (2010) the model with the highest resolution and accuracy, prior to the anticipated availability of data from GOCE. The choice of these two models also permits a comparison of the approaches followed in their development. Such a comparison reveals the critical changes in model development, which were brought about by the availability of GRACE data on one hand, and of high quality $5' \times 5'$ area-mean gravity anomalies (from the combination of terrestrial and altimetry-derived data sources) on the other.

In theory, the estimation of a combination solution complete to some (arbitrary) high degree and order could be carried out as follows:

(a) Form separate normal equations from each individual data type, to a maximum degree and order that corresponds to the resolution of the available data and their sensitivity to the gravitational signal.
(b) Treat satellite altimeter data as "direct tracking" observations, i.e., ranges from the spacecraft to the ocean surface whose upper endpoint senses (through the orbit dynamics) attenuated gravitational signals (static and time-varying), while their lower endpoint senses the combined effects of geoid undulation, DOT as well as tides and other time varying effects, without any attenuation. In this manner, altimeter data contribute to the estimation of the satellite's orbit, as well as the estimation of the DOT and of the potential coefficients.
(c) Combine the various normal equations (with appropriate relative weights) and invert the resulting system, to estimate the combination solution to its high degree, along with its full error covariance matrix.

Such an "ideal" approach would permit the most rigorous modeling of the observables and would allow the greatest flexibility in terms of data weighting. A combination solution to degree 360, if performed as outlined above, would require

the formation of full (symmetric) normal matrices (from satellite altimetry and surface gravimetry) for approximately 130,000 parameters (considering *only* the gravitational potential coefficients). For maximum degree 2,160, there would be approximately 4.7 million such parameters involved. Such computational tasks are beyond our present computational capabilities. Therefore, at present, one may choose between, or combine, two main solution strategies to attack the problem:

- ***Solution Strategy (A)*** Apply the "ideal" estimation strategy outlined above, to obtain a combination solution for the lower degree part of the field, up to a maximum degree that is computationally manageable. Apart from reasons of computational feasibility, this maximum degree should enable the appropriate modeling of the gravitational signal contained in the currently available satellite tracking data. Furthermore, since the DOT signal is of long wavelength nature, the benefits of "direct" altimetry are almost entirely retained here. To avoid aliasing effects however, the contribution to the altimetry and surface gravity data from the coefficients beyond the solved-for degree has to be filtered out of the data prior to the normal equation formation. This may be done using a pre-existing high-degree solution (Pavlis 1988; Denker and Rapp 1990). Hereon, we will refer to this type of solution as the *comprehensive* low degree combination model. The obvious shortcomings of this approach are the relatively low maximum attainable degree (approximately 200 at present) and its computational demands. Some models developed using this approach (or similar ones) include JGM-1 and JGM-2 (Nerem et al. 1994) and JGM-3 (Tapley et al. 1996), the part of EGM96 up to degree and order 70 (see Lemoine et al. 1998, Chap. 7), and EIGEN-GL04C (Förste et al. 2008).
- ***Solution Strategy (B)*** Consider that one is willing to make the following two approximations:

  (i) The orbits of the altimeter satellites, whose data are included in the combination solution, are perfectly known (at least radially). This approximation is justifiable if one is working with altimeter satellites supported by T/P-class precise orbit determination. Moreover, after the availability of GRACE-based gravitational models for precise orbit determination of altimeter satellites, errors arising from gravitational model inaccuracies do not dominate the orbit error budget of these satellites. Errors due to, e.g., mis-modeling of non-conservative forces acting on the spacecraft are likely to be more significant nowadays. In this regard, to allow the orbits of altimeter satellites to contribute (through their dynamics) to the determination of gravitational parameters within a combination solution may not be a desirable approach nowadays, because the effects of orbit errors of non-gravitational origin could corrupt the solved-for gravitational parameters.

  (ii) The DOT is known a priori, e.g., from an Ocean Circulation Model (OCM) or from a previous low degree comprehensive combination solution.

The implication of (i) is that satellite altimetry does not have to be treated as "direct" tracking anymore, which simplifies the problem considerably, since now

orbit dynamics are not involved in the altimeter data processing. One is left with a "surface" problem, where the geoid (N) and the DOT (ζ) signals have to be separated, given the "observed" SSH (h), which is their sum. If in addition, the approximation (ii) is introduced, then altimetry contributes to the combination solution "observed" geoid heights (N) over (parts of) the ocean.

In addition to the above two approximations, a key issue here is that altimetric information may also be provided in a *gridded* form. This is possible through the use of a Mean Sea Surface (MSS), obtained from multiple altimetric missions. The success of T/P has significantly improved the accuracy of MSS data sets (especially at long wavelengths). This is accomplished by adjusting the SSH data from other altimetric missions (e.g., ERS-1, ERS-2, GEOSAT, SEASAT), to the surface defined by T/P, using cross-over minimization techniques. Such MSS data sets have been developed by, e.g., Yi (1995), Kim et al. (1995), Anzenhofer et al. (1996) and Wang (2001), and more recently by Andersen et al. (see Chap. 9 for details). One may also have available gridded, altimetry-derived gravity anomaly values.

Such values have been estimated using various techniques, on an ocean-wide basis by, e.g., Rapp and Basic (1992), Andersen and Knudsen (1998), Trimmer and Manning (1996) and Sandwell and Smith (1997, 2009) among others.

The two simplifying approximations discussed above and particularly the availability of altimetric information in gridded form (especially in the form of gravity anomalies), make applicable an alternative class of high-degree combination solution techniques. These, combine the satellite-only information, with potential coefficient information obtained from the analysis of *complete*, regular grids of functional(s) of the disturbing potential (e.g., N, Δg), and are based on the highly efficient harmonic analysis algorithms originally studied and put forward by Colombo (1981a). These algorithms exploit the regularity of the data grids and the symmetry properties of Legendre and trigonometric (sine/cosine) functions. Using Fast Fourier Transform (FFT) techniques, one may process data arrays residing over latitude bands that are symmetric with respect to the equator, in a highly efficient manner. Estimators of this type are the (simple) Numerical Quadrature (NQ), the Block-Diagonal (BD) least-squares adjustment, and the Optimal Estimation (OE) technique. Models developed using the NQ approach include OSU86E/F (Rapp and Cruz 1986a) and OSU89A/B (Rapp and Pavlis 1990). BD techniques of varying sophistication have been used to develop GPM2 (Wenzel 1985), DGFI92A (Gruber and Bosch 1992), GFZ95A (Gruber et al. 1996), and EGM2008 (Pavlis et al. 2008). OE was used to develop the OSU86C/D models (Rapp and Cruz 1986b).

In the following sections we discuss in some detail the development approaches used for EGM96 and EGM2008 respectively.

6.6.1 EGM96

The two solution strategies (A) and (B) discussed above have their respective advantages and disadvantages. EGM96 (Lemoine et al. 1998) employed a comprehensive

solution to degree 70, augmented by a BD solution from degree $n = 71$–359, while the $n = 360$ coefficients were obtained from a NQ model. In the following sections we describe in some detail the "building blocks" that were used to form the EGM96 high-degree model. Although EGM96 has by now been surpassed in terms of performance by more recent models like EGM2008, its development strategy still serves as a didactic example of the particular techniques that were used to model and combine optimally the data that were available at the time of its development.

6.6.1.1 The EGM96S Satellite-Only Model

The estimation of potential coefficients from satellite tracking data is a non-linear problem that involves the simultaneous estimation of the orbit, tracking station coordinates, tide parameters, polar motion and Earth rotation parameters as well as numerous nuisance parameters which may be measurement type or satellite specific (e.g., measurement biases and drifts, atmospheric drag and solar radiation pressure scale factors, etc.). The problem is further complicated by the fact that each satellite samples the gravitational field effects in a particular manner, dictated by its orbital characteristics (altitude, inclination, eccentricity) and the type of tracking data (e.g., Doppler versus ranges). Empirical acceleration parameters that may be necessary to estimate accurate orbits, many times absorb useful gravitational signal as well, so the analyst has to make appropriate trade-offs with extreme care, to ensure an optimum solution. A satellite-only solution involves the processing of tracking data segmented initially by "arcs" of various time spans depending on the satellite and the tracking data type. Once the estimation of the initial state parameters for an orbital arc has converged, normal equations for all the parameters (arc-specific and common) are formed. EGM96S involved the formation of approximately 2,000 such normal equation sets. These were subsequently combined by satellite and/or measurement type, while arc-specific parameters were successively eliminated through back-substitutions. Thus, one was left with "combined" normal equations, which now involved only the parameters common to all satellites and all data types. In EGM96S, this process resulted in approximately 40 sets of "combined" normal equations, which involved $\sim$12,300 parameters. Addition of these normal equations (appropriately weighted) resulted in a single, final set of normal equations. Its inversion defined the satellite-only model and its associated error covariance matrix.

The most critical aspect in this combination of normal equations is the weight assigned to each set of them. In a relative sense, weights should be such that the solution does not "over-fit" any particular satellite/data type at the expense of the others. In an absolute sense, they should yield an a posteriori error covariance matrix, which would accurately reflect the quality of the model. To "calibrate" the weights one may use the subset solution technique of Lerch (1991). One data type (or satellite) at a time is withheld from the solution, and the changes of the potential coefficients are compared to the changes predicted by the corresponding formal error estimates (complete versus subset solution). Weight calibration is a time consuming, iterative task and requires one to start with a preliminary set of

weights, which should be close enough to the optimal set, to ensure convergence and minimize the number of iterations needed to achieve it. The experience of the analyst is indispensable here. This subset solution technique ensures primarily the internal consistency of the solution. Comparisons with external data, independent from the satellite-only model, provide the best means to test the reliability of the propagated error estimates of the model, in an absolute sense.

Particularly valuable to the development of EGM96S were SST data (high-low mode) from the GPS satellites to T/P, EUVE and GPS/Met, as well as TDRSS tracking of EUVE. These data provide continuous, precise tracking of the low orbiter and are more sensitive to high frequency geopotential effects, than traditional (pre-CHAMP) tracking data types. Calibration of the weights of these data proved to be a particularly challenging task.

The development of an accurate and well-calibrated satellite-only model is the most critical (and arguably the most complicated) part of the combination model development. Satellite-only models that were developed before the dedicated gravity-mapping missions (CHAMP, GRACE, and GOCE) include EGM96S (Lemoine et al. 1998) and GRIM4-S4 (Schwintzer et al. 1997).

It is important to recognize here that the normal equation matrices associated with these pre-GRACE satellite-only models were fully occupied, and rather ill-conditioned due to high correlations present among the coefficients of higher degrees. The inversion of these matrices usually required the use of some a priori constraint in the form of a power law (e.g., Kaula's rule). In order to preserve the integrity of the least-squares adjustment used to derive the combination solution, one had to consider the satellite-only normal equations in their complete (fully-occupied) form. Any block-diagonal approximation of these normal equations would result in estimation errors that could not be tolerated. This situation changed dramatically with GRACE, due to its global coverage and uniform data accuracy, which simplified the development of combination solutions dramatically, as we will see when we discuss the development of EGM2008.

6.6.1.2 The EGM96 Comprehensive Low-Degree Combination Solution

This solution involves the combination of the final satellite-only normal equations with normal equations developed from terrestrial gravity data and from satellite-altimeter data treated as "direct" tracking. Since altimetry enters here as direct tracking, and since the surface gravimetric data are introduced as a totally independent data type (i.e., no error correlation between the surface gravity, altimetry and satellite tracking data is considered), the surface gravity normal equations have to be developed based on gravimetric information independent of both the tracking and of the altimeter data. This requires the exclusion of any altimeter-derived anomalies from the file used to develop the surface gravity normal equations. The requirement for independence from the tracking data is slightly violated because of the way that "fill-in" anomalies are computed. In the following sections we describe the development of the surface gravity and altimetry normal equations.

The Surface Gravity (Low Degree) Normal Equations

The gravity potential of the Earth, W, is defined to be the sum of the gravitational potential, V, given in (6.1), plus the centrifugal potential Φ arising due to the rotation of the Earth. Consider the gravity potential U of a rotating equipotential ellipsoid of revolution (*Somigliana-Pizzetti* normal field). The disturbing potential $T(r_P, \theta_P, \lambda_P)$ is defined as (Heiskanen and Moritz 1967, Eq. 2-137):

$$T(r_P, \theta_P, \lambda_P) = W(r_P, \theta_P, \lambda_P) - U(r_P, \theta_P). \tag{6.3}$$

Due to (6.1) we have:

$$T(r_P, \theta_P, \lambda_P) = \frac{GM}{r_P} \sum_{n=2}^{\infty} \left(\frac{a}{r_P}\right)^n \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta_P, \lambda_P), \tag{6.4}$$

where we have assumed that the ellipsoid has the same mass and rotational speed as the actual Earth, and is centered at the Earth's center of mass. In (6.4), C_{nm} are now the remainders of the coefficients appearing in (6.1), after subtraction of the even degree zonal harmonic coefficients of the normal gravitational potential. Consider a quantity Δg^c that fulfills:

$$\Delta g^c = -\left(\frac{\partial T}{\partial r}\right)_Q - \frac{2}{r_Q} T, \tag{6.5}$$

where Q is a point on the *telluroid* (Heiskanen and Moritz 1967, p. 292). Substitution of (6.4) into (6.5) yields:

$$\Delta g^c(r_Q, \theta_Q, \lambda_Q) = \frac{GM}{r_Q^2} \sum_{n=2}^{\infty} (n-1) \left(\frac{a}{r_Q}\right)^n \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta_Q, \lambda_Q). \tag{6.6}$$

The quantity Δg^c is related to the *Molodensky* surface free-air gravity anomaly $\Delta g \quad (= |\vec{g}_P| - |\vec{\gamma}_Q|)$, obtained from scalar gravimetry, by:

$$\Delta g^c = \Delta g - \left(\varepsilon_h + \varepsilon_\gamma + \varepsilon_P\right)_Q . \tag{6.7}$$

$(\varepsilon_h + \varepsilon_\gamma + \varepsilon_P)_Q$ are ellipsoidal corrections (Pavlis 1988). These, along with atmospheric and other corrections are applied to the observed gravity anomalies beforehand. Equation 6.6 refers to point values. Gravitational model estimation currently employs area-mean values over equi-angular cells, although (as Jekeli 1996 has pointed out), the use of area-mean values defined over spherical caps may be a preferable approach. In addition, the gravity anomaly may be analytically continued from the telluroid to the reference ellipsoid. Analytical continuation may be done

using a Taylor expansion approach, whereby the anomaly on the ellipsoid Δg^e is related to Δg^c by:

$$\Delta g^e = \Delta g^c - \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k \Delta g^e}{\partial h^k} \cdot h^k. \tag{6.8}$$

If the Taylor series in (6.8) is truncated to the linear term only, one obtains the (linear) gradient solution to the downward continuation problem:

$$\Delta g^e \approx \Delta g^c - (\partial \Delta g^e / \partial h) \cdot H^*. \tag{6.9}$$

The normal height, H^*, is seldom available in practice. It is usually approximated by the orthometric height H. This approximation however, introduces non-negligible systematic errors in the analysis of surface gravimetric data (Pavlis 1988). The free-air gravity anomaly gradient $(\partial \Delta g^e / \partial h)$ may be evaluated using detailed elevation information (assuming linear correlation between the free-air anomaly and the elevation), or from a preliminary high-degree model (from which one may also compute higher-order terms in the Taylor series of (6.8), in an iterative fashion). Downward continuation can also be performed by the iterative numerical solution of Poisson's integral (see Wang 1987, 1988 for more details). These reductions, and the discretization of the area-mean values over equi-angular cells, produce a grid of values on the ellipsoid, which may be modeled using solid spherical harmonics as:

$$\overline{\Delta g}^e_{ij} = \frac{1}{\Delta \sigma_i} \frac{GM}{(r_i^e)^2} \sum_{n=2}^{M} (n-1) \left(\frac{a}{r_i^e} \right)^n \sum_{m=-n}^{n} C_{nm} \cdot I Y_{nm}^{ij}, \tag{6.10}$$

where the subscripts (i, j) identify the location of the cell in a two-dimensional array, defined by parallels and meridians, covering the ellipsoid. r_i^e is the geocentric distance to the center of the (i, j)th cell, and:

$$\Delta \sigma_i = \Delta \lambda \int_{\theta_i}^{\theta_i + 1} \sin \theta d\theta = \Delta \lambda \cdot (\cos \theta_i - \cos \theta_{i+1}), \tag{6.11}$$

$$I Y_{nm}^{ij} = \int_{\theta_i}^{\theta_i + 1} \overline{P}_{n|m|}(\cos \theta) \sin \theta d\theta \cdot \int_{\lambda_j}^{\lambda_j + 1} \begin{Bmatrix} \cos m\lambda \\ \sin |m| \lambda \end{Bmatrix}^{d\lambda} \begin{matrix} \text{if } m \geq 0 \\ \text{if } m < 0. \end{matrix} \tag{6.12}$$

Equation 6.10 represents a mean value whose frequency content is restricted to maximum degree and order M. However, real data (e.g., the $1^\circ \times 1^\circ$ mean values that were used in EGM96) are not band limited. To reduce aliasing effects, one may remove from the equi-angular mean values, the contribution beyond the solved-for degree M, using a preliminary high-degree model complete to some degree $N_{\max}$ (obviously one needs $N_{\max} \gg M$). Schematically:

$$\overline{\Delta g}^e (n = 2 \to M) \approx \overline{\Delta g}^e (n = 2 \to \infty) - \overline{\Delta g}^e (n = M + 1 \to N_{\max}). \tag{6.13}$$

Equation 6.10 is the mathematical model that underlies the observation (and subsequently the normal) equations formed from terrestrial ($1^\circ \times 1^\circ$) gravity data. Details on the normal equation formation can be found in Pavlis (1988). The weight assigned to each individual gravity anomaly should be considered carefully, so that the terrestrial-only solution does not over-fit the areas covered with the most accurate and dense gravity data. Note also that (6.10) assumes that the entire signal present in terrestrial gravity data is of gravitational origin, i.e., (6.10) does not account for any systematic errors that may be present in (near) global anomaly databases.

The "Direct" Altimetry Normal Equations

The altimeter range measurement, ρ_{alt}, can be modeled as Marsh et al. (1990):

$$\rho_{alt} = h_{sat} - (h + \Delta h + \varepsilon) + b, \tag{6.14}$$

where ρ_{alt} is the observed range (corrected for instrument offsets) from the instantaneous sea surface to the satellite's center of mass, and:

h_{sat} is the distance from the surface of the reference ellipsoid to the satellite's center of mass,

h is the instantaneous sea surface height above (or below) the reference ellipsoid,

Δh is the sum of various instrumental, environmental and geophysical corrections,

ε is the instrument noise, and,

b is a bias term arising from the constant and time-varying instrument bias.

The instantaneous sea surface height can be modeled as:

$$h = N(n = 2 \rightarrow M) + N(n = M + 1 \rightarrow \infty) + \zeta_t + t_b + t_o + h_a, \tag{6.15}$$

where:

$N(n = 2 \rightarrow M)$ is the geoid undulation contribution up to degree and order M,

$N(n = M + 1 \rightarrow \infty)$ is the geoid undulation contribution beyond degree and order M,

ζ_t is the instantaneous Dynamic Ocean Topography,

t_b, t_o are the solid-Earth and ocean tides, respectively, and,

h_a is the ocean's response to the atmospheric loading.

Notice that the absence of zero degree from the geoid undulation implies that the bias term b also contains the difference between the semi-major axis of the adopted reference ellipsoid, and that of an *ideal* mean-Earth ellipsoid, with respect to which the geoid undulation averages globally to zero. Combination of (6.14) and (6.15), yields an observation equation which relates ρ_{alt} to the quantities of interest. h_{sat} depends on the potential coefficients, as well as on the initial state vector, and possibly other parameters that influence the orbit dynamics. Therefore, starting from some approximate values, altimeter data can be used to differentially

improve the satellite's orbit (primarily in the radial direction), the low degree part of the potential coefficients (the $N(n = 2 \rightarrow M)$ part) and to estimate the DOT (ζ_t). The contribution $N(n = M + 1 \rightarrow \infty)$ can be filtered out (approximately) from the altimeter data using again a preliminary high-degree model, in a fashion similar to what was described above for the gravity anomalies. Δh is usually provided along with the data (Geophysical Data Records) or is computed from suitable models. h_a may be approximated as an "inverted barometer" response, while b, t_b and t_o may be included in the differential correction (estimation) process. The differential improvement of orbital parameters can be performed using numerical integration, in the exact same manner as is done for other tracking data types (Marsh et al. 1990; Nerem et al. 1994). Alternatively, linear perturbation theory may be used, to improve the radial orbit accuracy (Denker and Rapp 1990; Rapp et al. 1991).

The DOT, ζ_t, is composed of time-invariant and time-dependent parts, i.e.:

$$\zeta_t = \bar{\zeta} + \zeta(t), \tag{6.16}$$

where $\bar{\zeta}$ represents the mean value over some time interval and $\zeta(t)$ may contain annual, semi-annual and seasonal periodic constituents, as well as quasi-periodic effects (e.g., El Niño or La Niña effects). When data from non-contemporaneous missions are analyzed, more than one set of $\bar{\zeta}$-related parameters may be necessary. $\bar{\zeta}$ and $\zeta(t)$ may be represented in terms of surface spherical harmonics. In this case, the data gap generated by the presence of land areas (where $\bar{\zeta}$ and $\zeta(t)$ are not defined) requires some special consideration, in order to prevent large oscillations from occurring over land areas in the recovered $\bar{\zeta}$ and $\zeta(t)$ fields. Such oscillations may be avoided using, e.g., some a priori constraint, or some appropriately selected fictitious values over land, or by employing some alternative representation for these fields. Alternative representations for these ocean-specific signals have been proposed and studied by Hwang (1991) and Sanchez et al. (1997).

The normal equations from the satellite tracking data, the surface gravity data and the "direct" altimetry, can now be combined to estimate the low-degree comprehensive combination solution and its associated error covariance matrix. In the case of EGM96 this solution extended to degree and order 70. The maximum degree of the $\bar{\zeta}$ and $\zeta(t)$ representations which can be resolved, depends primarily on the accuracy of the satellite-only model and of the available marine gravity data. $\bar{\zeta}$ models to degree 20 and $\zeta(t)$ representations to degree 10 for annual and semi-annual constituents were estimated in EGM96. The relative weighting of the three sources of gravitational information (satellite, surface gravity, and altimetry) are again critical, especially when one employs long time spans of altimeter data acquired over repeat ground tracks. Examples of comprehensive low-degree models include EGM96 (to degree 70), TEG-3 (Tapley et al. 1997) and GRIM4-C4 (Schwintzer et al. 1997). The last model however, did not incorporate altimetry as "direct" tracking; it combined the GRIM4-S4 normal equations with normal equations obtained from the analysis of a global $1^\circ \times 1^\circ$ grid of mixed Δg (mostly over land) and N values obtained from altimetry, where the Levitus (1982) model was used to define a priori the DOT.

6.6.1.3 The High-Degree Combination Solution

As discussed in Sect. 6.6, a combination solution to high-degree (e.g., 360 or higher) may be performed by combining the satellite-only normal equations, with gravitational information obtained from the analysis of *complete* global grids of functionals of the field (e.g., N and/or Δg observations). These approaches rely on the exploitation of symmetry properties of the data grids, and take advantage of the applicability of FFT algorithms for the efficient formation of normal equations from the gridded data. Two of the techniques originally studied by Colombo (1981a), the (simple) Numerical Quadrature and the Block-Diagonal least-squares adjustment, were applied during the development of EGM96. These techniques combined the satellite-only information, with a global $30' \times 30'$ "merged" file of mean Δg (appropriately corrected and reduced to the ellipsoid). Other groups (e.g., GFZ) have variously incorporated global sets of *both* N and Δg data simultaneously into the high-degree combination solution. A disadvantage of their approach is that extensive areas have to be "filled-in" with synthetic pseudo-observations particularly of N, but also of Δg, to achieve global complete coverage in the respective files (Gruber et al. 1996). We review next the techniques implemented during the EGM96 model development.

The Numerical Quadrature (NQ) Technique

The orthogonality relations of the surface spherical harmonics (Heiskanen and Moritz 1967, Sect. 1–13) constitute the underlying principle of the NQ approach. In theory, one has:

$$C_{nm} = \frac{1}{4\pi\gamma(n-1)} \iint_{\sigma} \Delta g(\theta,\lambda) Y_{nm}(\theta,\lambda) d\sigma. \tag{6.17}$$

Equation 6.17 requires the existence of gravity anomalies continuously covering the sphere. In practice, one has discrete area-mean values of gravity anomalies, on an equi-angular grid on the ellipsoid ($\overline{\Delta g}^{e}_{ij}$). The discretization of the surface integral in (6.17), and the consideration of the ellipticity of the surface on which the $\overline{\Delta g}^{e}_{ij}$ values reside (see Jekeli 1988), lead to:

$$C^{T}_{nm} = \frac{1}{4\pi a\gamma(n-1)} \sum_{i=0}^{N-1} r_i^e \sum_{k=0}^{s'} \frac{L_{nmk}}{\bar{S}_{n-2k,|m|}(b/E)} \frac{\overline{IP}^{i}_{n-2k,|m|}}{q^{i}_{n-2k}} \cdot \sum_{j=0}^{2N-1} \overline{\Delta g}^{e}_{ij} \begin{Bmatrix} IC \\ IS \end{Bmatrix}^{j}_{m} \quad \begin{matrix} m \geq 0 \\ m < 0 \end{matrix} \quad . \tag{6.18}$$

The complete derivation of (6.18) can be found in Rapp and Pavlis (1990). The estimation of the complete high-degree set of geopotential coefficients is performed

here as a two-step procedure. First, the global set of $\overline{\Delta g}^{e}_{ij}$ provides, through (6.18), a "terrestrial" estimate, C^{T}_{nm}, of those harmonic coefficients present in the satellite-only model. In addition, (6.18) is used to propagate the error variances of $\overline{\Delta g}^{e}_{ij}$ to C^{T}_{nm}, and thus form their complete error covariance matrix, $\mathrm{Cov}(C^{T}_{nm})$. Based on the harmonic coefficients of the satellite-only model, C^{S}_{nm}, and their associated error covariance matrix, $\mathrm{Cov}(C^{S}_{nm})$, and their "terrestrial" counterparts, a least-squares adjustment is performed to estimate a unique set of coefficients (and its associated error covariance matrix), essentially as a weighted average of the two independent estimates. The formulation of this adjustment process is described in full detail in Rapp and Pavlis (1990, Sect. 2.3). This adjustment provides also a global set of adjusted gravity anomalies. In a second step, the adjusted gravity anomalies are input to (6.18) to yield the complete set of harmonic coefficients up to the high degree (360 or higher). The error variances of these higher-degree coefficients may be obtained by quadratic addition of the propagated anomaly error and an estimate of the sampling error (ibid., (50)–(53)). This general procedure was originally proposed by Kaula (1966) and has been used in several high-degree models developed at The Ohio State University (Rapp 1981; Rapp and Cruz 1986a; Rapp and Pavlis 1990).

Composite quadrature weights $1/q^{i}_{n}$ were introduced by Colombo (1981a, p. 76) as an efficient way of approximating the harmonic analysis results obtainable using Optimal Estimation (their latitude dependence was suggested by Katsambalos (1979). Pavlis (1996) introduced the following set of composite de-smoothing factors q^{i}_{n}, which avoid the discontinuities of Colombo's (ibid.) original set:

$$q^{i}_{n} = \begin{cases} (\beta^{i}_{n})^{2} & 0 \le n \le L/2 \\ (\beta^{i}_{n})^{L/n} & L/2 < n \le L \\ \beta^{i}_{L} & L < n \end{cases} . \tag{6.19}$$

$L(= \pi/\Delta\lambda)$ is the Nyquist degree implied by the sampling interval $\Delta\lambda$, and β^{i}_{n} is the Pellinen operator computed by:

$$\beta^{i}_{n} = \frac{1}{(1-\cos\psi^{i}_{0})}\frac{1}{(2n+1)}\left[P_{n-1}(\cos\psi^{i}_{0}) - P_{n+1}(\cos\psi^{i}_{0})\right], \tag{6.20}$$

where ψ^{i}_{0} is the semi-aperture of a spherical cap having the same area as the equi-angular block on the ith latitude band. It is computed by Colombo (1981a, p. 85):

$$\psi^{i}_{0} = arc\cos\left[\frac{\Delta\lambda}{2\pi}(\cos\Delta_{i+1} - \cos\Delta_{i}) + 1\right], \tag{6.21}$$

where δ is the reduced co-latitude (Heiskanen and Moritz 1967, Sect. 1-19). Introduction of the de-smoothing factors of (6.19) enabled Pavlis (1996) to extend

NQ models (developed using $30' \times 30'$ $\overline{\Delta g}^{e}_{ij}$) to degrees higher than 360, without experiencing large jump discontinuities at the Nyquist degree (360) implied by the $30' \times 30'$ data sampling.

The Block-Diagonal (BD) Least-Squares Adjustment Technique

Equation 6.10 could be used to form normal equations from a global set of $30' \times 30'$ $\overline{\Delta g}^{e}_{ij}$ data. These normal equations could then be combined with the satellite-only normal equations, to yield the combination solution. Rapp (1967) proposed originally this approach. However, to implement this technique to a high degree, one has to deal with the extremely high computational demands of such a task. This may be accomplished by forming, instead of the full (symmetric) "terrestrial" normal matrix, a suitable approximation of it. This approximation should be simple enough, to allow numerical implementation, and, on the same time, rigorous enough to maintain the most important characteristics of the full matrix. Colombo (1981a) has shown that if:

(a) The data reside on a surface of revolution (e.g., a rotational ellipsoid),
(b) The grid is complete and the longitude increment constant,
(c) The data weights are longitude-independent,
(d) The data weights are symmetric with respect to the equator, then, zero elements in the normal equations formed in the least-squares estimation will occur as prescribed by (see also (Pavlis 1988) for details):

$$[\mathbf{N}]_{C_{nm}\,C_{rs}} = 0 \quad if \quad \{m \neq s\} \; or \; \{m = s \; and \; n - r = 2k + 1\}\,. \tag{6.22}$$

Note that in this notation the order subscript is a signed integer, whose sign identifies the type of coefficient (positive: cosine, negative: sine). If condition (d) does not hold true, then:

$$[\mathbf{N}]_{C_{nm}\,C_{rs}} = 0 \quad if \quad \{m \neq s\}\,. \tag{6.23}$$

The sparsity patterns implied by (6.22) and (6.23) will be referred to as BD1 and BD2 respectively. In addition, a BD3 pattern will be considered defined by:

$$[\mathbf{N}]_{C_{nm}\,C_{rs}} = 0 \quad if \quad \{|m| \neq |s|\}\,, \tag{6.24}$$

which admits non-zero off-diagonal elements across coefficients of different type within a given order. It is instructive to consider the computational efficiency implied by these patterns. Table 6.2 provides relevant statistics for a solution complete from degree and order 0 to degree and order 360, excluding degree $n = 1$ terms. Such a solution involves 130,318 unknown coefficients, and the upper (or lower) triangular part of the (symmetric) full "terrestrial" normal matrix has 8,491,455,721 elements.

Table 6.2 Statistics of normal matrices related to an expansion complete to $N_{max} = 360$ (excluding degree $n = 1$ coefficients) using different sparsity patterns

Statistic	Sparsity pattern		
	BD1	BD2	BD3
Total number of non-zero elements	7,905,721	15,746,100	31,362,241
Percentage of full matrix elements	0.09	0.19	0.37
Number of blocks	1,440	721	361
Num. of unknowns in largest block	181	360	718
Num. of elements in largest block	16,471	64,980	258,121

The enormous computational savings that can be inferred from Table 6.2 make the BD approximations very attractive estimation strategies. These savings however come at the expense of the rigor exercised in the development of the model. The real-world anomaly data to be analyzed comply *only* with the conditions (a) and (b) above (in fact, to comply even with the (a) and (b) conditions, "filling-in" techniques and analytical continuation have to be employed, since the original $30' \times 30'$ data file is neither complete, nor residing on any surface of revolution). Furthermore, the normal equations of the EGM96S satellite-only model do not conform to any such sparsity pattern. BD3 is the most rigorous of the three patterns, while being well within our present computational capabilities. In EGM96 we therefore chose to form the "terrestrial" normal equations according to BD3. We did not however alter the data weights to enforce compliance with (c) or (d). Furthermore, the satellite-only normal equations were *not* truncated, since this was found to degrade the combination solution unacceptably, at the lower degrees (Lerch et al. 1993).

The BD technique may be viewed as an intermediate type of approach between the rigorous *comprehensive* least-squares adjustment procedure and the NQ procedure. The BD approach combines some of the advantages of the other two approaches, while avoiding their critical shortcomings. Pavlis et al. (1996a) discuss some of the analytical differences between the NQ and the BD techniques, both from the harmonic analysis and from the combination solution points of view. Three important aspects of the BD approach require some discussion here.

1. ***Ordering and Grouping of the Unknowns.*** The particular ordering of the unknown potential coefficients within the vector of parameters has a tremendous impact on the efficiency with which the combined (satellite-only plus terrestrial) normal equations can be inverted. To illustrate this, let us consider for a moment that a "high-degree" solution complete to $N_{max} = 6$ is to be developed, in a least-squares combination with a satellite-only model complete to degree $N_{sat} = 4$. In this case, the terrestrial normal equations involve 46 unknowns (complete set to $N_{max} = 6$, excluding $n = 1$ terms), while the satellite-only normal equations 22 unknowns. The unknown coefficients are ordered first by increasing order (m), then by type (C then S), and lastly by increasing degree (n). This is denoted ordering pattern "V" in Pavlis (1988, Table 6.3). Adhering to the sparsity pattern BD3, the *terrestrial* normal equations take the form shown in Fig. 6.3a, where

gray areas indicate non-zero elements. This type of normal matrix can be set up and inverted very efficiently, thus providing the terrestrial-only estimates of the coefficients and their associated BD error covariance matrix. For an analysis to $N_{\max} = 359$, (which is the maximum degree resolvable from $30' \times 30'$ mean values using least-squares) this matrix contains 360 diagonal blocks, the largest one having dimension 716×716 (corresponding to order $m = 1$), while the smallest one having dimension 2×2 (corresponding to $m = 359$).

However, if one conforms to this ordering of unknowns, the *combined* (terrestrial plus satellite-only) normal equations take the form shown in Fig. 6.3b. In this figure, black areas indicate the non-zero elements in the combined normal equations, which arise from the satellite-only normal equation contribution (overlaid on the structure of Fig. 6.3a). It is obvious that the "V" type of ordering of the unknowns creates a large (although sparse) block in the combined normal equations, which would have to be treated as a full matrix. In the real-world (EGM96) case, where $N_{sat} = 70$, this block would have dimension $45,787 \times 45,787$. Clearly, a different ordering of the unknowns is required, whereby the coefficients present in the satellite-only model would be grouped together. Two ways to accomplish this are:

Forward grouping	**Reverse grouping**
Group 1: $n \leq N_{sat}$, $m \leq N_{sat}$	Group 1: $n > N_{sat}$, $m > N_{sat}$
Group 2: $n > N_{sat}$, $m \leq N_{sat}$	Group 2: $n > N_{sat}$, $m \leq N_{sat}$
Group 3: $n > N_{sat}$, $m > N_{sat}$	Group 3: $n \leq N_{sat}$, $m \leq N_{sat}$

Inside each group, the coefficients are ordered following the same pattern "V" as before. Figure 6.3c, d show the structure of terrestrial and combined normal equations respectively, for the forward grouping, while Fig. 6.3e, f for the reverse grouping. Bosch (1993) studied the structure of the combined normal equations resulting from the forward grouping and proposed an algorithm for the solution of such a system. However, the reverse grouping possesses a very significant advantage over the forward one. Namely, the Cholesky factor of the matrix in Fig. 6.3f preserves the structure of the upper (or lower) part of the original matrix. This enables a very efficient solution of the combined system, and provides the possibility to compute selected elements of the inverse of the combined normal equations (error covariance matrix). This was recognized by Chan and Pavlis (1995), and independently by Schuh (1996).

2. ***Reference Values and Aliasing Effects.*** Consider the rigorous (complete) set of normal equations obtained from the merged $30' \times 30'$ $\overline{\Delta g}^{e}_{ij}$ data, denoted by:

$$\mathbf{N} \cdot \hat{\mathbf{X}} = \mathbf{U}, \tag{6.25}$$

where $\hat{\mathbf{X}}$ represents the adjusted coefficients of the disturbing potential, i.e., remainders after subtraction of the even zonal harmonics of the normal ellipsoidal field. The BD3 truncated version of the normal equation system may be written as:

$$\tilde{\mathbf{N}} \cdot \hat{\tilde{\mathbf{X}}} = \mathbf{U}. \tag{6.26}$$

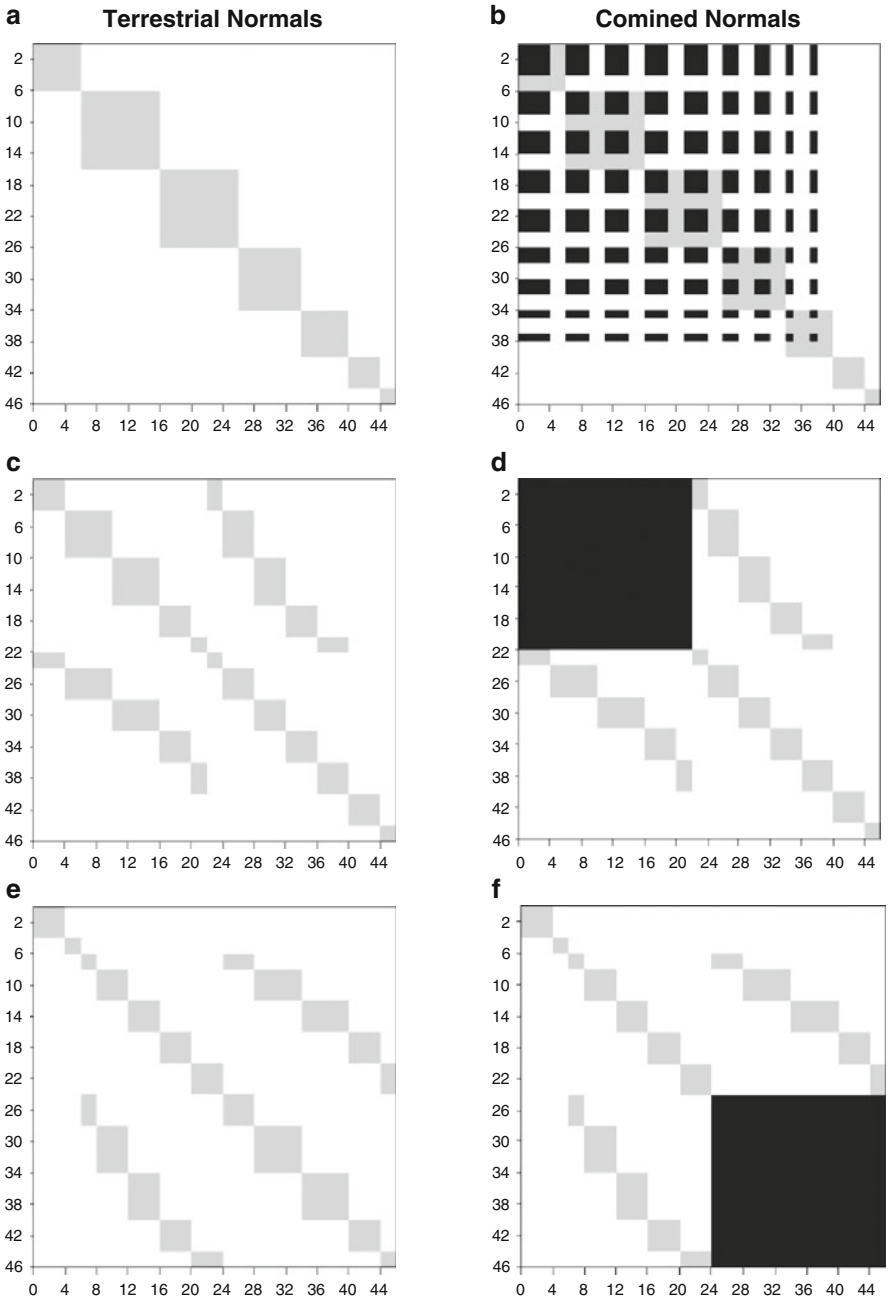


Fig. 6.3 Structure of terrestrial and combined normal equations for different orderings and groupings of the unknown coefficients

Notice that in both cases the right hand side vector **U** is the same (and is computed rigorously). The difference in the estimates of the unknowns (rigorous minus BD3 approximation) is therefore:

$$\mathbf{dX} = \hat{\mathbf{X}} - \hat{\tilde{\mathbf{X}}} = (\mathbf{N^{-1}} - \tilde{\mathbf{N}}^{-1}) \cdot \mathbf{U}. \tag{6.27}$$

Equation 6.27 indicates that the magnitude of **dX** may be reduced either by reducing the magnitude of the term $(\mathbf{N^{-1}} - \tilde{\mathbf{N}}^{-1})$ (i.e., by providing a better approximation to the normal matrix), or, for a given approximation of the normal matrix, by reducing the magnitude of **U**. One may reduce the magnitude of **U** by modeling residual anomalies, after subtraction of a high-degree reference model, instead of the original $30' \times 30'$ $\overline{\Delta g}^{e}_{ij}$ values, which refer to the normal (ellipsoidal) field. Moreover, as it is discussed by Pavlis et al. (1996a), the BD solution (unlike the NQ one) may be affected significantly at the high degrees by aliasing. This arises from signal present in the $30' \times 30'$ data, which corresponds to harmonics beyond those solved-for. In order to reduce the magnitude of **dX** in (6.27), and also reduce aliasing effects, Pavlis et al. (1996b) suggested removing a "reference" $\overline{\Delta g}^{e}_{ij}$ value from the original $30' \times 30'$ anomalies. This reference value can be computed from a preliminary NQ model to $N_{\max} = 460$. The BD combination solution is then performed using residual anomalies, and yields a set of coefficient corrections to $N_{\max} = 359$. Addition of the reference model coefficients (to $N_{\max} = 359$) yields the final BD high-degree expansion. This is the procedure that was followed in the development of EGM96, for the portion of that model from degree $n = 71$ to $n = 359$.

3. ***Use of an A Priori Constraint.*** Un-modeled long wavelength systematic errors that are (still) present in global gravity databases, coupled with our inability to numerically account for error correlations between the $30' \times 30' \overline{\Delta g}^{e}_{ij}$ data, necessitate some down-weighting of these data in present combination solutions. This aims to preserve the highly accurate long wavelength information contributed by the satellite-only normal equations. This down-weighting, in conjunction with the use of diagonal anomaly weight matrices, has the undesirable side effect that the propagated errors of the combination solution at the high degrees (n > 250) are too pessimistic. This was the case for both the OSU89A/B models (Rapp and Pavlis 1990), and for the OSU91A model (Rapp et al. 1991). The BD approach offers a possible (at least partial) remedy to this problem. This may be accomplished by introducing some a priori information for the coefficient *corrections* (relative to the NQ reference model), for n > 70. This a priori information may be provided in the form of some anomaly degree variance model. Although this approach helps produce a more realistic error spectrum at the higher degrees (n > 250), it is a rather simplistic and empirical way of attacking the underlying problem. Furthermore, it has the undesirable side effect of "dampening" also the power of the signal itself, a characteristic of the EGM96 model that attracted some criticism (see also Jekeli 1999b). More study is needed to develop better ways of treating systematic errors in global gravity anomaly data sets either by direct estimation and/or by using more sophisticated error modeling.

6.6.2 EGM2008

The main reason for the choice of the solution strategy implemented in the development of EGM96 was the fact that its satellite-only component (EGM96S) was accompanied by a variance-covariance matrix that was fully-occupied. This was due to the fact that the heterogeneous tracking data from the (approximately) 40 satellites that were used to derive EGM96S to degree and order 70, were incapable of de-correlating adequately the spherical harmonic coefficients up to this degree and order. Therefore, in order to preserve the integrity of the least-squares adjustment used to combine EGM96S with the surface gravity and altimetry data, one had to consider the satellite-only normal equations in their complete (fully-occupied) form. Any block-diagonal approximation of these normal equations would result in estimation errors that could not be tolerated. This situation changed dramatically with GRACE. Due to the global coverage and uniform accuracy of the GRACE data, the corresponding normal equation matrix could be safely approximated by a block-diagonal matrix, without significant loss of accuracy. In addition, after the availability of satellite-only models from GRACE, there is really no need to incorporate altimeter data into the combination solution in the form of "direct" tracking. Instead, a preliminary model based on GRACE data and a MSS, can be used to estimate a preliminary model of the DOT. This DOT model could then be used to correct the altimeter data, and estimate from them an ocean-wide set of altimetry-derived gravity anomalies. These anomalies can be "merged" with terrestrial and airborne data to form a complete global gravity anomaly grid. The gravitational information implied by these gridded data could then be combined (in a least-squares adjustment) with the satellite-only model from GRACE, to derive the combination solution, up to the high degree (2,159), in a single step. The entire process may be iterated, using the high-degree combination solution to derive the next estimate of the DOT, and so on. This is essentially the approach that was used to develop EGM2008 (Pavlis et al. 2008). Two iterations of the estimation of the altimetry-derived gravity anomalies were performed, which was sufficient for the process to converge. Despite its iterative nature, this approach permits the development of very high-degree combination solutions in an efficient manner, and moreover in a single adjustment step, thereby avoiding the "piece-wise" nature of models like EGM96.

In terms of data complement and solved-for parameters, EGM2008 resembles the OSU89A/B solutions (Rapp and Pavlis 1990). The gravitational information of a satellite-only model (accompanied by its complete error variance-covariance matrix) is combined with the corresponding information from the analysis of a complete set of area-mean gravity anomalies given over a global equi-angular grid, to estimate a set of potential coefficients complete to a harmonic degree commensurate with the resolution of the gravity anomaly data. In EGM2008 (as in OSU89A/B), a model of the DOT was not estimated simultaneously with the potential coefficients, in contrast to the strategy used in the development of EGM96. In the following sections we describe in some detail the "building blocks" that were used to estimate the EGM2008 very high-degree model.

6.6.2.1 The ITG-GRACE03S Satellite-Only Model

The satellite-only model that was used in the development of EGM2008 is designated ITG-GRACE03S (Mayer-Gürr 2007). This model was developed at the Institute of Theoretical Geodesy of the University of Bonn, in Germany. The ITG-GRACE03S model was based on 57 months of GRACE Satellite-to-Satellite Tracking (SST) data. No other data were used in the development of ITG-GRACE03S. A short-arc analysis approach was used for the development of ITG-GRACE03S, as described by Mayer-Gürr et al. (2007).

ITG-GRACE03S is complete to spherical harmonic degree and order 180, and was accompanied by its full error variance-covariance matrix. Due to the global coverage resulting from the near-polar orbits of the two GRACE spacecraft, and due to the uniform accuracy of the GRACE data, this error variance-covariance matrix is diagonally-dominant. Numerical experiments indicated that a block-diagonal approximation of this matrix according to the BD1 scheme (see (6.22) would be sufficient to maintain the essential characteristics of the errors associated with the ITG-GRACE03S model, without any appreciable loss of accuracy in the development of the combination gravitational solution. This simplified tremendously the numerical implementation of the combination solution, as we discuss in Sect. 6.6.2.3.

The ITG-GRACE03S model was developed and was made available in terms of spherical harmonic coefficients. However, as we discuss next, the analysis of the terrestrial data, and the combination solution that led to the development of EGM2008, were implemented in terms of ellipsoidal harmonics (see also Jekeli 1988 for details). Therefore, in a first step, both the ITG-GRACE03S coefficient model and its associated error variance-covariance matrix were transformed from spherical to ellipsoidal harmonics, using the transformation formulas developed by Jekeli (1988) and implemented by Gleason (1988). The transformation from spherical to ellipsoidal harmonic coefficients is given in Gleason (1988, Eq. 2.8). The reverse (ellipsoidal to spherical) transformation is given in Gleason (1988, Eq. 2.10). Both are linear transformations that preserve the harmonic order but not the harmonic degree. It is very important to recognize that both transformations preserve the structure of the BD1 block-diagonal pattern of the error variance-covariance matrix. This aspect of the transformations is of critical importance to the computational efficiency of both the least-squares adjustment necessary to derive the combination solution, and of the subsequent error propagation associated with the final combined solution. In the development of EGM2008, we first transformed the full error variance-covariance matrix of ITG-GRACE03S from the spherical to the ellipsoidal harmonic representation, and then approximated the resulting matrix to one conforming to the BD1 block-diagonal pattern.

The outcome from this "pre-processing" of the ITG-GRACE03S information is a set of ellipsoidal harmonic coefficients of this satellite-only model, $C_{nm}^{S,e}$, complete to degree and order 180, accompanied by the BD1 approximation of its error variance-covariance matrix, $\mathrm{Cov}(C_{nm}^{S,e})$.

6.6.2.2 The Block-Diagonal (BD) Least-Squares "Terrestrial" Coefficient Estimates

The second "building block" of EGM2008 is the gravitational information obtained from the analysis of a complete global equi-angular $5' \times 5'$ grid of area-mean gravity anomalies. These anomalies have been corrected for ellipsoidal corrections, and have been analytically continued to the surface of the reference ellipsoid. They are denoted by $\overline{\Delta g}^e_{ij}$ and represent exactly the same type of observations (albeit at a different grid size) as the $\overline{\Delta g}^e_{ij}$ of section "The Numerical Quadrature (NQ) Technique", where we discussed the development of EGM96. In the notation of (6.10), considering the small latitudinal extent of the $5' \times 5'$ cell, the small and regular latitudinal variation of r^e within the cell can be safely ignored (see also Rapp and Pavlis 1990, p. 21,887), so that we may approximate:

$$\overline{(r\Delta g)}^e_{ij} \approx r^e_i \cdot \overline{\Delta g}^e_{ij}, \tag{6.28}$$

where r^e_i is the geocentric distance to the center of the (i, j)th cell. The product $r^e_i \cdot \overline{\Delta g}^e_{ij}$, defined over the surface of the reference ellipsoid, can be expanded into surface ellipsoidal harmonic functions (Heiskanen and Moritz 1967, Sect. 1-20), as:

$$r^e_i \cdot \overline{\Delta g}^e_{ij} = \frac{1}{\Delta\sigma_i} \frac{GM}{a} \sum_{n=2}^{M} (n-1) \sum_{m=-n}^{n} C^{T,e}_{nm} \cdot IY^{ij}_{nm}, \tag{6.29}$$

where δ is the reduced co-latitude (Heiskanen and Moritz 1967, Sect. 1-19) and:

$$\Delta\sigma_i = \Delta\lambda \int_{\delta_i}^{\delta_{i+1}} \sin\delta d\delta = \Delta\lambda \cdot (\cos\delta_i - \cos\delta_{i+1}), \tag{6.30}$$

$$IY^{ij}_{nm} = \int_{\delta_i}^{\delta_{i+1}} \overline{P}_{n|m|}(\cos\Delta) \sin\delta d\delta \cdot \int_{\lambda_j}^{\lambda_{j+1}} \left\{ \begin{matrix} \cos m\lambda \\ \sin|m|\lambda \end{matrix} \right\} d\lambda \begin{matrix} \text{if } m \geq 0 \\ \text{if } m < 0 \end{matrix}. \tag{6.31}$$

The quantity $r^e \Delta g^e$ represents a harmonic function, and, under the approximation of (6.28), so does the quantity $r^e_i \cdot \overline{\Delta g}^e_{ij}$. This allows one to relate the *ellipsoidal* harmonic coefficients $C^{T,e}_{nm}$ of (6.29), to the corresponding *spherical* harmonic coefficients $C^{T,s}_{nm}$, using the exact transformations derived by Jekeli (1988) and implemented and verified by Gleason (1988). Note that our $C^{T,s}_{nm}$ and $C^{T,e}_{nm}$ coefficients are related to the corresponding $\overline{g}^s_{n,m}$ and $\overline{g}^e_{n,m}$ coefficients of Gleason (ibid.) by:

$$\left\{ \begin{matrix} \overline{g}^s_{n,m} \\ \overline{g}^e_{n,m} \end{matrix} \right\} = \frac{GM}{a^2}(n-1) \cdot \left\{ \begin{matrix} C^{T,s}_{nm} \\ C^{T,e}_{nm} \end{matrix} \right\}. \tag{6.32}$$

Based on (6.29), one forms a system of observation equations that can be written in vector format as:

$$\mathbf{v} = \mathbf{A} \cdot \hat{\mathbf{x}} - \mathbf{L_b}, \tag{6.33}$$

where $\mathbf{L_b}$ is the vector of observations $\overline{\Delta g}^e_{ij}$, $\mathbf{v}$ is the vector of corresponding residuals, $\mathbf{A}$ is the design matrix whose elements are formed based on (6.29), and $\hat{\mathbf{x}}$ represents the vector of estimated coefficients $C^{T,e}_{nm}$. The least-squares solution, $\hat{\mathbf{x}}$, which minimizes the quadratic form $\mathbf{v^T P\, v}$, is given by Uotila (1986):

$$\left.\begin{array}{ll} \hat{\mathbf{x}} = \mathbf{N^{-1}U} & \text{(a)} \\ \mathbf{N} = \mathbf{A^T P\, A} & \text{(b)} \\ \mathbf{U} = \mathbf{A^T P\, L_b} & \text{(c)} \end{array}\right\}, \tag{6.34}$$

where $\mathbf{P}$ is the weight matrix associated with the observations $\overline{\Delta g}^e_{ij}$. $\mathbf{P}$ was assumed diagonal, with elements equal to the reciprocal of the error variance associated with each individual gravity anomaly observation, i.e.:

$$\mathbf{P} = \sigma_0^2 \cdot \begin{pmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_K^2} \end{pmatrix}, \tag{6.35}$$

where K is the total number of observations, and σ_0^2 is the a priori variance of unit weight. For the complete, global equi-angular $5' \times 5'$ grid of area-mean gravity anomalies used here, $K = 2160 \times 4320 = 9331200$. The assumption that the gravity anomaly errors are uncorrelated is made out of necessity, rather than desire. It is extremely difficult to estimate the error correlations between the gravity anomalies with any degree of accuracy. It is also practically impossible to handle numerically a full (symmetric) weight matrix of dimension 9,331,200. Even the estimation of realistic error variances for the gravity anomalies is itself a very challenging task. These error variances affect critically the error variance-covariance matrix of the "terrestrial" coefficients $C^{T,e}_{nm}$, which is given by:

$$\mathrm{Cov}(C^{T,e}_{nm}) = \sigma_0^2 \cdot \mathbf{N^{-1}}. \tag{6.36}$$

The gravity anomaly error variances should be such that they represent realistic estimates of the accuracy of the data, not their precision. They should produce an error variance-covariance matrix $\mathrm{Cov}(C^{T,e}_{nm})$ that would permit the combination of $C^{T,e}_{nm}$ with the satellite coefficients $C^{S,e}_{nm}$ using well "calibrated" *relative* weights. The gravity anomaly error estimates, as well as the error estimates associated with the satellite information, should also be realistic in an *absolute* sense. Otherwise, the a posteriori error estimates associated with the combination solution will not reflect adequately the real accuracy of the combined model.

With the complete, global equi-angular $5' \times 5'$ grid of area-mean gravity anomalies as input, the expansion of (6.29) was extended to maximum degree and order $M = 2159$, in ellipsoidal harmonics. This is the maximum degree for which the system of (6.29) still maintains full rank (Colombo 1981a). The "terrestrial" normal equations were approximated by their BD1 counterpart. Although the weights associated with the gravity anomalies do not strictly comply with the requirements for a BD1 sparsity pattern (see section "The Block-Diagonal (BD) Least-Squares Adjustment Technique"), the geographic variation of these weights do not produce significant departures from such a pattern. This is mainly due to the uniformity of the errors of altimetry-derived gravity anomalies, which cover approximately 70% of the Earth's total area.

It should also be emphasized here that the residuals appearing in (6.33) represent a measure of "goodness of fit" and are not necessarily representative of the errors of the gravity anomaly data (Pavlis 1988). In fact, if the gravity anomaly data were limited in spectral content and contained contributions *only* from (a subset of) the solved-for harmonics appearing in (6.29), these residuals would have been identically zero (to the level of the numerical noise).

6.6.2.3 The Least-Squares Combination Solution

The least-squares combination solution coefficient set, $C_{nm}^{C,e}$, is determined essentially as the weighted average of the satellite-only estimate, $C_{nm}^{S,e}$, and of the "terrestrial" estimate, $C_{nm}^{T,e}$, each of these two *independent* estimates being weighted by the inverse of its respective error variance-covariance matrix, according to:

$$\begin{aligned} C_{nm}^{C,e} = & \left\{ \left[\mathrm{Cov}(C_{nm}^{S,e}) \right]^{-1} + \left[\mathrm{Cov}(C_{nm}^{T,e}) \right]^{-1} \right\}^{-1} \cdot \\ & \left\{ \left[\mathrm{Cov}(C_{nm}^{S,e}) \right]^{-1} \cdot C_{nm}^{S,e} + \left[\mathrm{Cov}(C_{nm}^{T,e}) \right]^{-1} \cdot C_{nm}^{T,e} \right\}. \end{aligned} \tag{6.37}$$

The error variance-covariance matrix of $C_{nm}^{C,e}$, $\mathrm{Cov}(C_{nm}^{C,e})$, is given by:

$$\mathrm{Cov}(C_{nm}^{C,e}) = \left\{ \left[\mathrm{Cov}(C_{nm}^{S,e}) \right]^{-1} + \left[\mathrm{Cov}(C_{nm}^{T,e}) \right]^{-1} \right\}^{-1}. \tag{6.38}$$

It is important to recognize here that the BD1 approximation of both the satellite-only and the "terrestrial" error variance-covariance matrices permits the evaluation of the combination solution in a fashion that is extremely fast and numerically economic. This is because the linear system representing the entire combination solution is comprised of uncorrelated BD1-type blocks that can be inverted independently of each other. For the solution to degree and order 2,159, the largest (symmetric) matrix that needs to be stored and inverted is of the order of 1,080, a task that is trivial for the currently available computers.

Evaluation of (6.29), using the combined solution coefficients, $C_{nm}^{C,e}$, in the place of $C_{nm}^{T,e}$, yields the set of adjusted area-mean gravity anomalies $\widehat{\overline{\Delta g}}_{ij}^{e}$, as:

$$r_i^e \cdot \widehat{\overline{\Delta g}}_{ij}^{e} = \frac{1}{\Delta\sigma_i}\frac{GM}{a}\sum_{n=2}^{M}(n-1)\sum_{m=-n}^{n} C_{nm}^{C,e}\cdot IY_{nm}^{ij}. \tag{6.39}$$

The residual gravity anomalies v_{ij} resulting from the least-squares adjustment that produced the combination solution are then computed as the difference between these adjusted anomalies and the original (input) values:

$$v_{ij} = \widehat{\overline{\Delta g}}_{ij}^{e} - \overline{\Delta g}_{ij}^{e}. \tag{6.40}$$

These residual anomalies are due to any existing differences between the satellite-only and the "terrestrial" estimates of the gravity anomalies. The values of these residual anomalies are affected directly by the weights used in the combination solution for the satellite-only estimate relative to its "terrestrial" counterpart.

A final step towards the estimation of the combination high-degree solution is the transformation of the *ellipsoidal* harmonic coefficients $C_{nm}^{C,e}$, and of their associated error variance-covariance matrix $\mathrm{Cov}(C_{nm}^{C,e})$, to their *spherical* counterparts, $C_{nm}^{C,s}$ and $\mathrm{Cov}(C_{nm}^{C,s})$. This is performed again using the ellipsoidal-to-spherical transformation formulas of Jekeli (1988) and Gleason (1988). Due to the fact that this transformation preserves the order but not the degree, an ellipsoidal harmonic expansion complete to degree and order 2,159, as in the case of EGM2008, produces a corresponding spherical harmonic coefficient set that extends up to degree 2,190. The "extra" coefficients are linear combinations of the lower degree coefficients (Jekeli 1988). Such "extra" coefficients are of negligible effect for expansions to degree 360 or so (e.g., EGM96), but cannot be omitted in expansion that extend to degree 2,159 (e.g., EGM2008). In such very high-degree expansions, omission of these "extra" coefficients will result in unacceptable modeling errors, especially over high latitude areas (see also Holmes and Pavlis 2007 for details).

6.6.2.4 Error Propagation

Users of a high-resolution global gravitational model require geographically specific estimates of the error associated with various gravitational functionals (e.g., gravity anomalies, height anomalies, deflections of the vertical) computed from the model. These estimates are composed of the commission and the omission error implied by a specific model. The commission (or propagated) error is due to the fact that a model that is based on actual observations can never be error-free since the data supporting its development can never be error-free. The omission (or truncation) error is due to the fact that a model can only have finite resolution; therefore it will always omit a portion of the Earth's true gravity field spectrum, which extends to infinity. Rigorous computation of the commission error implied

by any model requires the complete error variance-covariance matrix of its defining parameters. In principle, given this matrix, one can compute the commission error of various model-derived functionals, using error propagation. The error variance-covariance matrix of a spherical harmonic model complete to degree and order 2,159 has dimension ∼4.7 million. The computation of such a matrix is beyond the existing computing technology. Even for expansions to degree and order 360, like EGM96, which involve approximately 130,000 parameters, the formation of the normal equation matrix, its inversion, and the subsequent error propagation using the resulting error variance-covariance matrix is a formidable computational task. For EGM96 (Lemoine et al. 1998), such error propagation was only possible for the portion of the model extending to degree and order 70. For EGM2008, which extends to degree 2,159 in ellipsoidal harmonics, the alternative error propagation technique that was developed and implemented by Pavlis and Saleh (2005) was used. This technique is capable of producing geographically specific estimates of a model's commission error, *without* the need to form, invert, and propagate large matrices. Instead, this technique uses integral formulas with band-limited kernels and requires as input the error variances of the gravity anomaly data that are used in the development of the gravitational model.

The main idea behind the technique of Pavlis and Saleh (2005) is the realization that in combination solutions like EGM96 and EGM2008, the satellite-only information influences the combined model only up to a relatively low degree, which is the maximum degree of the satellite-only solution. Up to this maximum degree, the combined solution is the outcome of a least-squares adjustment. However, the higher degree and order portion of the combined gravitational model (beyond the range of influence of the satellite information), is determined *solely* from a complete, global grid of area-mean gravity anomaly data. Therefore, beyond the maximum degree and order of the available satellite-only solution, there is little need to form complete normal matrices, since no "adjustment" takes place within this degree range. The merged (terrestrial plus altimetry-derived) area-mean gravity anomalies are the only data whose signal and error content determine the model's signal and error properties over this degree range. This fact enables high-degree error propagation, *with* geographic specificity, through the use of integral formulas with band-limited kernels, *without* the need to form, invert, and propagate extremely large matrices. We illustrate next the technique introduced by Pavlis and Saleh (2005), using geoid undulations as an example of a model-derived quantity.

Consider the gravity anomaly computed from a combined model as being composed of two separate spectral parts:

$$\widehat{\Delta g} = \widehat{\Delta g}_L + \widehat{\Delta g}_H = \sum_{n=2}^{L} \widehat{\Delta g}_n + \sum_{n=L+1}^{H} \widehat{\Delta g}_n, \tag{6.41}$$

where, L and H stand for *Low*- and *High*-degree, respectively. L denotes the maximum degree of the satellite-only model used to develop a combined solution which extends to degree and order H. The corresponding geoid undulation is then:

$$\widehat{N} = \widehat{N}_L + \widehat{N}_H = \sum_{n=2}^{L} \widehat{N}_n + \sum_{n=L+1}^{H} \widehat{N}_n, \tag{6.42}$$

and can be written as (Heiskanen and Moritz 1967, Eq. 2-163b):

$$\widehat{N} = \frac{R}{4\pi\gamma} \iint_{\sigma} \widehat{\Delta g}\, S(\psi)\, d\sigma. \tag{6.43}$$

The Stokes function $S(\psi)$ can also be decomposed into separate spectral components as (see ibid., Eq. 2-169):

$$\begin{aligned} S(\psi) &= \sum_{n=2}^{\infty} \frac{2n+1}{n-1} P_n(t) \\ &= \sum_{n=2}^{L} \frac{2n+1}{n-1} P_n(t) + \sum_{n=L+1}^{H} \frac{2n+1}{n-1} P_n(t) + \sum_{n=H+1}^{\infty} \frac{2n+1}{n-1} P_n(t) \\ &= S_L(\psi) + S_H(\psi) + S_\infty(\psi), \end{aligned} \tag{6.44}$$

where $t = \cos(\psi)$ and $P_n(t)$ is the Legendre polynomial of degree n. Substituting $S(\psi)$ in (6.43) by its three spectral components from (6.45), and considering (6.41), due to the orthogonality of spherical harmonics we have:

$$\widehat{N} = \frac{R}{4\pi\gamma} \iint_{\sigma} \left(\widehat{\Delta g}_L + \widehat{\Delta g}_H + 0 \right) \cdot [S_L(\psi) + S_H(\psi) + S_\infty(\psi)]\, d\sigma \quad \Rightarrow$$

$$\widehat{N} = \frac{R}{4\pi\gamma} \iint_{\sigma} \widehat{\Delta g}_L\, S_L(\psi) d\sigma + \frac{R}{4\pi\gamma} \iint_{\sigma} \widehat{\Delta g}_H\, S_H(\psi) d\sigma = \widehat{N}_L + \widehat{N}_H. \tag{6.45}$$

Therefore, a strict, degree-wise separation of spectral components can be achieved by restricting the spectral content of the kernel function accordingly, *as long as* the integration is performed globally. The *High*-degree band-limited version of Stokes's equation:

$$\widehat{N}_H = \frac{R}{4\pi\gamma} \iint_{\sigma} \widehat{\Delta g}_H\, S_H(\psi)\, d\sigma\ , \tag{6.46}$$

implies, for *uncorrelated* errors of $\widehat{\Delta g}_H$, the following error propagation formulas:

$$\left.\begin{aligned} \sigma^2(\widehat{N}_H) &= \left(\frac{R}{4\pi\gamma}\right)^2 \iint_{\sigma} \sigma^2(\widehat{\Delta g}_H)\, S_H^2(\psi)\, d\sigma \qquad (a) \\ \sigma_{12}(\widehat{N}_H) &= \left(\frac{R}{4\pi\gamma}\right)^2 \iint_{\sigma} \sigma^2(\widehat{\Delta g}_H)\, S_H(\psi_1)\, S_H(\psi_2)\, d\sigma \qquad (b) \end{aligned}\right\}. \tag{6.47}$$

Equation 6.47a provides the error variance of the high-degree geoid undulation component, while (6.47b) the error covariance of the same component between two points located at ψ_1 and ψ_2 spherical distance respectively. Discretized versions of (6.47a, b) allow the computation of $\sigma^2(\widehat{N}_H)$ and $\sigma_{12}(\widehat{N}_H)$ from $\sigma^2(\widehat{\Delta g}_H)$ through global convolutions. One can implement (6.47a) using the 1D FFT approach of Haagmans et al. (1993), with H covering the degree range where the merged (terrestrial plus altimetry-derived) Δg define solely the solution. The geoid error covariances from (6.47b) can also be computed using global convolution, although with considerably less efficiency compared to the computation of error variances for points on regular grids. This approach is applicable to any functional f, related to Δg by an integral formula. Pavlis and Saleh (2005) provide the functional relationships required to propagate a model's error onto gravity anomalies, gravity disturbances, geoid undulations, and the components of the deflection of the vertical. Equations like (6.47a, b) employ the spherical approximation, which is considered quite adequate for error propagation work. Apart from this, these equations are rigorous, and their numerical implementation is only subject to discretization errors. Finally, the band limiting of integration kernels removes the singularity at the origin of kernels like Stokes's and Vening Meinesz's, therefore the innermost zone effects require no special treatment.

If we assume that the error correlation between $\widehat{\Delta g}_L$ and $\widehat{\Delta g}_H$ is negligible due to orthogonality, then the total error variance of a field functional, f, at the geographic location (r, φ, λ), as computed from a specific gravitational model, can be written as:

$$\begin{aligned}\sigma_f^2(r,\varphi,\lambda) \approx\ & \sigma_f^2(r,\varphi,\lambda)\_commission\_L \\ & +\sigma_f^2(r,\varphi,\lambda)\_commission\_H\ . \\ & +\sigma_f^2(r,\varphi,\lambda)\_omission\end{aligned} \tag{6.48}$$

$\sigma_f^2(r,\varphi,\lambda)\_commission\_L$ can be computed by propagation of the complete error variance-covariance matrix resulting from the least-squares adjustment that produced the combined solution, employing, e.g., the 2D FFT approach of Haagmans and Van Gelderen (1991). $\sigma_f^2(r,\varphi,\lambda)\_commission\_H$ can be computed by global convolution based on an integral formula as we illustrated above for the case of geoid undulations. Finally, $\sigma_f^2(r,\varphi,\lambda)\_omission$ may be estimated using, e.g., some local covariance model. This approach does not require one to form, invert, and propagate extremely large matrices. Figure 6.4 shows the propagated error of the geoid undulations computed from the EGM2008 model up to degree and order 2,159. This computation was performed on a global $5' \times 5'$ grid. Corresponding computations were also performed for gravity anomalies and for the deflection of the vertical components. In this fashion, the estimation of the propagated error of some specific functional at an arbitrary (φ, λ) location can be easily performed using interpolation, given the pre-computed global $5' \times 5'$ grid of the propagated error of the functional in question.

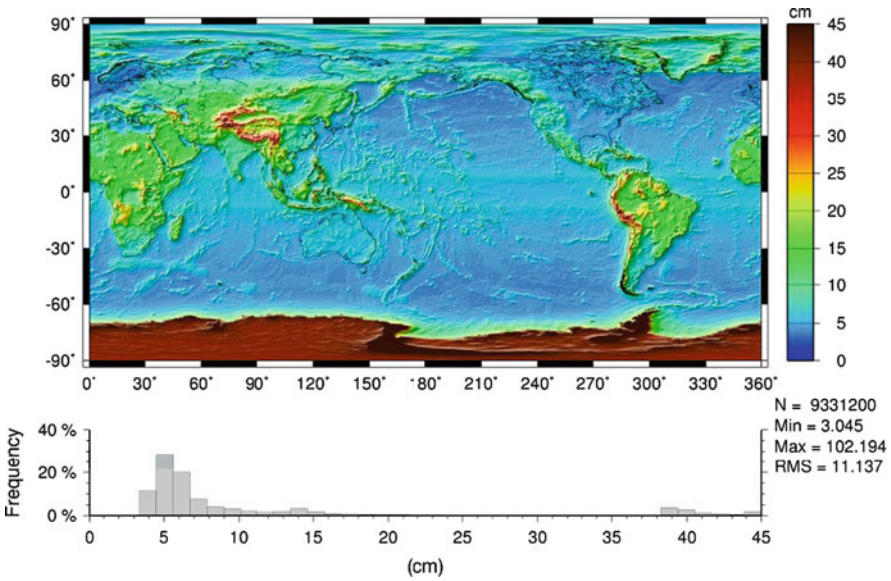


Fig. 6.4 Propagated error estimates (in centimeters) of the geoid undulations computed using the EGM2008 model to degree and order 2,159

6.6.2.5 Accuracy Assessment

The propagated error estimates of any global gravitational model depend strongly on the error estimates that were assigned to the data used in its development. Many times these data error estimates differ significantly from the true accuracy of the data. In addition, certain assumptions that may have been introduced within the development of a model (e.g., the assumption that the errors of the data are uncorrelated) could affect significantly the reliability of the model's propagated error estimates. In contrast, comparisons of model-derived quantities with data *independent* from the model, are in principle capable of revealing the true accuracy of a model. Of course, this requires the independent data that are used to test a model to be of significantly higher accuracy than the model-derived quantities. Such comparisons with independent data serve two general purposes:

(a) Evaluation of the accuracy of a model and inter-comparison of the performance of competing models.
(b) "Calibration" of the propagated error estimates. The comparison between the observed discrepancies between model-derived quantities and independent data, and the propagated errors of the model, allows one to test the veracity of the propagated error estimates, and to "calibrate" these estimates so that they match the observed performance of the model.

Several different data types, withheld from a model's development, are used for these purposes. These independent data have different spectral sensitivities and/or

Table 6.3 GPS/Leveling comparisons over CONUS

Model (Nmax)	Bias removed		Linear trend removed	
	Number passed edit	Weighted std. deviation (cm)	Number passed edit	Weighted std. deviation (cm)
EGM96 (360)	4,096	21.4	4,092	18.2
GGM02C_EGM96	4,169	18.9	4,165	17.6
EIGEN-GL04C (360)	4,167	19.5	4,163	18.1
EGM2008 (360)	4,185	17.6	4,181	16.4
EGM2008 (2,190)	4,201	7.1	4,197	4.8
USGG03 ($1' \rightarrow$10,800)	4,201	9.1	4,197	5.8

occupy different geographic regions. Tests that are usually employed here include satellite orbit determination and comparisons with tracking data, comparisons with geoid undulations obtained from GPS and leveling data (see e.g., Pavlis et al. 1993), comparisons with deflections of the vertical obtained from astrogeodetic techniques (Jekeli 1999b), comparisons employing altimetry and general ocean circulation models, etc. A useful practice, introduced during the development of EGM96, and used also during the development of EGM2008, is to invite a voluntary evaluation working group, independent of a model's developers, that evaluates and provides feedback to the model's developers regarding candidate preliminary solutions, as well as the final outcome from a modeling effort, in a manner as objective as possible. These groups usually work under the auspices of the International Association of Geodesy (IAG) and upon completion of a certain evaluation effort they report their findings in IAG-sponsored publications, which can be accessed freely by the public. In the case of EGM2008, the results from such an evaluation of both a preliminary version of the model (PGM2007A) (Pavlis et al., 2007b), as well as the final version of it, by 25 different international investigating teams are reported in *Newton's Bulletin No. 4*, which is jointly published by the Bureau Gravimétrique International (BGI) and the International Geoid Service (IGeS).

As an example of the evaluation of the EGM2008 and other models, Table 6.3 summarizes the results from the comparison of geoid undulations computed from GPS positioning and spirit leveling to model-derived values, over the conterminous United States (CONUS). A (thinned) set consisting of 4201 GPS/Leveling stations was used in this comparison. A ± 2 m editing criterion was applied to the differences between model-derived values and GPS/Leveling estimates. The analysis was done on a State by State basis, and the conversion from height anomalies to geoid undulations (Rapp, 1997b) was applied using a set of spherical harmonic coefficients of the elevation implied by the DTM2006.0 database (see Sect. 6.7), to a degree commensurate to the maximum degree of the gravitational model being tested. It is noteworthy that in this comparison, the EGM2008 model (which was developed based on $5' \times 5'$ area-mean gravity anomalies) performs better than the detailed ($1' \times 1'$) gravimetric geoid (USGG03), computed at the National Geodetic Survey (NGS) of the United States, using the most detailed *point* gravity anomaly data available for the area.

6.7 Data Requirements and Data Availability

The development of a GGM of very high degree and order requires a global set of gravity anomalies defined over a grid whose cell size is commensurate with the maximum degree of the expansion (e.g., $5' \times 5'$ for expansions to degree and order 2,160). One can form such a global grid, by merging gravity anomaly data obtained from terrestrial, ship-borne, air-borne, and satellite altimeter measurements. In addition to these data, elevation information in the form of a global Digital Topographic Model (DTM) is required. The resolution of this DTM should be considerably higher than the resolution of the gravity anomaly data grid to be compiled. We review next the essential aspects of these data requirements and describe the data that were available for the development of the EGM2008 model.

6.7.1 Elevation Data

The pre-processing and analysis of the detailed surface gravity data necessary to support the development of a GGM to harmonic degree and order 2,160, depends critically on the availability of accurate topographic data, at a resolution sufficiently higher than the resolution of the area-mean gravity anomalies, which will be used eventually for the development of the GGM. In Lemoine et al. (1998, Sect. 2.1) *Factor* discusses some of the uses of such topographic data within the context of the development and the subsequent use of a high-resolution GGM. These include the computation of Residual Terrain Model (RTM) effects, the computation of analytical continuation terms (g_1), the computation of Topographic/Isostatic gravitational models that may be used to "fill-in" areas void of other data, and the computation of models necessary to convert height anomalies to geoid undulations (Rapp, 1997b). For these computations to be made consistently, it is necessary to compile first a high-resolution global Digital Topographic Model (DTM), whose data will support the computation of all these terrain-related quantities.

For EGM96 (Lemoine et al. 1998), which was complete to degree and order 360, a global digital topographic database (JGP95E) at $5' \times 5'$ resolution was considered sufficient. JGP95E was formed by merging data from 29 individual sources, and, as acknowledged by its developers, left a lot to be desired in terms of accuracy and global consistency. Since that time, and thanks primarily to the Shuttle Radar Topography Mission (SRTM) (Werner 2001), significant progress has been made on the topographic mapping of the Earth from space. During approximately 11 days in 2000 (February 11–22), the SRTM collected data within latitudes 60°N and 56°S, thus covering approximately 80% of the total landmass of the Earth with elevation data of high, and fairly uniform, accuracy. Rodriguez et al. (2005) discuss in detail the accuracy characteristics of the SRTM elevations. Comparisons with ground control points whose elevations were determined independently using kinematic GPS positioning, indicate that the 90% absolute error of the SRTM elevations ranges

from ±6 to ±10 m, depending on the geographic area (ibid., Table 2.1). Additional information regarding the SRTM can be obtained from the web site of the United States' Geological Survey (USGS) (http://srtm.usgs.gov/), and from the web site of NASA's Jet Propulsion Laboratory (http://www2.jpl.nasa.gov/srtm).

In preparation for the development of the EGM2008 model, we compiled DTM2006.0 by overlying the SRTM data over the data of DTM2002 (Saleh and Pavlis 2003). In addition to the SRTM data, DTM2006.0 contains ice elevations derived from ICESat laser altimeter data over Greenland (Ekholm, personal communication 2005) and over Antarctica (DiMarzio, personal communication 2005). Over Antarctica, data from the "BEDMAP" project (http://www.antarctica.ac.uk/aedc/bedmap/) were also used to define ice and water column thickness. Over the ocean, DTM2006.0 contains essentially the same information as DTM2002, which originates in the estimates of bathymetry from altimetry data and ship depth soundings of Smith and Sandwell (1997). DTM2006.0 was compiled in $30'' \times 30''$ resolution (providing height and depth information only), and in $2' \times 2'$ and $5' \times 5'$ resolutions, where lake depth and ice thickness data are also included. DTM2006.0 is identical to DTM2002 in terms of database structure and information content. Pavlis et al. (2007a) provide details about the DTM2006.0 database and its use towards the development and implementation of the EGM2008 model.

6.7.2 Terrestrial Gravity Anomaly Data

For the development of EGM2008, terrestrial gravity anomaly data were compiled in the form of $5' \times 5'$ area-mean values. These values were estimated from point gravity measurements using Least Squares Collocation (LSC) (Moritz 1980), following the general approach described by Kenyon and Pavlis (1996). Ship-borne data were also used (primarily near the coasts), as well as airborne measurements where such measurements were available. Over certain areas, the terrestrial gravity data were limited to a resolution corresponding to $15' \times 15'$ area-mean values. In order to compile a global dataset with as much as possible uniform spectral content, capable of supporting the estimation of potential coefficients to degree 2,160, the spectral content of these gravity anomalies beyond degree 720 (corresponding to the $15' \times 15'$ resolution), was augmented with the gravitational information obtained from a global set of gravity anomalies implied by the Residual Terrain Model (RTM) effect (Forsberg 1984). This approach was initially tested and verified over areas where high quality gravity data are available (USA, Australia), as Pavlis et al. (2007a) discuss in more detail. The gravity anomalies synthesized in this fashion were designated as "fill-in" data.

Despite the improvements in gravity anomaly resolution, coverage, and accuracy that were realized during the EGM2008 modeling effort, there are still many areas of the globe (most notably Antarctica) where gravity anomaly data are sparse, poor in accuracy, or completely non-existent. In addition, the coverage and

quality of the available marine gravity data leave a lot to be desired. Pavlis (1988) demonstrated that long-wavelength errors present in the available marine gravity anomalies are a major contributor to the inconsistencies observed between satellite-only and surface gravity-only solutions. Marine gravity data are important to aid the separation (at least over short wavelengths) between the geoid undulation and the DOT signals, within the altimetry-derived sea surface height measurements. Efforts should therefore continue to try and improve the present status of the marine gravity data availability and quality.

6.7.3 Altimetry-Derived Gravity Anomalies

Two sources of altimetry-derived gravity anomalies were used for the compilation of the global $5' \times 5'$ area-mean gravity anomaly file used for the development of EGM2008. One set was estimated at the Danish National Space Center (DNSC), and was made available near the end of year 2007. This set was (internally) designated DNSC07C. DNSC07C is a predecessor of the DNSC08GRA data set that is described by Andersen et al. (2010b). The other set of altimetry-derived gravity anomalies was estimated in a collaborative effort between the Scripps Institution of Oceanography (SIO) and the National Oceanic and Atmospheric Administration (NOAA). The SIO/NOAA set is a predecessor of the data set described by Sandwell and Smith (2009). Preliminary tests performed during the development of EGM2008 indicated that the DNSC07C dataset performed better than the SIO/NOAA set near the coastlines, with the opposite behavior being observed over open ocean areas. Accordingly, the two altimetry-derived sets were "spliced" together, so that over a zone of ~ 190 km from the coast the DNSC07C set was used, followed by a "transition" zone of ~ 85 km where a weighted mean anomaly value computed from the two estimates (using complementary weights that vary linearly as one moves away from the coast), leading finally to 100% use of the SIO/NOAA values over the open ocean.

6.7.4 The Merged $5' \times 5'$ Area-Mean Gravity Anomaly File

To implement the BD estimation technique discussed in Sect. 6.6.2.2, one has to set up a global, complete file of $5' \times 5'$ area-mean gravity anomalies. Since the estimator does not allow for overlapping (duplicate) data input, one has to select for each $5'$ cell on the ellipsoid, the most accurate anomaly estimate out of multiple data that may be available for that cell (e.g., marine and altimetry-derived values). Rapp and Pavlis (1990) discuss such kind of data selection and merging algorithm. In the development of EGM2008, a similar (but not identical) algorithm was used. This process resulted in a complete global grid (9,331,200 values) of $5' \times 5'$ area-mean gravity anomalies, which were then input to the BD high-degree model estimator. Table 6.4 summarizes the overall statistics of this merged file.

Table 6.4 Statistics of the $5' \times 5'$ area-mean gravity anomalies of the merged file used to develop the EGM2008 model (Units are mGal)

Data source	% area	Minimum	Maximum	RMS	RMS σ
ArcGP	3.0	−192.0	281.8	30.2	3.0
Altimetry	63.2	−361.8	351.1	28.4	3.0
Terrestrial	17.6	−351.9	868.4	41.2	2.8
Fill-in	16.2	−333.0	593.5	46.8	7.6
Non Fill-in	83.8	−361.8	868.4	31.6	2.9
All	100.0	−361.8	868.4	34.5	4.1
(φ, λ)		19.4°, 293.5°	10.8°, 286.3°		

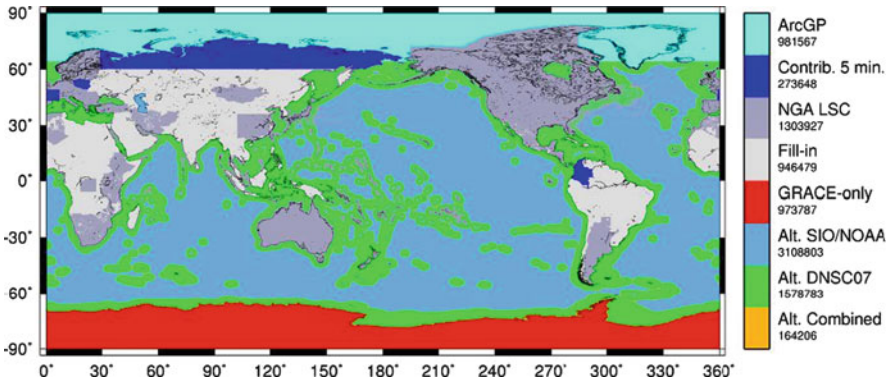


Fig. 6.5 Geographic distribution and source identification of the $5' \times 5'$ area-mean gravity anomalies in the merged file used to develop the EGM2008 model

Some noteworthy aspects of this merged file include the extensive use of $5' \times 5'$ area-mean gravity anomalies from the Arctic Gravity Project (ArcGP) (Kenyon and Forsberg 2008), and the avoidance of use of any Topographic/Isostatic mean anomalies (Pavlis and Rapp 1990). Over Antarctica, the $5' \times 5'$ area-mean gravity anomalies were synthesized purely on the basis of the ITG-GRACE03S (Mayer-Gürr 2007) model. This makes the EGM2008 model completely free of any isostatic hypothesis, at the cost of producing a smooth field over Antarctica (since ITG-GRACE03S is complete only up to degree and order 180). Figure 6.5 shows the geographic distribution and source identification of the $5' \times 5'$ area-mean gravity anomalies in the merged file used to develop the EGM2008 model.

6.8 Use of Global Gravitational Models and of Their By-Products

The estimated coefficients of the high-degree combination solution, $C_{nm}^{C,s}$, allow the user to compute the various functionals of the gravitational potential (e.g., gravity anomalies, height anomalies, deflections of the vertical), on or above the physical surface of the Earth, using harmonic synthesis. A versatile computer program

(HARMONIC_SYNTH), written in FORTRAN, which can be used to perform such harmonic synthesis, in various modes (e.g., for randomly scattered locations, for grids of point and/or area-mean values) was made available by Holmes and Pavlis (2006). This program, accompanied by test input and output files, and associated documentation is freely available from:

http://earth-info.nga.mil/GandG/wgs84/gravitymod/new_egm/new_egm.html

With regards to geoid computations, the user should also pay attention to some important issues related to the Permanent Tide, and the Geodetic Reference System (GRS) to which the computed geoid undulations (and/or height anomalies) refer. For example, in applications involving ellipsoidal heights obtained from space techniques (e.g., GPS), the user should be aware of the fact that the International Earth Rotation and Reference Systems Service (IERS), reports positions with respect to a (conventional) "Tide-Free" (also known as "Non-Tidal") crust. Therefore, in order to maintain consistency, geoid undulations and/or height anomalies involved in computations that use positions derived from space techniques, should be computed in the same Tide-Free system. In contrast, in applications involving satellite altimetry the "Mean Tide" system is used. Therefore, geoid undulations that are to be subtracted from altimetry-derived sea surface heights, in order to estimate the dynamic ocean topography, should also be computed in the Mean Tide system. The definition of the three systems used with regards to the Permanent Tide (Tide-Free, Mean, and Zero), and the relationships between the geoid undulations expressed in different systems is discussed in Lemoine et al. (1998, Chap. 11). This chapter is also available on-line from:

http://cddis.nasa.gov/926/egm96/doc/S11.HTML

In the same chapter, the issue of expressing the geoid undulations and/or height anomalies with respect to a specific GRS is discussed. In the case of EGM2008, the conversion from an "ideal" mean-Earth ellipsoid (whose semi-major axis remains numerically unspecified), in the Tide-Free system, and the WGS84 GRS, involves the application of a zero-degree height anomaly equal to −41 cm. The zero-degree height anomaly (ζ_z). that was computed when the WGS84 EGM96 geoid was released was equal to −53 cm (Lemoine et al. 1998, Chap. 11). The primary reason for the change in the numerical value of ζ_z from the EGM96 days to the current best estimate, is the discovery by Ouan-Zan Zanife (CLS, France) of an error in the Oscillator Drift correction applied to TOPEX altimeter data (Fu and Cazenave 2001, p. 34). The erroneous correction was producing TOPEX sea surface heights, biased by approximately 12–13 cm.

Under:

http://earth-info.nga.mil/GandG/wgs84/gravitymod/egm2008/egm08_wgs84.html

the user can find a modified version of the HARMONIC_SYNTH program, specifically designed to compute geoid undulations at arbitrarily scattered locations, in the Tide-Free system, with respect to the WGS84 GRS. In the same web site, the user can also find pre-computed global grids of these geoid undulations, at both $1' \times 1'$ and $2.5' \times 2.5'$ grid-spacing, as well as a FORTRAN program to interpolate from these grids.

6.9 Temporal Variations

The topic of temporal gravity field variations, although outside the main scope of the present discussion, cannot be omitted. Non-tidal temporal gravity field variations originate from mass redistribution within the entire solid Earth-Ocean-Atmosphere-Hydrosphere-Cryosphere system. Some of these variations have strong seasonal signals (e.g., variations in the atmosphere and hydrosphere) while others are episodic (e.g., redistribution of mass due to seismic activity). Until recently, Satellite Laser Ranging (SLR) data were the only source of information based on which temporal variations in a handful of very low-degree harmonic coefficients of the gravitational field could be determined (see e.g., Cheng et al. 1997). As a result of the success of the GRACE mission, this situation has changed dramatically in recent years. GRACE offers the capability of constant monitoring of gravitational field variations, with a temporal resolution of approximately 1 month, and a spatial resolution of approximately 400 km. This has opened up an entirely new area of geodetic research and of geodetic contributions towards the establishment of an Earth Observing System, especially in view of its importance in areas related to global Climate Change (e.g., polar ice melting). Under:

http://www.csr.utexas.edu/grace/publications/citation.html

the interested reader can find a plethora of publications involving the use of GRACE data to address a wide variety of science topics.

6.10 Outlook

It is becoming increasingly clear these days that the demarcation between global and regional (or local) gravimetric approximation studies is shifting (if not disappearing altogether). The satellite data that have become available from missions like GRACE and GOCE is prompting some geodesists that used to focus their efforts on local gravimetric studies, to consider also global problems. On the other side, the increasing availability of detailed gravimetric data prompts some global modelers to increase the resolution of their models, effectively "stepping" into the spectral regime that was considered traditionally part of the regional or local approximation work. Mathematical innovations that could facilitate the bridging of any existing gap between these two regimes and provide (better) solutions to some of the problems identified before are therefore highly desirable.

Improvements in gravimetric data coverage and quality are still necessary over vast areas, especially in Antarctica, South America, Africa, and parts of Asia. Airborne gravimetric surveys have provided a wealth of data over remote areas that are very difficult to access and survey otherwise (e.g., Greenland). Such data acquisition techniques currently offer the best means of filling-in the existing gravimetric data gaps.

Innovative analysis techniques have been developed and are constantly being refined. These techniques, and the availability of ever more capable computers, have enabled geodesists to process vast amounts of data on a more or less routine basis these days. But the geodesist's "appetite" for increased accuracy and resolution keeps challenging even some of the most capable computers that are available today.

While some of the traditional geodetic problems may have been solved to a satisfactory degree of accuracy (which is indicative of the progress made within the discipline), the important role of geodesy in the monitoring of the evolving Earth System opens up new possibilities for innovative work. The detection and monitoring of minute changes in the gravitational field is quickly becoming a valuable tool for the study of Climate Change. So, while the character of global gravimetric problems may be changing, new challenges arise, and the future of the discipline seems to this author to be limited only by the imagination and innovation of its practitioners.

Chapter 7
Geoid Determination by 3D Least-Squares Collocation

Carl Christian Tscherning

7.1 Outline of the Chapter

The use of 3D Least-Squares Collocation (LSC) for the determination of a regional or local approximation to the anomalous (gravity) potential as implemented by the GRAVSOFT Fortran programs is described. The software also implements the remove-restore method so that gravity variations outside the region of computation are accounted for by subtracting the contribution of an Earth Gravity Model (EGM) and so that statistical homogenisation is achieved by removing the contribution of topographic short wavelength features. It is also described how LSC may be used to determine parameters like a height datum off-set or to detect possible gross errors. Examples using data from New Mexico, USA, illustrates the use of the method.

7.2 Introduction

The purpose of this chapter is to provide a guide to gravity field modelling, and especially to geoid determination, using least-squares collocation as implemented by the GRAVSOFT Fortran programs (Forsberg and Tscherning 2008). GRAVSOFT includes both programs for 3D and 2D methods for geoid determination. Here we will only consider the 3D methods. The primary difference between 3D and 2D methods is that when using 3D methods no data are "moved" from the surface of the Earth to the ellipsoid or the sphere.

The reader is supposed to be familiar with Parts I–III of this book. However the theory will be reviewed briefly in order to fix the terminology. So, for example when using the term **geoid**, we will mean the quasi-geoid, i.e. the surface having the distance from the ellipsoid equal to the height anomaly, ζ, evaluated at the surface of the Earth.

The general methodology for (regional or local) gravity field modelling is as follows:

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_7,

A: Transform all data to a global geodetic datum (GRS80/WGS84)
B: Convert geoid heights to height anomalies (in regions where orthometric heights are used).
C: Use the remove-restore method.

C1: Remove the effect of a global Earth gravity field model (EGM, a spherical harmonic expansion)
C2: Remove the effect of the topography from the data.

This will produce what we will call *residual* data.

D: Estimate (one or more) empirical covariance function(s) for the residual data in the region in question.
E: Determine an analytic representation of the empirical covariance function(s).
F: Make a homogeneous selection of the data to be used for geoid determination, check for gross-errors (make a contour map of data), verify error estimates of data,
G: Determine using LSC a (regional) residual gravity field approximation. Compute estimates of the residual height anomalies and their errors and of contingent parameters.
H: If the error is too large, and more data is available, add new data and repeat G.
I: Check model, by comparison with data not used to obtain the model.
J: Restore the effect of the EGM and of the residual topography.
K: Convert height anomalies to geoid heights if orthometric heights are used.

The whole process of 3D LSC may be carried through using the GRAVSOFT programs GEOCOL, EMPCOV, TC, TCGRID, COVFIT, SELECT, GEOIP and N2ZETA; see Appendix 4. If 2D LSC or Stokes formulae are to be used, other programs are available; see Fig. 7.1.

GRAVSOFT includes supporting data from New Mexico, USA, which can be used to test the programs and procedures. They have here been used to illustrate the use of LSC. A Python (http://www.python.org) interface has been developed, which has been used in the examples described below. See also Appendix 1

7.3 Theory

The anomalous gravity potential, T, is equal to the difference between the gravity potential W and the so-called normal potential U, $T = W - U$. T is a harmonic function, and may as such be approximated using a global gravity field model (an expansion in solid spherical harmonics), see (3.143).

$$T_{EGM}(r, \bar{\varphi}, \lambda) = GM \sum_{l=2}^{N} \sum_{m=-l}^{l} \bar{C}_{lm} S_{lm}(r, \bar{\varphi}, \lambda) \tag{7.1}$$

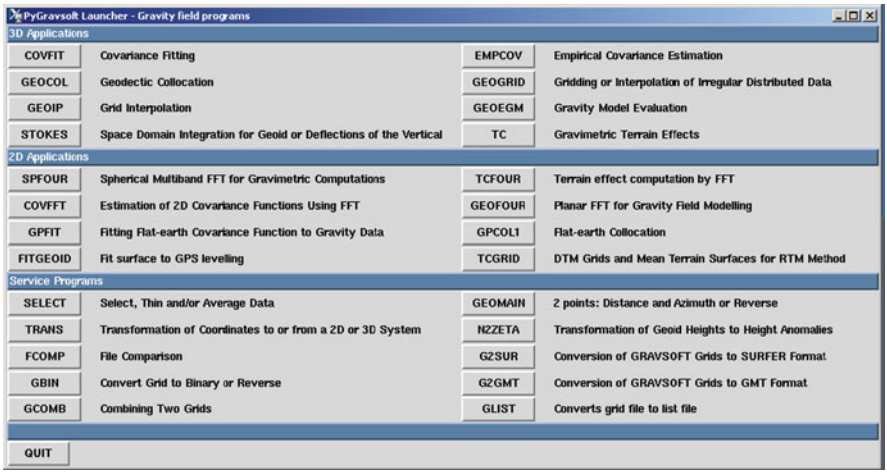


Fig. 7.1 GRAVSOFT Python Launcher, status Feb. 2009

with the solid spherical harmonics

$S_{lm}(r, \bar{\varphi}, \lambda) = \frac{1}{a}\left(\frac{a}{r}\right)^{l+1} Y_{lm}(\bar{\varphi}, \lambda)$, and
GM the product of the mass of the Earth and the gravitational constant
r = distance from origin, $\bar{\varphi}$ = geocentric latitude, λ = longitude,
a = semi-major axis or scale factor, Y_{lm} = normalized surface spherical harmonics,
$\bar{C}_{\mathrm{lm}}$ normalized coefficients with the coefficients of the normal potential U subtracted.

The solid spherical harmonics are orthogonal base-functions in a Hilbert space with an isotropic inner-product, harmonic down to a so-called Bjerhammar-sphere totally enclosed in the Earth see Part III. T will not necessarily be an element of such a space, but may – as mentioned above – be approximated arbitrarily well with such functions.

We will in the following use spherical approximation, i.e. we put the geocentric latitude $\bar{\varphi}$ equal to the geodetic latitude φ and r = $\bar{R}+h$, where $\bar{R}$ is the mean radius of the Earth and h is the ellipsoidal height. If the ellipsoidal height, is unknown, the orthometric height, H, is used. (Note, that if spherical approximation (optionally) is not used then the Bjerhammar sphere must have a radius smaller than the semi-minor axis of the Earth.)

The space will have a reproducing kernel, which is a function of two points in space P, Q. (The coordinates of Q will be distinguished from these of P with an $'$). The kernel will be (see (2.181) and Part III, Sect. 12.4)

$$K(P,Q) = \sum_{l=2}^{\infty} \sigma_l^2 \sum_{m=-l}^{l} S_{lm}(r, \bar{\varphi}, \lambda) S_{lm}(r', \bar{\varphi}', \lambda')$$

$$= \sum_{l=2}^{\infty} (2l+1)\sigma_l^2 \left(\frac{a^2}{rr'}\right)^{l+1} P_l(\cos\psi),$$

where P_l are Legendre polynomials, σ_l^2 are positive constants (degree variances) and ψ the spherical distance from P to Q. (7.2)

(In Part I, the degree-variances are denoted k_l). If the degree-variances are selected equal to simple polynomial functions in the degree l multiplied by exponential expressions like q^l, where $q < 1$ then K(P,Q) may be represented by a closed expression. The most simple example is $\sigma_l^2 = q^l$, where we get well-known expressions related to the reciprocal distance. Below we will describe other simple models for the degree-variances which enables the infinite sum to be evaluated using a closed expression.

In a reproducing kernel Hilbert space one may determine approximations to the elements (here the anomalous potential, T) from data for which the associated linear functionals are bounded. The relationship between the data and T are expressed through linear functionals L_i, (see Part I, Sect. 2.3) so that

$$y_i = L_i(T) + A_i^T X + e_i \tag{7.3}$$

where y_i is the ith data element, L_i the functional, e_i the error, A_i a vector of dimension k and X a vector of parameters also of dimension k. $A_i^T X$ may for example express the effect of a datum-shift or of a bias and tilt in the data, see Part I, Chap. 2.

The GRAVSOFT programs may use or estimate many kinds of gravity data including gravity gradients at satellite altitude (not discussed here) and spherical harmonic coefficients. Here we will only consider

- The height anomaly: $\varsigma = T/\gamma$, where γ is normal gravity,
- The gravity anomaly: $\Delta g = g(P) - \gamma(Q)$,
- The meridian deflection of the vertical: $\xi = \Phi - \varphi$, where Φ is the astronomical latitude,
- The prime-vertical deflection of the vertical: $\eta = (\Lambda - \lambda)\cos(\varphi)$, where Λ is the astronomical longitude,

or mean values of these quantities being represented by an average of point-values. The point Q is a point with latitude and longitude equal to P, but having ellipsoidal height equal to the orthometric height of P. φ is the geodetic latitude if the point is on the ellipsoid. When evaluating gravity anomalies or deflections of the vertical using an EGM, they are computed without using spherical approximation. Deflections are evaluated as the spatial angles between the gravity vector computed from the EGM and the normal field gravity vector at the same point.

The linear functionals are given in Part I, (2.36) and (2.48). However for the gravity anomaly we use the further approximation

$$\Delta g = -\frac{\partial T}{\partial r} - \frac{2}{r}T \tag{7.4}$$

i.e. spherical approximation. (This approximation is only used on the residual quantities, see Sect. 7.4).

An optimal approximation to T using error-free data in a geocentric system may then be obtained using that the observations are given by, $\mathrm{L_i}(T) = \mathrm{y_i}$. The "optimal" solution is the projection on the n-dimensional sub-space spanned by the so-called representers of the linear functionals, $\mathrm{L_i}(\mathrm{K}(\mathrm{P}, \mathrm{Q})) = \mathrm{K}(\mathrm{L_i}, \mathrm{Q})$. The projection is the intersection between the subspace and the subspace which consist of functions which agree exactly with the observations, : $\mathrm{L_i}(T) = \mathrm{y_i}$. (See Part III, Chap. 12).

$$\tilde{T}(P) = \{K(P, L_i)\}^T \{K(L_i, L_j)\}^{-1} \{y_j\} \tag{7.5}$$

If the data contain noise, then the elements σ_{ij} of the variance-covariance matrix of the noise-vector is added to $\mathrm{K}(\mathrm{L_i}, \mathrm{L_j})$. The solution will then both minimize the square of the norm of T and the noise variance. If the noise is zero, the solution will agree exactly with the observations. This is the reason for the name collocation. Upper limits for the approximation error may be calculated if the norm of T is known, (see Tscherning 1985).

If we want to minimize the mean-square error, the reproducing kernel must be selected so that it approximates the so-called empirical covariance function, COV(P,Q). This function is equal to the reproducing kernel given above, and having the degree-variances derived from T equal to

$$\sigma_l^2 = \left(\frac{GM}{\bar{R}}\right)^2 \sum_{m=-l}^{l} (\bar{C}_{lm})^2 \left(\frac{\bar{R}}{a}\right)^{2l+2} \tag{7.6}$$

The selection of the reproducing kernel in this way, also is an implicit selection of the mathematical structure (inner-product) in the Hilbert Space. We will obtain approximations to the anomalous potential with a smoothness resembling the one observed. ***We will in the following use terms from statistical theory even if everything (except random errors) here is deterministic.***

The normal equation matrix may now be expressed using covariances:

$$\bar{C} = \{COV(L_i, L_j) + \sigma_{ij}\} \tag{7.7}$$

The result in terms of predictions is

$$L(\tilde{T}) = \{b_i\}^T \{COV(L, L_i)\} = \{y_j\}^T \bar{C}^{-1} \{COV(L, L_i)\} \tag{7.8}$$

and error estimates

$$\sigma(L)^2 = COV(L, L) - \{COV(L, L_i\}^T \bar{C}^{-1} \{COV(L, L_j)\} \tag{7.9}$$

In the diagonal of the normal equation matrix we find the sum of the data variance $C(L_i, L_i)$ and the noise variance. We can say that we here have a "natural" balance between the signal and the noise. If the observation equation contain parameters, slightly more complicated equations for the prediction and the error-estimate are obtained, see Part I.

The covariances are computed using the "law" of covariance propagation, i.e. COV(L_i, L_j) = $L_i(L_j$(COV(P,Q))), where COV(P,Q) is the basic "potential" covariance function. COV(P,Q) is an isotropic reproducing kernel with the degree-variances given by (7.6).

Example 1. If we want to derive the gravity anomaly covariance function for gravity anomalies in two points P and Q then we must apply the functionals given in (7.4) on $K(P, Q)$. This is

$$
\begin{aligned}
COV(\Delta g(P), \Delta g(Q)) &= \left(-\frac{\partial}{\partial r} - \frac{2}{r} ev_P\right)\left(-\frac{\partial}{\partial r'} - \frac{2}{r'} ev_Q\right) \sum_{i=2}^{\infty} \sigma_i^2 \left(\frac{R^2}{rr'}\right)^{i+1} \\
P_i(\cos\psi) &= \left(-\frac{\partial}{\partial r'} - \frac{2}{r'} ev_Q\right) \sum_{i=2}^{\infty} \sigma_i^2 \frac{i-1}{r} \left(\frac{R^2}{rr'}\right)^{i+1} \\
P_i(\cos\psi) &= \sum_{i=2}^{\infty} \sigma_i^2 \frac{(i-1)^2}{rr'} \left(\frac{R^2}{rr'}\right)^{i+1} P_i(\cos\psi) \qquad (7.10)
\end{aligned}
$$

The quantities $COV(L, L)$, $COV(L, L_i)$ and $COV(L_i, L_j)$ may all be evaluated by the sequence of subroutines COVAX, COVBX and COVCX which form a part of the programs GEOCOL and COVFIT for the funtionals listed above. For further details see Tscherning (1976, 1993).

7.4 The Remove-Restore Method

The least-squares collocation solution is giving the minimum mean square error in a very specific sense, namely as the mean over all data-configurations which, by a uniform rotation around the Earth's centre, may be mapped into each other. So if this should work locally, we must make all areas of the Earth look alike, as see from the gravity field standpoint. This is obviously not possible, but may be achieved to a certain degree due to the use of the remove-restore method.

This is done by removing as much as we know, and later adding it. We obtain a field which is statistically more homogeneous and more smooth than before.

First we may remove the contribution T_{EGM} from a known EGM like the EGM96, complete to degree N = 360 (Lemoine et al. 1998). Secondly we may remove the effect of the local topography, T_M, using Residual Terrain Modelling (RTM, see Forsberg and Tscherning 1981). We will then be left with a **residual** field,

Table 7.1 GRAVSOFT data from the New Mexico Area to be used in LSC geoid examples

Type	Format (see Forsberg and Tscherning 2008)	Error	File-name
Free-air gravity (Fig. 7.6)	Number, latitude, longitude, altitude, anomaly	0.2 mgal	nmfa
Deflections of the vertical	Number, latitude, longitude, altitude, ξ, η	0.5 arcsec	nmdfv
Height anomalies	Number, latitude, longitude, altitude, ζ	0.02 m	nmzeta
Digital terrain model	Grid-label, data in North-South, East-West	5 m	nmdtm

with a smoothness in terms of standard deviation of gravity anomalies between 50% and 25% less than the original standard deviation, see Table 7.2. The removal and later restoration of the contribution from T_{EGM} has furthermore the effect, that gravity field information outside the data-area is implicitly accounted for. It also has the effect that the covariance functions will have a smaller correlation distance as compared to the global covariance function, compare (Tscherning and Rapp 1974, Fig. 7.1) and Fig. 7.6, thus making the solution of the normal-equations in (7.8) more stable.)

The residual quantities are then

$$y_{ir} = y_i - L_i(T_{EGM}) - L_i(T_M) = L_i(T) - L_i(T_{EGM}) - L_i(T_M) + e_i + A_i^T X \quad (7.11)$$

Example 2. We compute residual gravity and height anomalies using the EGM96 spherical harmonic expansion and the New Mexico DTM, cf. Table 7.1. The free-air gravity anomalies are shown in Fig. 7.2.

The program GEOCOL is used to subtract the contribution from EGM96 using the Python interface module GEOEGM, see Fig. 7.1. The coefficients of the EGM96 model are found in the file data/EGM96, see Appendix 1. The difference file is denoted nmfa-egm96.dat for the free-air anomalies and nmzeta-egm96.dat for the height anomalies. The gravity differences are shown in Fig. 7.3. Note the increased smoothing compared to Fig. 7.2.

The RTM contribution is computed and subtracted using the program TC. First a reference terrain model must be constructed using the program TCGRID with the file nmdtm, cf. Table 7.1, as basis. Such files are stored as nmdtm5 and nmdtm30. The results are stored in files named nmfa-egm96-tc.dat and nmzeta-egm96-tc.dat. The residual gravity anomalies are shown in Fig. 7.4. The results are summarized in Table 7.2.

The degree-variances will be changed up to the maximal degree, N, of the spherical harmonic series (in the Examples N = 360), contingently to a smaller value, if the series is not agreeing well with the local data (i.e. if no or little data in the area were used when the series were determined). The first of N new degree-variances will depend on the error of the coefficients of the series. We will here suppose that the degree-variances at least are proportional to the so-called error-degree-variances so that for i = 2, . . ., N we have

$$(\sigma_l^{err})^2 = \alpha \left(\frac{GM}{a}\right)^2 \sum_{m=-l}^{l} (\sigma_{lm}^{EGM})^2 \left(\frac{a}{\bar{R}}\right)^{2l+2} \quad (7.12)$$

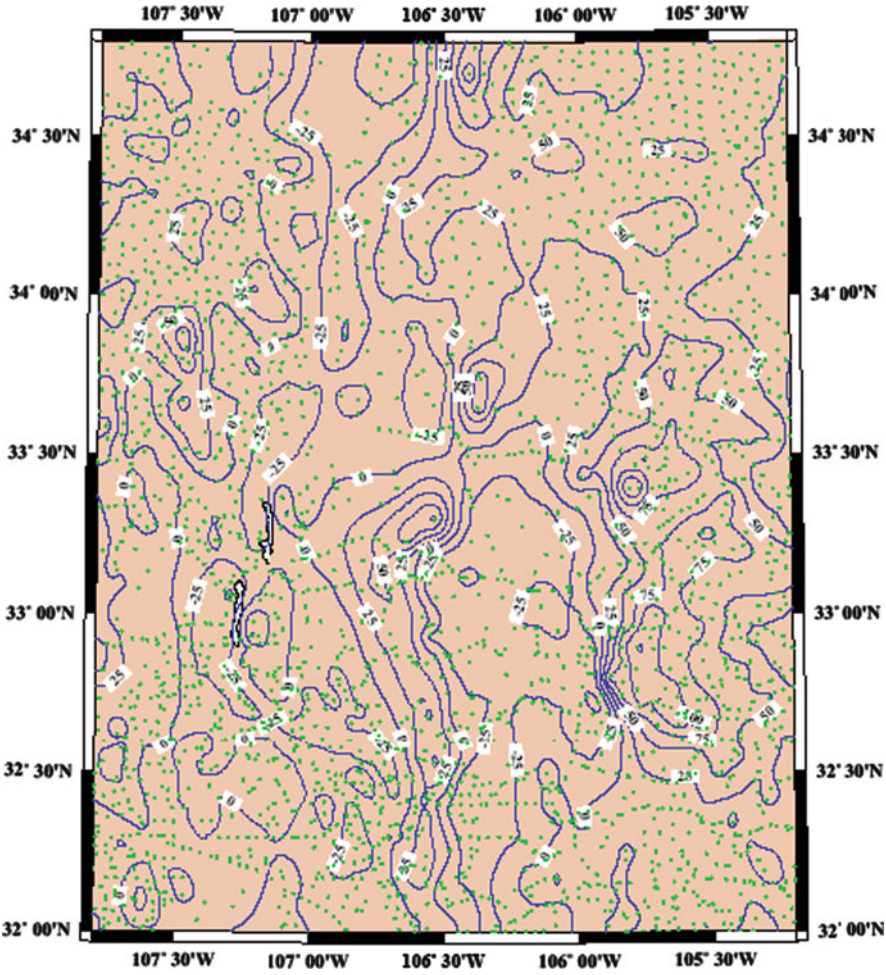


Fig. 7.2 Free-air gravity anomalies/mgal

Note that the error-degree variances refer to the mean-earth sphere. They may be evaluated using the GRAVSOFT program degv.for (not shown in Fig. 7.1), which will produce the degree-variances for gravity anomalies in units of mgal^2. (Error gravity anomaly degree-variances for EGM96 are found in the file data/egm96.edg, cf. Appendix 1.) The reason why gravity error-degree variances are used is that these quantities express how much gravity anomaly power is left within a certain degree after having subtracted the EGM. For EGM96 it is 0.2 mgal^2 at degree 360.

The scaling factor α must be determined from the data (in the program COVFIT, see later).

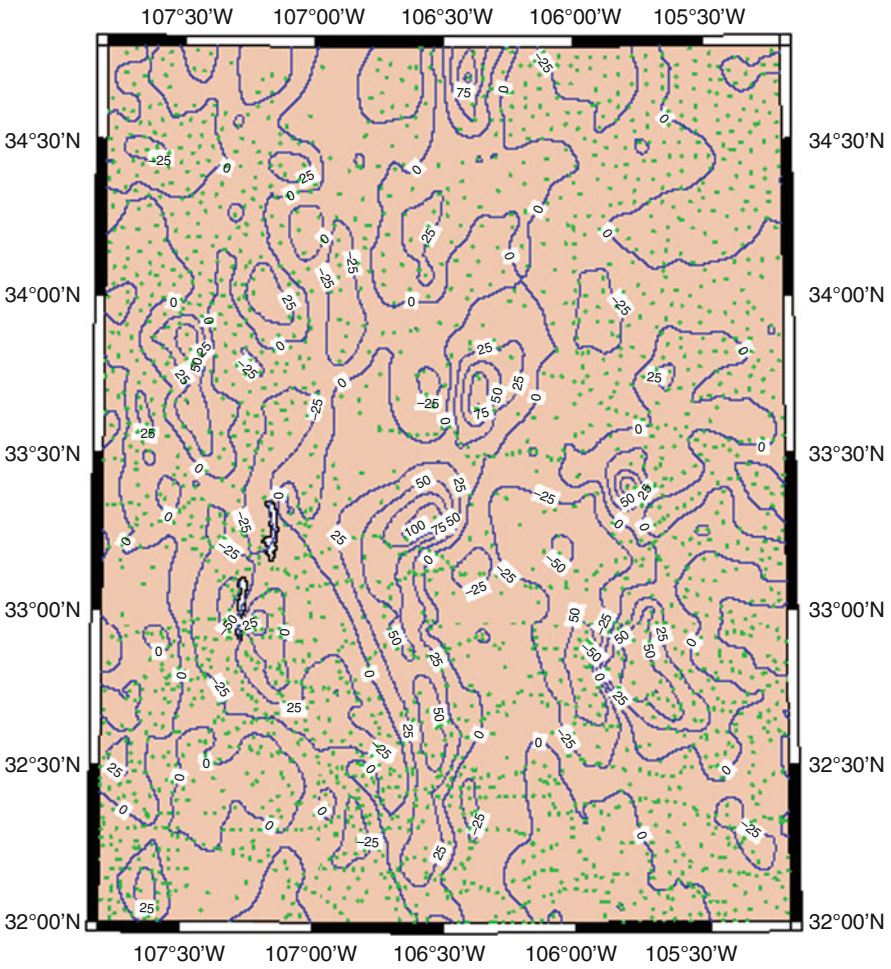


Fig. 7.3 Gravity anomalies – EGM96/mGgal

7.5 Covariance Function Estimation and Representation

The global covariance function used in LSC is equal to a triple integral

$$COV(P, Q) = \frac{1}{8\pi^2} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} T(P)T(Q)\, d\alpha \cos\varphi \; d\varphi \, d\lambda \tag{7.13}$$

where α is the azimuth between P and Q and φ, λ are the coordinates of P. The point Q has a fixed spherical distance from P. Note that this is a global expression,

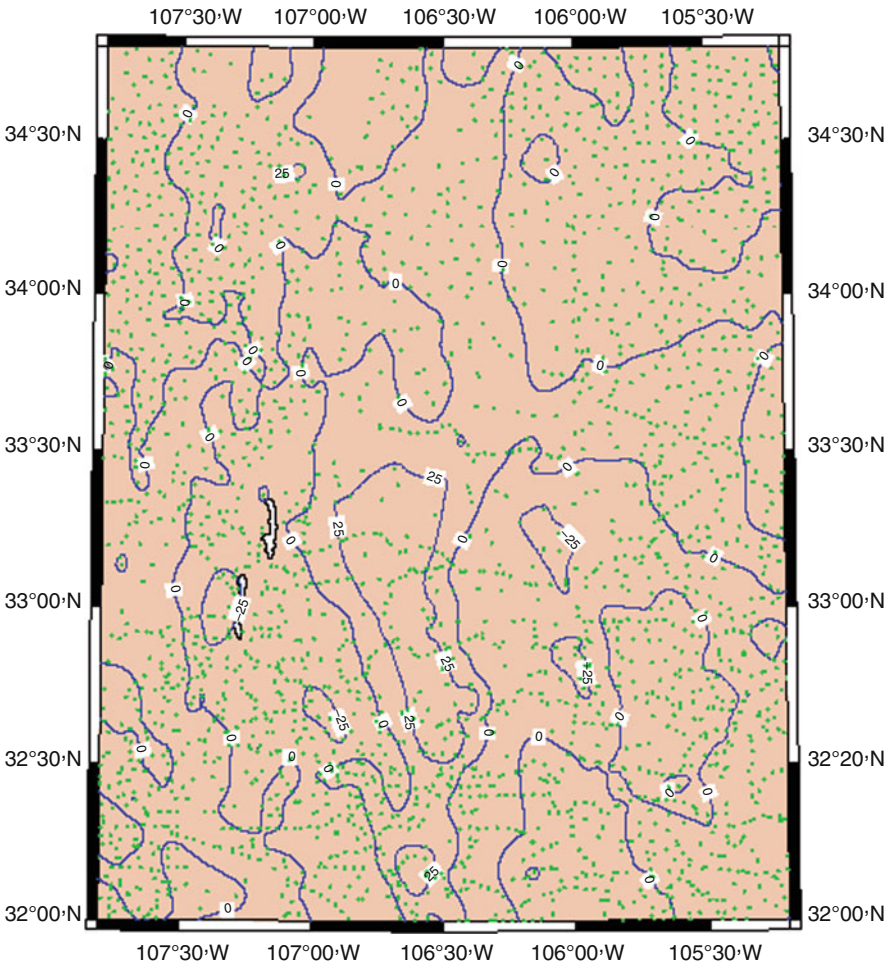


Fig. 7.4 Gravity anomalies – EGM96 – TC/mGgal

and that it will only depend on the radial distances r, r′ of P and Q and of the spherical distance ψ between the points. We have in the program COVFIT used the original definition of the covariance function as an integral in order to determine weights of estimated empirical covariances as a function of the mean data spacing (see Knudsen 1987a). In COVFIT this is done by giving as input to the program the boundaries of the data area and the data-spacing. From this the optimal number of products (see (7.15)) used to evaluate the integral may be compared to the actual number of products (M in (7.15)).

In practice the covariance function must be estimated from the kind of data we have most of: gravity anomalies; but sometimes also height anomalies are available, cf. Table 7.1. But as seen from (7.10), we may determine from a gravity

Table 7.2 Statistics of residual quantities. We have also included a comparison with EGM2008 (Pavlis et al. 2008)

Δg (mgal)	Mean	Standard dev.	Minimum	Maximum
Original data	9.2	30.4	−58.7	162.5
-EGM96	−2.9	21.3	−74.8	126.4
-EGM96-TC	0.3	13.1	−41.0	45.0
-EGM08	−2.2	7.9	−44.6	70.1
ζ (m)				
Original data	−24.27	1.08	−20.92	−25.06
-EGM96	0.04	0.16	−0.30	0.34
-EGM96-TC	0.20	0.14	−0.09	0.42
-EGM08	0.21	0.04	0.14	0.28

anomaly covariance function the basic covariance function by multiplying the degree-variances by $((\mathrm{i}-1)/\mathrm{R})^2$.

In a local area we will implicitly regard all data outside the area as having the same statistical characteristics as the data in the area, so that we may estimate the gravity anomaly covariance function by taking a sum of products of the data in the area grouped according to an interval $\Delta\psi$ of spherical distance (also denoted the sampling interval size),

$$\psi_i - \Delta\psi/2 \;\le\; \psi \;\le\; \psi_i + \Delta\psi/2 \tag{7.14}$$

Note that two intervals may be merged, so that the sampling interval becomes the double. In the program EMPCOV this may be done a number of times, as an aid in selecting the right size of the sampling interval. Hence the estimated covariance is

$$COV_{est}(\psi_i, r_m, r_m) = \frac{1}{M}\sum_{n=1}^{M} \Delta g(P)\Delta g(Q) \tag{7.15}$$

where M is the number of products from the ith sampling-interval and r_m is the mean altitude. In the calculations the covariance will be regarded as referring to the mean height, which for the New Mexico dataset (Table 7.1) is approximately 1, 700 m.

Example 3. (To be evaluated using a hand-held calculator!) We compute a table of empirical covariance values using the data in Appendix 1. Note the location of the first zero-point for the covariance function and the correlation distance (the distance to where the covariance first time becomes half the variance (and we have 50% correlation).

We then predict using collocation the gravity anomaly in a point with latitude 56.65 and longitude 10.0 from point 18 only, and then from points 17 and 18. We regard the data as error-free. We also compute the error-estimate. The result is also in Appendix 1.

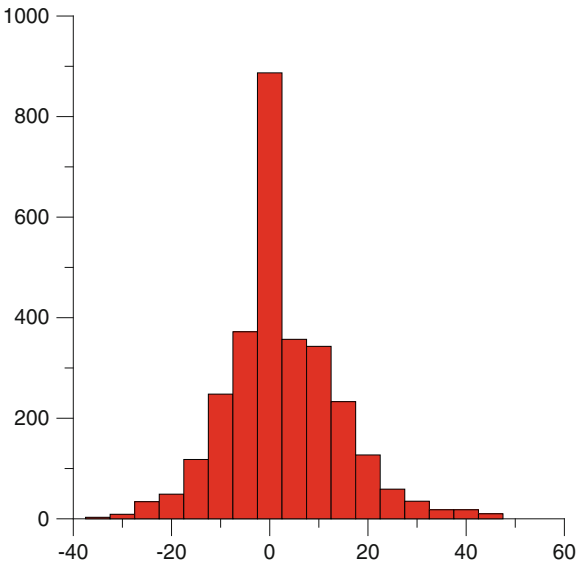


Fig. 7.5 Histogram of the 2,920 residual free-air gravity anomalies in 5 mgal bins

Example 4. We compute the empirical gravity anomaly covariance function using the program EMPCOV for the New Mexico area for the anomalies from which both the contribution of EGM96 and the RTM-effects have been subtracted (input file nmfa-egm96-tc.dat). A sampling interval of 2.5 arcmin is used. The estimated covariances are shown in Fig. 7.5. In the output from EMPCOV is also written the data boundaries and the mean spacing. For the data used it is minimum and maximum latitude of 31.68 and 34.81 degrees and minimum and maximum longitude of −107.82 and −105.19 degrees. The mean altitude is 1614.9 m, and the mean linear spacing is 0.049 degrees. These numbers will be used in the program COVFIT to determine the number of products as mentioned above.

Take note of the correlation distance ψ_1, i.e. the distance where the covariance becomes equal to 50% of the variance $C_0 = COV(0, r, r)$.

We see from (7.10), that if we can find the gravity anomaly degree-variances, we also can find the potential degree variances. However, we also see that we need to determine infinitely many quantities in order to find the covariance function.

The solution to this problem is to use a so-called degree-variance model, i.e. a functional dependence between the degree and the degree-variances. In the program COVFIT, three different models (1, 2 and 3) may be used. The main difference is related to whether the (potential) degree-variances go to zero like n^{-2}, n^{-3} or n^{-4}. The best model (see Tscherning and Rapp 1974 and Appendix 1) is of the type 2,

$$\sigma_n^2 = \frac{A}{(n-1)(n-2)(n+B)} \left(\frac{R_B}{\bar{R}} \right)^{2n+2} \tag{7.16}$$

WHERE R_B is the radius of the Bjerhammar-sphere, A is a constant in units of $(m/s)^4$, B an integer. If a spherical harmonic series expansion (EGM) is used, B is typically put equal to a small number like 4, while in the original work it was put equal to 24, so that the low-degree degree-variances could be modelled appropriately. This model is simultaneously a Reproducing Kernel in a Hilbert Space of functions harmonic outside the Bjerhammar-sphere, see Part III, Sect. 12.4.

The complete model used is

$$COV(\psi, r, r') = \alpha \sum_{n=2}^{N} (\sigma_i^{err})^2 \left(\frac{\bar{R}^2}{rr'}\right)^{n+1} P_n(\cos\psi) + \sum_{n=N+1}^{\infty} \frac{A}{(n-1)(n-2)(n+4)} \left(\frac{R_B^2}{rr'}\right)^{n+1} P_n(\cos\psi) \quad (7.17)$$

The actual modelling of the empirically determined values is done using the program COVFIT. The factors α, A and R_B need to be determined (and the first index N + 1 must be fixed). (However, instead of the factor A the gravity anomaly variance at zero altitude $C_0(\Delta g)$ is used, because this quantity is more meaningful to the user.)

The program makes an iterative non-linear adjustment for the Bjerhammar-sphere radius, and linear for the two other quantities (see Knudsen 1987a). Unfortunately sometimes the iteration may behave irregularly (e.g. result in a Bjerhammar-sphere radius larger than R). This may occur, if the data has a very inhomogeneous statistical character. Therefore simple histograms are always produced together with the covariances (in EMPCOV) in order to check that the data distribution is reasonably symmetric, if not normal, see Fig. 7.6.

Example 5. We compute using COVFIT an analytic representation for the empirical covariance function.

Gravity error-degree-variances for the EGM96 coefficients are found in the file data/egm96.edg. The estimated and the fitted covariance values are shown in Fig. 7.6. The resulting values are $\alpha = 0.2837$, the depth to the Bjerhammar-sphere of 792.72 m and the gravity variance C_0 at zero altitude equal to 334.36 $mgal^2$. These values are used when running GEOCOL in the following examples.

COVFIT may also be used to tabulate the analytic covariance function. The result corresponding to Example 5 are found in Table 7.3.

The numerical evaluation of the expression for the covariance function is rather time-consuming because it involves the summation of a Legendre-series up to degree N, in the examples equal to 360. However, the covariances of geoid heights, gravity anomalies and deflections of the vertical may be tabulated. This option is available in COVFIT and in GEOCOL. The selection of the optimal table entries is complicated. The solution recommended for the moment is trial and error. The tool here is COVFIT, which may compute the differences between tabulated and

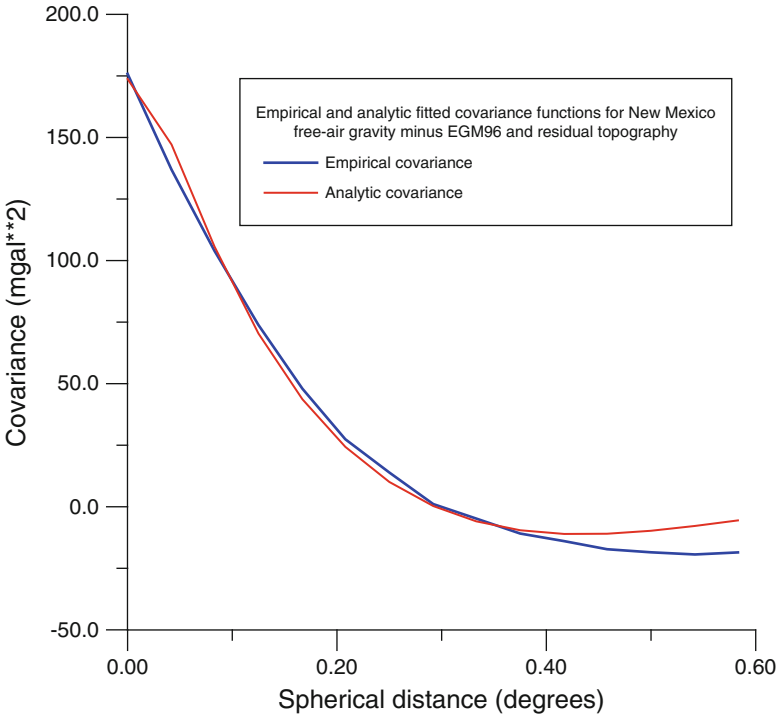


Fig. 7.6 Empirical and analytic fitted covariance functions

correctly evaluated quantities. (Unfortunately the Python interface cannot be used here).

The covariances may also be tabulated in 1-dimension, i.e. only as a function of spherical distance. This is very useful if we deal with data at a constant altitude, e.g. ocean data (at height zero).

7.6 Conversion from Geoid Heights to Height Anomalies

The conversion of geoid heights N_{P_0} to height anomalies ζ_P at altitude H is needed in 3D LSC, because all quantities must be related to points outside the masses. An approximate equation is given in Part I, (2.74).

$$N_{P_0} - \zeta_P \approx \frac{\Delta g_B}{\gamma_0} H \tag{7.18}$$

It involves the Bouguer anomaly Δg_B in the same points as the geoid heights, and this quantity must be determined by prediction from other gravity data. The GRAVSOFT program GEOCOL may be used to do this, applying the analytic

Table 7.3 Table of model-covariances of height anomalies, gravity anomalies and deflections of the vertical. Height 1,700 m. Model degree-variances equal to A/((i−1)(i−2)(i+4)). Error degree-variances used from degree 2 to 360 with scale factor 0.2555. Error degree variances from EGM96. Depth to Bjerhammar − sphere = −819.00 m, variance of point gravity anomalies at 0 height $C_0(\Delta g) = 335.41\,\text{mgal}^2$, the factor A, divided by R_E^2 is = 452.52 mgal^2

KP =	1	3	3	6	6	6
KQ =	1	3	1	6	3	1
ψ(deg.)	m^2	$mgal^2$	m*mgal	$arcsec^2$	arcsec*mgal	arcsec*m
0.00	0.0476	174.15	2.058	3.878	0.000	0.000
0.05	0.0463	139.00	1.885	2.749	−11.167	−0.092
0.10	0.0430	90.20	1.535	1.307	−13.884	−0.146
0.15	0.0387	53.43	1.167	0.318	−13.132	−0.167
0.20	0.0342	27.60	0.837	−0.309	−11.188	−0.167
0.25	0.0298	10.08	0.566	−0.678	−8.886	−0.153
0.30	0.0260	−1.15	0.358	−0.860	−6.597	−0.132
0.35	0.0227	−7.65	0.209	−0.907	−4.512	−0.108
0.40	0.0201	−10.61	0.112	−0.857	−2.730	−0.084
0.45	0.0182	−11.03	0.058	−0.744	−1.300	−0.062
0.50	0.0167	−9.71	0.038	−0.593	−0.233	−0.044
0.55	0.0157	−7.36	0.043	−0.428	0.483	−0.030
0.60	0.0150	−4.53	0.062	−0.264	0.885	−0.021
0.65	0.0146	−1.66	0.088	−0.117	1.019	−0.016
0.70	0.0141	0.91	0.115	0.004	0.939	−0.014
0.80	0.0133	4.40	0.152	0.153	0.377	−0.019
0.90	0.0120	5.20	0.152	0.174	−0.353	−0.028
1.00	0.0102	3.75	0.117	0.100	−0.910	−0.036
1.10	0.0081	1.11	0.061	−0.014	−1.117	−0.038
1.20	0.0061	−1.51	0.003	−0.121	−0.961	−0.035

covariance function determined in Example 5. A GRAVSOFT module N2ZETA may then be used to make the conversion from geoid heights to height anomalies.

7.7 LSC Geoid Determination from Residual Data

We now have discussed all the tools available for using LSC: residual data and a covariance model. The rest is to establish the normal equations (7.7), solve the equations, (7.8) and compute predictions and error estimates, (7.8) and (7.9). This is done using GEOCOL.

However, as realized from (7.7) we have to solve a system of equations as large as the number of observations. This is one of the key problems with using the LSC method. The problem may be reduced by using mean values of data in the border area. Also, if the observations are clustered, we may not need all observations. Rules for the necessary data density (d) as a function of the correlation length ψ_1 of the covariance function are given in Tscherning (1985, p. 330). Suppose we want

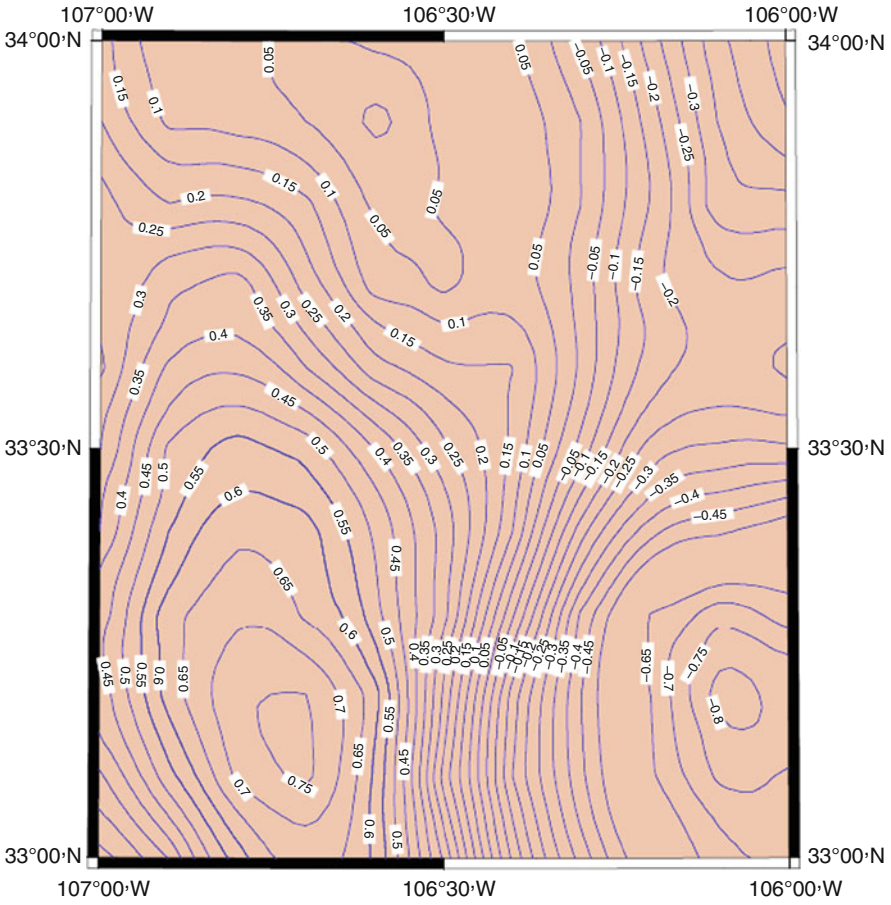


Fig. 7.7 Error estimates of reduced height anomalies calculated from gravity only

to determine geoid height differences with an error of 10 cm over 100 km. This corresponds to an error in deflections of the vertical of 0.2". This is equivalent to that we must be able to interpolate gravity anomalies with a mean error of 1.2mgal. The rule-of-thumb for the square of the error is

$$e_d^2 \approx C_0(d \cdot 0.3/\psi_1)^2 \tag{7.19}$$

Example 6. We use the residual gravity variance $C(0, r_m, r_m) = 175\,\text{mgal}^2$ and the correlation distance determined in Example 4 of $\psi_1 = 10\,\text{km}$ for the determination of the needed data spacing. A mean spacing of about 3 km is needed to obtain the result. In reality we have too few data in the dataset denoted nmfa, which has a mean spacing of 5 km. From Fig. 7.7 we see that the estimated error is between 0.07 and 0.13 m.

If we have more data than deemed necessary the program SELECT may be used for the selection of points as close a possible to the nodes of a grid having the required data spacing, and which covers the area of interest. The area covered should be larger than the area in which the geoid is to be computed. Data in a distance at least equal to the distance for which gravity and geoid becomes less than 10% correlated, (cf. Table 7.3) should be included.

When data have been selected it is recommended to prepare a contour plot of the data such as Fig. 7.1. This will show whether the data should contain any gross-errors. LSC may also be used for the detection of gross-errors, see Tscherning (1991).

An input file for the program GEOCOL must then be prepared, or the program may be run interactively or using the Python interface. However, in order to compute height-anomalies at terrain altitude, a file with points consisting of number, latitude, longitude and altitude must be prepared. This may be prepared using the program GEOIP, and input from a digital terrain model.

Example 7. We prepare a file named nm.h covering the area bounded by 33.0° and 34.0° in latitude and −107.0° and −106.0° in longitude consisting of sequence number, latitude, longitude and height given in a grid with 0.1 degree spacing. The program GEOIP is used with input from nmdtm. This will produce a grid-file. This must be converted to a standard point data file (named data/nm.h2) using the program GLIST.

When using GEOCOL the following must be specified (see Appendix 3)

The coordinate system used (GRS80),
The constants defining the covariance model and contingently its tabulation
The input data files (nmfa-egm96-tc.dat or nmzeta-egm96-tc.dat)
The files containing the points in which the predictions should be made (nm.h2).

Several options must be selected such as whether error-estimates should be computed or whether we want statistics to be output.

Example 8. We run the program GEOCOL (geocol17) with the selected gravity data for the prediction of the height anomalies in the file nmzeta-egm96-tc.dat, and compare the input and the predicted values. The result is found in Table 7.4.

The table shows that the height-anomalies have a large off-set. We may determine this by adding the data as additional observations, and use this to determine a bias value. The reason for the bias is the difference in semi-major axis of GRS80 and the scale factor (semi-major axis) of EGM96 which is 0.7 m. The remaining part is possibly due to a vertical datum off-set caused by sea-surface topography.

Example 9. We run again the program GEOCOL, but now with additional data from the file nmzeta-egm96-tc.dat, and re-use the already reduced normal equations. We define that one bias parameter must be determined. (This is done automatically if the Python interface is used, see Appendix 4). We predict residual height anomalies in the points of the file nm.h2, see Fig. 7.8, and add back the contribution from the

Table 7.4 Results of predicting reduced height anomalies from reduced gravity anomalies and from both reduced gravity anomalies and height anomalies. In the last case, a bias was estimated also (see last row), which made the error-estimates of the calculated height anomalies larger. The error of the height anomalies was set to 0.02 m (All units m)

Height anomaly (m) prediction from gravity	Observed residual	Predicted	Difference	Error estimate
Mean	−0.897	0.000	−0.897	0.057
Standard deviation	0.159	0.146	0.052	0.015
Maximum	−0.633	0.210	−0.789	0.091
Minimum	−1.267	−0.320	−0.987	0.046
From gravity and height anomalies				
Mean		−0.897	0.000	0.082
Standard deviation		0.147	0.040	0.000
Maximum		−0.672	0.062	
Minimum		−1.218	−0.062	
Estimated bias		−0.918		0.041

topography and EGM96. The total quasi-geoid is shown in Fig. 7.9. We also use the solution to predict the height anomalies in the observation points, cf. Table 7.4.

When the LSC-solution has been made, the RTM contribution to the geoid must be determined. Here the program tc may be used with the file nm.h2 defining the points of computation. The LSC determined residual geoid heights and the associated error-estimates (computed from gravity only) are shown in Figs. 7.7 and 7.8. Figure 7.7 shows the corresponding error-estimates after height-anomalies have been added.

If mean gravity anomalies, deflections of the vertical or GPS/levelling determined geoid-heights were to be used, they could easily have been added to the data. It would not be necessary to recalculate the full set of normal-equations. Only the columns related to the new data need to be added. Likewise, an obtained solution may be used to calculate such quantities and their error-estimates (Fig. 7.10).

The use of deflections and geoid heights (e.g. from satellite altimetry) may require that parameters such as datum shifts and bias/tilts are determined. These possibilities are also included in GEOCOL (see Tscherning 1985).

Example 10. We want to detect suspected gross-errors by comparing the differences between observed quantities and predicted quantities to the estimated error. We use GEOCOL (cf. Example 8) for this purpose by predicting reduced gravity anomalies (data/nmfa-egm96-tc0.dat) and comparing these with values predicted from an identical file, but named data/nmfa-egm96-tc.dat. A file name for a file to hold suspected gross errors must be input. The Python interface for finding gross-errors is found in Appendix 4.

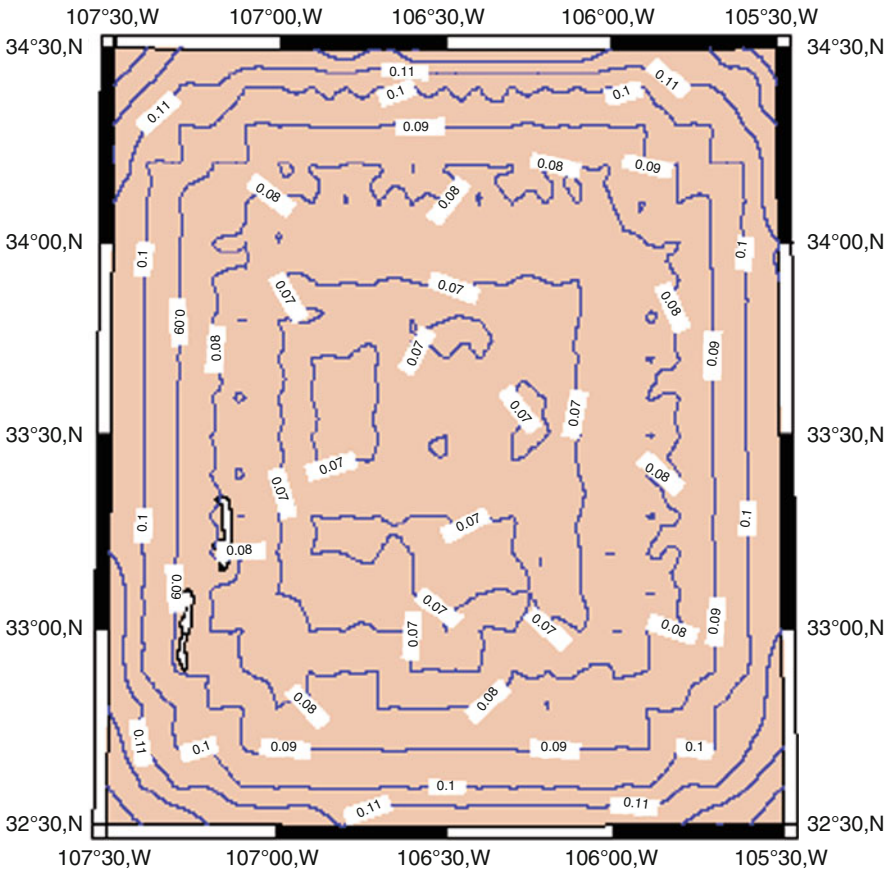


Fig. 7.8 Reduced height anomalies (units: m)

7.8 Conclusion

We have now gone through all the steps from data to predicted height anomalies or geoid heights. The examples describes the use of gravity data only and GPS/levelling derived height anomalies. Deflections of the vertical and gravity disturbances (see Tscherning et al. 2001) could have been used as well.

The difficult steps in the application of LSC is the estimation of the covariance function and subsequent selection of an analytic representation.

The flexibility of the method is very useful in many circumstances, and is one of the reasons why the method has been applied in many countries. If the reference spherical harmonic expansion is of good quality, only a limited amount of data outside the area of interest are needed in order to obtain a good solution. But if this is not the case, data from a large border-area must be used so that a vast

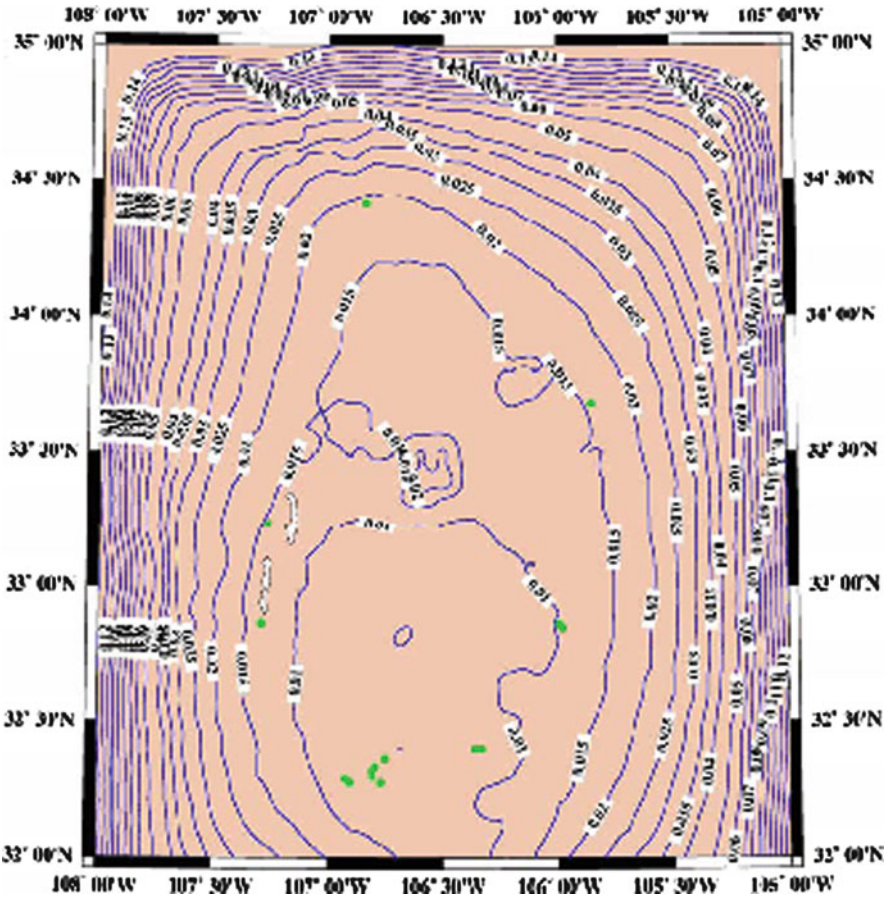


Fig. 7.9 Total height anomalies/m

computational effort is needed to obtain a solution. This may make it impossible to apply the method.

A way out is then to use the method only for the determination of gridded values, which then may be used with Fourier transform techniques (Schwarz et al. 1990) or Fast Collocation (Bottoni and Barzaghi 1993). Also the use of multiple processors is feasible, see Tscherning and Veicherts 2007.

Acknowledgements The original version of this chapter was prepared for the International School for the determination and Use of the Geoid, Milano, Oct., 1994. Thanks to the US National geodetic Survey for the permission to use the New Mexico data. Thanks to F. Sansò for valuable comments.

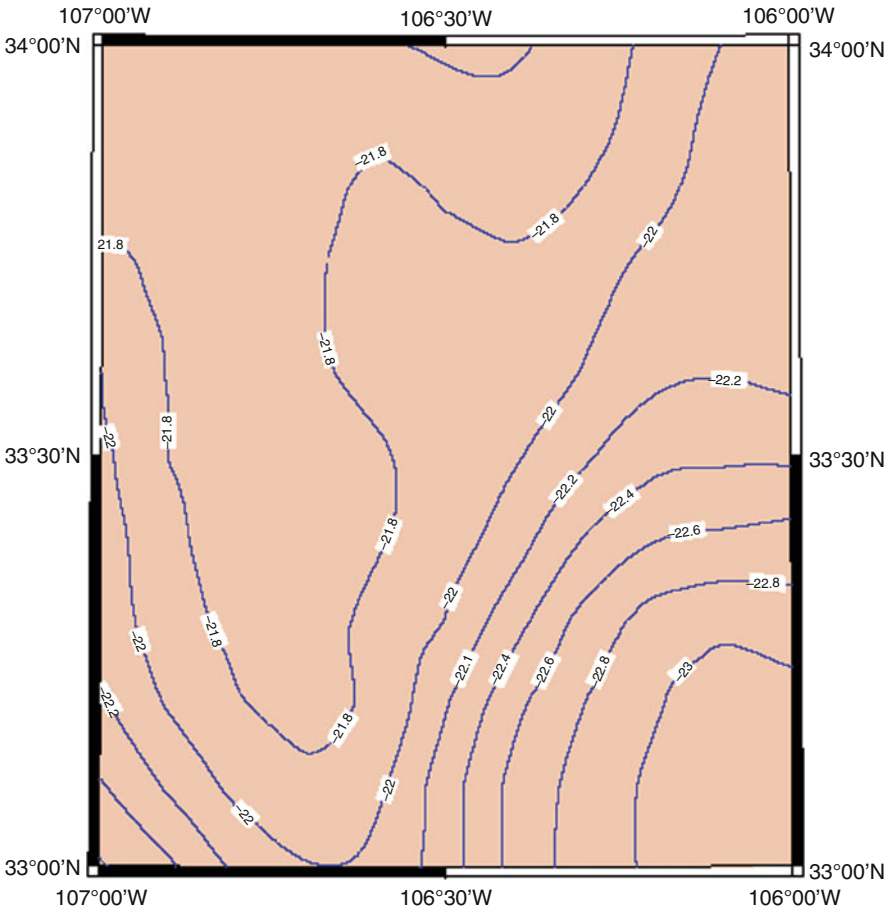


Fig. 7.10 Error estimates from gravity and height-anomalies

Appendix 1 GRAVSOFT Software and Data Organisation

The GRAVSOFT software and test-data are available free-of-charge for scientific and educational purposes. A Python interface has been developed to aid the user, who do not require the use of all options in the programs.

The Python modules are in a root-directory denoted pyGravsoft.

The New Mexico data (Table 7.1), the EGM coefficients and the associated error-degree-variances are stored in a sub-directory denoted "data". The Fortran programs are stored in a sub-directory "src" and the compiled Windows executables in another sub-directory "bin". Program documentation are stored in a sub-directory "doc".

The Python modules will generate a file with input data denoted <program-name>.inp and the screen output will besides being send to the screen be stored

in a file <program-name>.log. The results presented in all the numerical examples described in the chapter have been produced using the Python interface.

Appendix 2 Data and Result of Example 2

The following data is used, with format: number, latitude, longitude (degrees), altitude (m) and gravity anomaly in mgal.

11	56.0	10.0	0.0	4.0
12	56.1	10.0	0.0	2.0
13	56.2	10.0	0.0	0.0
14	56.3	10.0	0.0	−2.0
15	56.4	10.0	0.0	−4.0
16	56.5	10.0	0.0	−6.0
17	56.6	10.0	0.0	−8.0
18	56.7	10.0	0.0	−9.0
19	56.8	10.0	0.0	−7.0
20	56.9	10.0	0.0	−5.0
21	57.0	10.0	0.0	−3.0
22	57.1	10.0	0.0	−1.0
23	57.2	10.0	0.0	1.0
24	57.3	10.0	0.0	5.0
25	57.4	10.0	0.0	4.0

The resulting covariances are given in the following table:

Ψ		Covariance	
O	‘	$mgal^2$	Number of products
0	0.0	23.13	15
0	6.0	21.43	14
0	12.0	16.69	13
0	18.0	10.67	12
0	24.0	3.36	11
0	30.0	−4.40	10
0	36.0	−11.44	9
0	42.0	−15.25	8

We see that the correlation distance is $16'$ and that the first zero-point is located at $\psi = 27$. The mean value of −1.9 mgal has not been removed from the data.

We now predict the gravity anomaly with the location $\varphi = 56.65$ deg. and same longitude as the other points. The distance to point 18 is 0.05 deg. or $3'$, so the

covariance (obtained by linear interpolation) is $22.28\,\mathrm{mgal}^2$. Consequently, using (7.8) we have

$$\Delta g(\varphi = 56.65) = \mathrm{cov}(0.05\mathrm{deg.}) \bullet \; y(18)/cov(0.0) = \\ 22.28 \bullet (-9.0)/23.13 = -8.67\mathrm{mgal}.$$

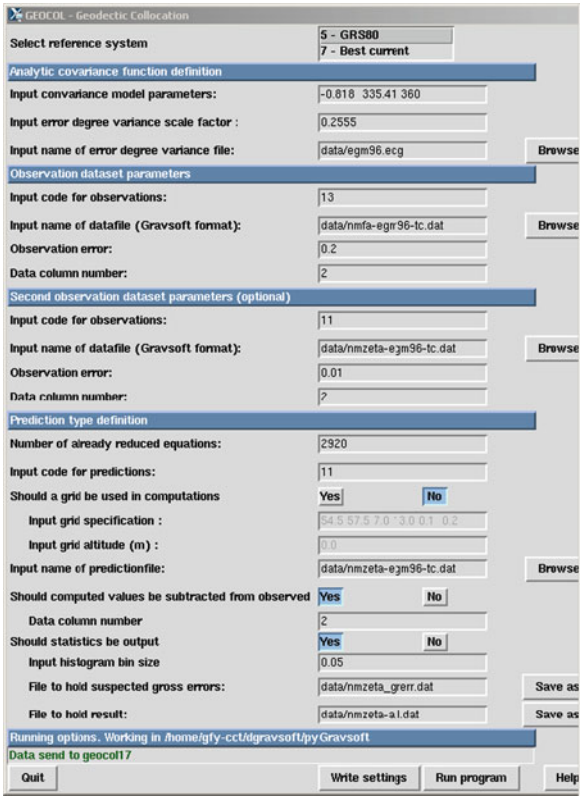
Adding another point (7.17) we obtain

$$\Delta g(\varphi = 56.65) = \begin{Bmatrix} \mathrm{cov}(0.05\,\mathrm{deg}) \\ cov(0.05\mathrm{deg}) \end{Bmatrix}^T \cdot \begin{Bmatrix} \mathrm{cov}(0.0) & \mathrm{cov}(0.05) \\ \mathrm{cov}(0.05) & \mathrm{cov}(0.0) \end{Bmatrix}^{-1} \cdot \begin{Bmatrix} -8.0 \\ -9.0 \end{Bmatrix} = \\ \begin{Bmatrix} 22.28 \\ 22.28 \end{Bmatrix}^T \cdot \begin{Bmatrix} 23.23 & 21.43 \\ 21.43 & 23.23 \end{Bmatrix}^{-1} \cdot \begin{Bmatrix} -8.0 \\ -9.0 \end{Bmatrix} = \\ \begin{Bmatrix} 22.28 \\ 22.28 \end{Bmatrix}^T \cdot \begin{Bmatrix} 0.2890 & -0.2666 \\ -0.2666 & 0.2890 \end{Bmatrix} \cdot \begin{Bmatrix} -8.0 \\ -9.0 \end{Bmatrix} = \begin{Bmatrix} 0.5 \\ 0.5 \end{Bmatrix}^T \begin{Bmatrix} -8.0 \\ -9.0 \end{Bmatrix} = -8.5\mathrm{mgal}$$

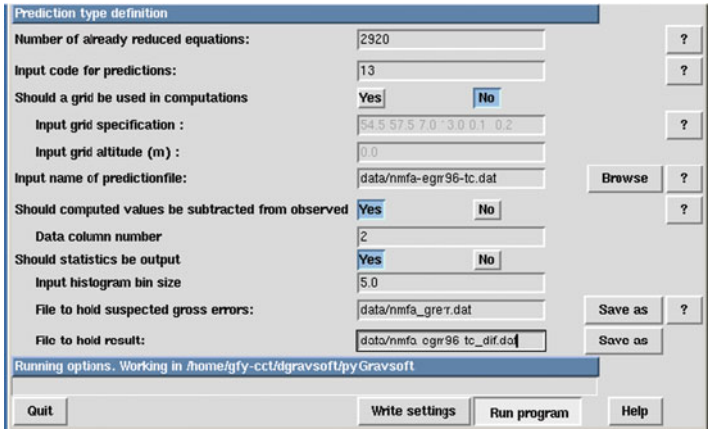
For the error-estimate we get

$$23.23 - \begin{Bmatrix} 0.5 \\ 0.5 \end{Bmatrix}^T \cdot \begin{Bmatrix} 22.28 \\ 22.28 \end{Bmatrix} = 0.95\mathrm{mgal}^2$$

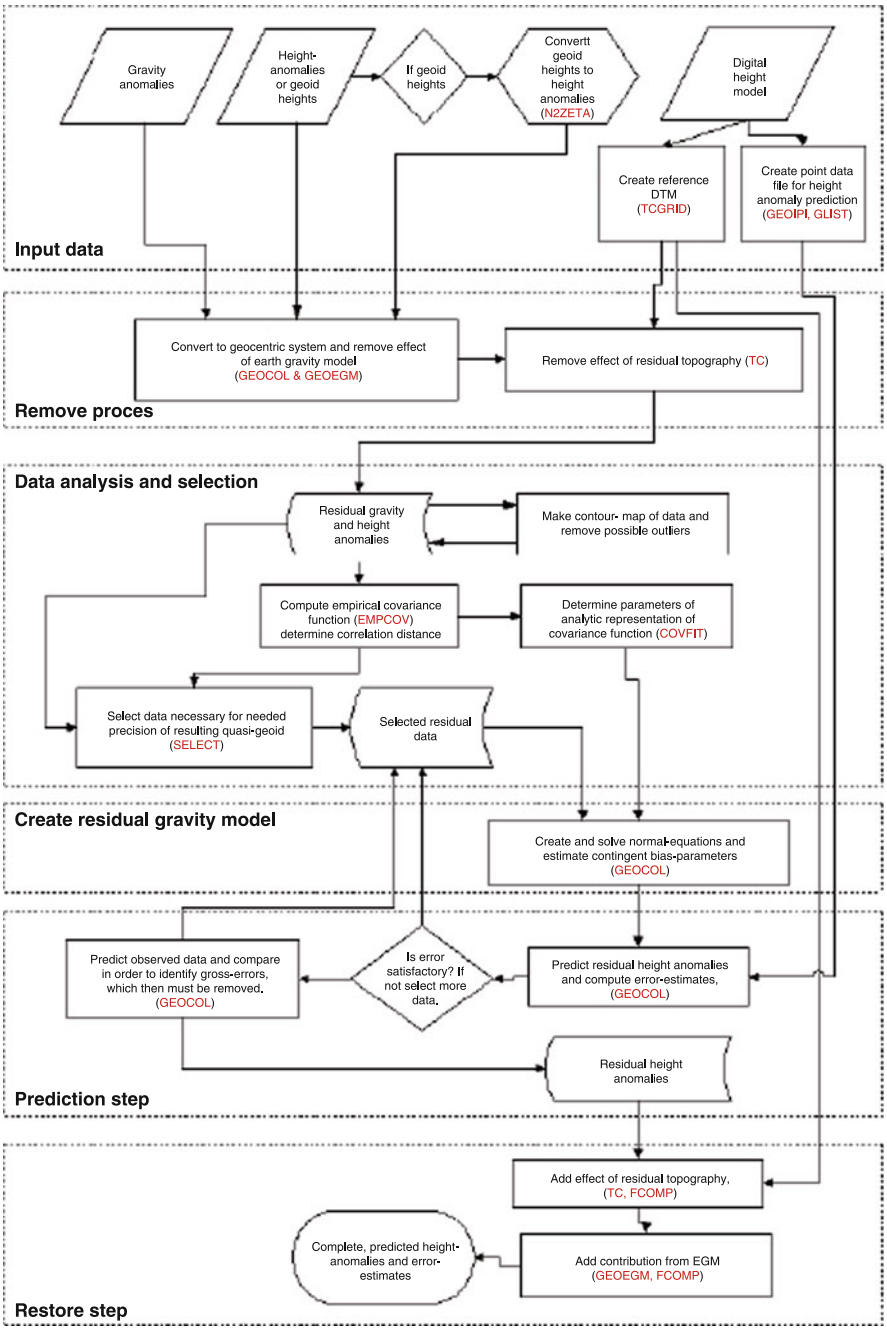
Appendix 3 Python Interface to GEOCOL for Height-bias Estimation



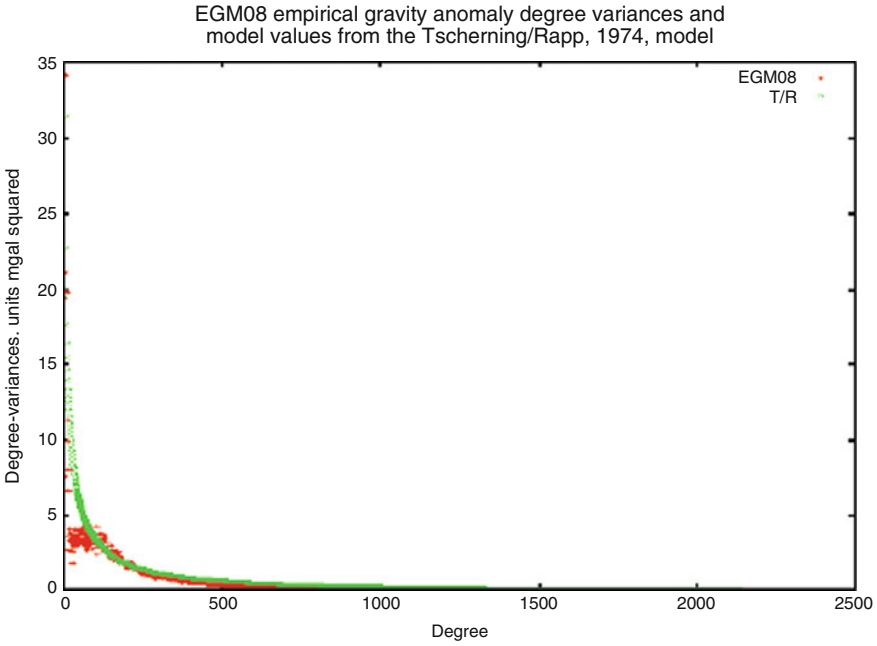
The result of the bias estimation is −0.78 m with a standard deviation of 0.09 m. No suspected gross-errors were detected. A similar detection was carried out for the gravity data, but no errors found.



Appendix 4 3D LSC GRAVSOFT Flowchart



Appendix 5 EGM98 and T/R Gravity Anomaly Degree Variances



Covariances between quantities KP and KQ, evaluated in P,Q, having spherical distance ψ and an azimuth equal to zero. Values of KP or KQ used signify 1 = height anomaly, 3 = gravity anomaly, 6 = component of deflection of the vertical in the direction between P and Q. Units: m, mgal and arcsec

Chapter 8
Topographic Reductions in Gravity and Geoid Modeling

Ilias N. Tziavos and Michael G. Sideris

8.1 Outline of the Chapter

This chapter focuses on a review of the conventional methods widely used for the computation of the effects of topography and bathymetry in geoid and quasi-geoid modeling. Terrain and bathymetry models of high-resolution and accuracy are used in order to provide the high-frequency content of the gravity field spectrum through the available mass reduction methods (e.g., terrain corrections, simple and refined Bouguer effects, residual terrain model, isostatic reduction schemes). Several other reduction schemes (e.g., the Rudzki and Poincaré and Prey reductions), which are briefly discussed herein, can be possible alternatives of computation of mass effects in gravity field modeling, although they are not commonly used in geodetic applications. The high-frequency contribution of the topographic and bathymetric effects to gravity-field related quantities (e.g., gravity anomalies, gravity disturbances, geoid undulations, deflections of the vertical, gravity gradients) is primarily due to the strong correlation of the short-wavelength gravity features with topography and bathymetry.

In the basic theory presented in this chapter, as well as in the practical computational examples discussed herein, emphasis is given to the terrain effects on geoid computations over continental areas, even though the same principles apply to the marine or oceanic environment, given the relative density changes and the typical representation of bathymetry by a digital depth model (*negative heights*) in correspondence with the representation of the visible topography by a digital terrain model. In the subsequent sections, the basics of gravity field modeling are briefly reviewed in connection with the solution of the geodetic Boundary Value Problem (BVP) based on Stokes's and Molodensky's theory. Then, the conventional mass reductions to gravity data are outlined along with the resulting gravity anomalies, which are employed in geoid/quasi-geoid calculations and other applications in geodesy and geosciences. Furthermore, the indirect effects caused by some of the aforementioned topographic reduction schemes are formulated and, finally, practical computational examples are presented on various mass reductions.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_8,

Although the majority of equations related to geoid and gravity field modeling have been given in other chapters and mainly in Part I (Chap. 4) of this book, some fundamental equations are repeated here in order to make as self-contained the chapter as possible. Nevertheless, only references are made to Part I (Chaps. 1 and 7) and Part II (Chap. 10) of the book, when the effects of different topographic reductions on geoid/quasi-geoid computations are discussed. An extensive and updated list of references is incorporated in the unified bibliography of the book, where the interested reader can find more theoretical details, numerical results and a variety of applications directly connected with the effects of topographic reductions in geoid computations and gravity field modeling at different scales.

8.2 Introduction

The reductions of gravity-field related quantities (e.g., gravity anomalies and disturbances, geoid heights, deflections of the vertical, gradients of the disturbing potential) for the effect of topographic and/or bathymetric masses play a crucial role in geodetic applications and particularly in geoid modeling.

In the solution of the geodetic BVP for the determination of precise geoid and quasi-geoid undulations using the Stokes or Molodensky approaches the masses are taken into account in a different way and play a particular role in the solution of the corresponding problem. Generally speaking, the gravitational attraction of topographic masses creates a strong gravity signal that dominates the gravity spectrum in shorter wavelengths and therefore the topograhy can be used to smooth the gravity field before any modeling process (Forsberg 1984, 2010), i.e., the gravity field may be smoothed by the terrain reductions. Additionally, the presence of the topography implies that the gravity observations (e.g., gravity anomalies) are given on a non-level surface and consequently the basic requirement of Stokes's theory is not valid and Helmert's or Molodensky's condensation methods should be applied to offset the non-level surface. It should be noted that in practical computations the mass reductions are considerably bigger than the non-level surface corrections in Molodensky's approach. The Molodensky-type corrections have no meaning in oceanic areas, where the geoid and quasi-geoid surfaces coincide and the available gravity observations refer to the geoid. However, the bathymetry has a strong effect on gravity data, that is comparable or even larger than the corresponding effect of the topography (Forsberg 1984, 1985, 2010). Although the effect of the bathymetry on gravity observables was neglected in the past, mainly due to the lack of detailed bathymetric data grids, in modern-day gravity field modeling the effects of bathymetry are seriously considered in numerical applications for the improvement of marine geoid models as high-resolution depth data are readily available.

The Digital Terrain and Bathymetry Models (DTMs and DBMs, respectively) play a crucial role in gravity field studies, since they provide the high-frequency content of the gravity field spectrum through the available mass reduction methods

(e.g., the simple and complete Bouguer effects, the classic terrain corrections, the residual terrain model (RTM) and the isostatic reduction schemes). This high-frequency contribution of the topographic effects to different gravity field constituents is due to the high correlation of the short-wavelength gravimetric features with the topography/bathymetry. According to Schwarz (1984), about 2% and 34% of the geoid height and gravity anomaly spectra, respectively, is contained in the high frequencies (harmonic degrees 360–36,000), where terrain effects play a significant role. Furthermore, gravity field and geoid/quasi-geoid approximation is based heavily on the well-known remove-restore procedure (Forsberg and Tscherning 1981; Forsberg 1993; Schwarz et al. 1990). In this scheme the topography/bathymetry data are used along with a Global Geopotential Model (GGM) to smooth the observations, to aid data gridding, transformations and predictions and eliminate aliasing (under-sampling) effects (Forsberg and Tscherning 1981; Forsberg 1985; Forsberg and Solheim 1988; Sjöberg 2005; Tziavos et al. 1992; Tziavos et al. 2010). In geophysics topographic reductions to gravity anomalies are used to gain insight of the mass distribution and lateral as well as radial density variations in the Earth's lithosphere and estimate the isostatic compensation of topographic features inside Earth's mantle in the form of Moho depths (see, e.g., Forsberg 1984; Huang et al. 2001; Kuhn 2003; Strykowski et al. 2005; Tziavos et al. 2010). Even though the needs for the computation of mass effects for these two branches of geosciences have different origins, they both require high DTM and DBM accuracy and resolution. Higher accuracy means that fewer and smaller errors are propagated in the final estimates (gravity, geoid and density variations) thus leading to better approximations of reality. Higher resolution means that aliasing effects are reduced and the spatial resolution of the estimated fields is increased so that a better picture of reality is gained.

A variety of methods can be used for the computation of all topographic effects on gravity field observables located either on a boundary surface or on its exterior. Numerical and spectral methods are currently widely used in mass effect modeling primarily in a grid-wise fashion. The numerical integration method (NIM) is a highly accurate but time consuming procedure, which can be used for point-wise or grid-wise computations. The rectangular prisms representation is a rigorous and useful model for numerical integration, but numerically unstable over large distances, where approximative formulas can be employed (Forsberg 1984). An advantage of the prism method in certain applications is the fact that it is originally designed for single point computation and thus works well and produces even better results, when detailed information around the computation point is available, besides the heights of the regular grid. An approximation model of the prism representation, i.e., the mass line model, trades computational efficiency for accuracy (Li 1993; Tziavos 1993; Li and Sideris 1994). Different alternatives of numerical integration (e.g., Gaussian quadrature) can sufficiently model the mass effects in local applications, but additional computational efforts are needed (Hwang et al. 2003). Regarding the spectral methods, the fast Fourier transform (FFT) technique is one of the most efficient tools for handling large amounts of height data, although special attention should be paid to the problems arising

from the numerical evaluation of this spectral approach (see, e.g., Forsberg 1984; Sideris 1984; Haagmans et al. 1993; Li 1993; Tziavos 1993; Tziavos et al. 1988; Li and Sideris 1994). Finally, different combination schemes are used for the treatment of height data towards the modeling of topographic effects. These combined methods are mainly based on the evaluation of the numerical integration in the intermediate zone around the computation point and the use of FFT in the rest of the area (e.g., Tziavos and Andritsanos 1998; Tsoulis 1999; Jekeli and Serpas 2003; Jekeli and Zhu 2006; Zhu and Jekeli 2009).

8.3 Topographic Reductions and Gravity Field Modeling

In this section the general integral formulas of the potential and the attraction of the Earth's topography are given along with a brief discussion on the data needed for their efficient evaluation. Then, different reduction schemes are discussed, that can be directly applied to gravity anomalies in order to produce residual gravity fields appropriate for interpolation, gridding or densification purposes as well as useful to gravity database generation and geophysical interpretation.

8.3.1 The Potential and the Attraction of the Earth's Topography

The topographic potential at a point $P(x_P, y_P, H_P)$ on the Earth's surface in a flat-Earth approximation can be expressed by Newton's integral as

$$T(x_P, y_p, H_P) = G \iint_E \int_0^H \frac{\rho(x, y, z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{1/2}} dx dy dz, \tag{8.1}$$

where G is the gravitational constant and $\rho(x, y, z)$ is the 3-dimensional (3D) density function, which can be moved out of the integral when it is assumed to be uniform within the masses.

The topographic vertical attraction at a point $P(x_P, y_P, H_P)$ on the Earth's surface is the negative first-order derivative of the potential of the topographic masses in the z-direction and may be expressed as

$$T_z(x_p, y_p, H_P) = G \iint_E \int_0^H \frac{\rho(x, y, z)(H_P - z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{3/2}} dx dy dz. \tag{8.2}$$

The topographic effect on gravity expressed by (8.2) can be split into two parts, i.e., the Bouguer plate effect $B(x_p, y_p, H_P)$ and the terrain correction $c(x_p, y_p, H_P)$, and, therefore, this equation can be rewritten as

$$T_z(x_p, y_p, H_P) = B(x_p, y_p, H_P) - c(x_p, y_p, H_P). \tag{8.3}$$

Details on the above mentioned formulas and analytical derivations are given in Chaps. 3 and 4 (Part I of the book), while additional information on their numerical implementation in the spatial and frequency domain are given in the following sections as well as in Chap. 10 (Sect. 10.4.2 in Part II of the book).

The high-resolution and accuracy modeling of the Earth's topography plays a fundamental role in the practical evaluation of the integral formulas given before, as well as in the determination of the gravity field constituents and especially in the computation of geoid and quasi-geoid heights. Attention has to be paid to the short-wavelength topographic and/or bathymetric effect in mountainous areas, where the different kind of mass reductions have a dominant contribution (see, e.g., Li and Sideris 1994).

In practical research applications the topography is usually represented by a set of rectangular prisms with the density of masses to be assumed as constant within each prism. Therefore, it has become apparent the need for a very high-resolution DTM to compute the terrain effects on gravity and the indirect effect on the geoid. This necessity is of main importance today that ultra-high resolution GGMs like EGM2008 (Pavlis et al. 2008) are available and high-resolution, e.g., $0.5'$–$1'$, geoid models are needed. If this information is not available and a coarser DTM is used, then the topographic effects computed are aliased due to the insufficient resolution (under sampling) of the topographic data used. Nevertheless, in several countries, even today, high-resolution local DTMs are not available due to either lack of data or confidentiality reasons. Furthermore, the DTMs available are usually not homogeneous, since they are derived by a (simple, in most cases) merging of available height data by the digitization of available historic topographic maps, which were originally produced by photogrammetric methods. Therefore, even though new, higher-resolution and higher-accuracy gravity field related data are available to the scientific community, the accuracy and resolution of the available DTMs and DBMs are not always adequate for the determination of precise geoid and gravity field models (e.g., Tziavos et al. 2010).

The problems described above improved significantly when high-resolution data of the Earth's topography with global homogeneous coverage were collected by recent dedicated satellite missions. In 2000, the Shuttle Radar Topographic Mission (SRTM) was launched on-board the space shuttle Endeavour and a wealth of data of the Earth's topography collected (see, e.g., Farr et al. 2007). This resulted in the release of a global $3''$ (roughly 90 m) SRTM DTM. Even higher-resolution DTMs ($1''$ globally, roughly 30 m) were made available from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), along with reflectance and temperature data of land surface (see, e.g., Yamaguchi et al. 1998).

Two SRTM-based DTMs, after a refinement by a national elevation model, were used in the numerical tests summarized below (see Sect. 8.6).

From the SRTM data resolution and the estimated horizontal and vertical accuracy (90% linear error), approximately at the level of 7 m in both directions (Farr et al. 2007), it became obvious that such a global DTM could offer great aid to local and regional gravity field and geoid determination. The latter refers to the use of the SRTM data to either fill-in gaps and densify local and regional/continental DTMs or as a stand-alone DTM for the computation of topographic reductions. In all cases the inherent problems of the SRTM elevations due to mountain shadowing and roof-top effects (SRTM is a digital surface model (DSM) more than a DTM) should be acknowledged. It should be mentioned here that a spherical harmonic expansion of the Earth's topography is recently available with a maximum degree of evaluation 2,160 (Pavlis et al. 2007a). This expansion can be used in combination with SRTM-based DTMs of finer resolution to model gravity field structures at scales shorter than those offered by EGM2008 by the aid of an appropriate mass reduction method (Hirt et al. 2010). More details are discussed in Sect. 8.4.4.

In sea areas or in coastal areas, the DTMs are represented by a combination of DTMs and DBMs available in the marine regions for the computation of mass reductions to gravity data. The latest bathymetry models widely used in gravimetric geoid determination are those developed by the Danish National Space Agency (Andersen and Knudsen 2008) and the Scripps Institute of Oceanography (Smith and Sandwell 1997). Both models with a resolution of, approximately, $1'$ globally (roughly 1.8 km), have been produced by the inversion of satellite altimetry measurements and differ only in terms of the methodology used for their development. In the former, the inverse Stokes method was applied to the altimetry data (see details in Part II, Chap. 10 of this book) and in the latter the bathymetric depths were derived from deflections of the vertical computed along the altimetric tracks.

Besides the DTM and DBM contributions, Digital Density Models (DDMs) are also of importance in gravimetric geoid computations (see, e.g., Tziavos and Featherstone 2001). Gravimetric geoid models typically use a constant topographic density in their computation. This was mainly due to the lack of detailed density models. As it has been reported in several studies, the actual density of the topographic masses may differ by more than 10% from a constant density assumption, mainly in areas with complicated geological structures (see, e.g., Martinec et al. 1994; Tziavos et al. 1996; Kuhn 2000; Makhloof 2007). This will introduce errors in mass reductions that will be propagated to geoid heights. Therefore, a 3D DDM would be ideally needed for the modeling of topographic and deeper masses (Li 1993; Pagiatakis et al. 1999; Huang et al. 2001), although two-dimensional (2D) density models are usually sufficient for geoid computations. The latter may be produced from density information extracted from geological maps. It is recommended that, if available, a reasonable DDM model be used in all steps of geoid modeling in order to further improve the accuracy of the computed geoid heights.

8.3.2 *Terrain Reductions for Gravity Densification and Gridding*

Many methods are available for the mathematical and physical treatment of the contribution of topography to gravity field related quantities, which are usually formed as reductions to the available input data. The difference between the various terrain reduction methods is based on the way each one treats the topographic masses outside the geoid and from a theoretical point of view they should all provide the same result. In numerical applications two important considerations are usually taken into account in the selection of the most appropriate topographic reduction method: (a) The magnitude of the indirect effect that should be restored to the reduced geoid heights and (b) the smoothness, the magnitude and the mean value of the reduced gravity anomalies that will be used for geoid height prediction using the Stokes's integral formula or other space and frequency based methods (e.g., Tziavos et al. 2010). The former is vital since larger indirect effects can result in larger prediction errors during their computation, thus larger errors will be propagated to the geoid height estimates. The latter refers to the smoothness of the residual gravity anomalies after the reduction for the topography for the sake of easiness and improved precision of prediction, gridding and interpolation operations. The necessity for a zero mean to the reduced data lies mainly to the requirement that the reduced field is treated as a stationary random process in least-squares collocation (LSC) based estimation problems (see details in Chap. 7, Part II of this book). If the reduced gravity anomaly field has zero mean, then the signal error can be regarded as free of biases and the interpretation as a random field is facilitated. Note that another necessary operation during the remove step of the remove-restore method, is the removal of a low-degree harmonic field, i.e., to reference the input data to some global geopotential model (e.g., EGM2008), which further reduces regional trends and contributes to the further smoothness of the reduced field. Extensive discussions on this subject can be found in, e.g., Moritz (1980), Forsberg (1993), Martinec and Vaníček (1994), Sideris (1994, 2010), Forsberg and Tscherning (1997), Tscherning (2010).

8.3.2.1 Bouguer Reduction

The complete or refined Bouguer reduction removes all the topographic masses above the geoid contained in the Bouguer plate as well as the irregular part of the topography deviating from the Bouguer plate, i.e., the so-called terrain correction, as it was mentioned in the previous section (see 8.1–8.3 and Fig. 8.1).

The Bouguer reduction and the terrain correction in (8.3) can be expressed, in an analogous way as the topographic vertical attraction (8.2), by the following integral equations, respectively:

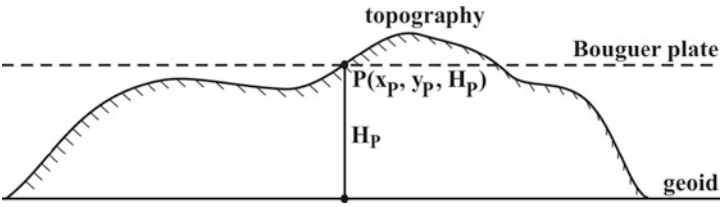


Fig. 8.1 The topography and the Bouguer plate

$$B(x_p, y_p, H_P) = G \iint_E \int_0^{H_P} \frac{\rho(x, y, z)(H_P - z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{3/2}} dxdydz, \quad (8.4)$$

$$c(x_p, y_p, H_P) = G \iint_E \int_H^{H_P} \frac{\rho(x, y, z)(H_P - z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{3/2}} dxdydz. \quad (8.5)$$

If the radius of the previously mentioned area is infinite, the Bouguer reduction in the case of a simple horizontal plate can be determined as:

$$B = 2\pi G\rho H_P. \quad (8.6)$$

Using the approximated value of G that has been already given in Chap. 1 (Sect. 1.2, in Part II of this book) and assuming a constant density $\rho = 2670\,\mathrm{kgm}^{-3}$ ($2.67\,\mathrm{gcm}^{-3}$), the simple Bouguer reduction reads

$$B = 0.1119 H_P, \quad (8.7)$$

that gives the reduction in mGal when H is given in meters.

The terrain correction formula (8.5) is a refinement of the simple case of the Bouguer plate, since it accounts for the surpluses and deficits of the actual Earth's topography from the aforementioned horizontal Bouguer plate (Fig. 8.2). Based on the kernel function of the terrain correction formula expressed by $(\Delta h / l^3)$, an area E of $100 \times 100\,\mathrm{km}$ may be considered big enough to get a reasonable accuracy in the computation of the terrain correction at a point P lying at the center of this area (Peng 1994).

The 2D linearly approximated formula for the terrain correction at a point P on a plane reference surface E is derived from (8.5) integrating with respect to z and in this case the triple integral of the terrain correction reads as follows (Sideris 1985):

$$\begin{aligned} c(x_p, y_p) &= G \iint_E \frac{-\rho(x, y)}{\left(l_0^2 + z^2\right)^{1/2}} \Big|_0^{\Delta H} dxdy \\ &= \iint_E \frac{\rho(x, y)}{l_0} \left[1 - \left[1 + \left(\frac{\Delta H}{l_0}\right)^2\right]^{-1/2}\right] dxdy, \end{aligned} \quad (8.8)$$

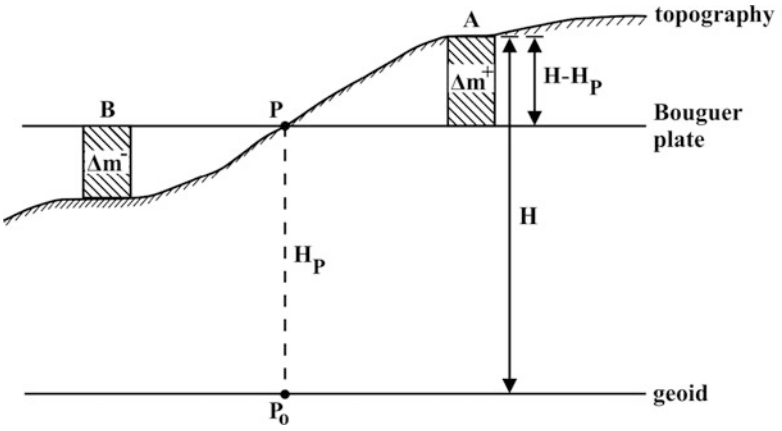


Fig. 8.2 The geometry of the planar Bouguer reduction and the terrain correction

where $\Delta H = H(x_P, y_P) - H(x, y)$ and $l_0^2 = (x_p - x)^2 + (y_p - y)^2$. For $(\frac{\Delta H}{l_0})^2 \leq 1$, the term $[1 + (\frac{\Delta H}{l_0})^2]^{-1/2}$ can be expanded into a series as follows:

$$\left[1 + \left(\frac{\Delta H}{l_0}\right)^2\right]^{-1/2} = 1 - \frac{1}{2}\left(\frac{\Delta H}{l_0}\right)^2 + \frac{1.3}{2.4}\left(\frac{\Delta H}{l_0}\right)^4 - \frac{1.3.5}{2.4.6}\left(\frac{\Delta H}{l_0}\right)^6 + \ldots. \tag{8.9}$$

Keeping the terms up to third order we get the following approximation of the terrain correction integral:

$$\begin{aligned} c(x_p, y_p) = {} & \frac{1}{2}G \iint_E \frac{\rho(x, y)(H_P - H)^2}{\left[(x_P - x)^2 + (y_P - y)^2\right]^{3/2}} dxdy \\ & - \frac{3}{8}G \iint_E \frac{\rho(x, y)(H_P - H)^4}{\left[(x_P - x)^2 + (y_P - y)^2\right]^{5/2}} dxdy \\ & + \frac{5}{16}G \iint_E \frac{\rho(x, y)(H_P - H)^6}{\left[(x_P - x)^2 + (y_P - y)^2\right]^{7/2}} dxdy \end{aligned} \tag{8.10}$$

If $\frac{\Delta H}{l_0} \ll 1$ a good approximation is obtained by $[1 + (\frac{\Delta H}{l_0})^2]^{-1/2} \approx 1 - \frac{1}{2}(\frac{\Delta H}{l_0})^2$ and by substituting it into (8.8) we finally get:

$$c(x_P, y_P) = \frac{1}{2}G \iint_E \frac{\rho(x, y)\left[H(x_P, y_P) - H(x, y)\right]^2}{\left[(x_P - x)^2 + (y_P - y)^2\right]^{3/2}} dxdy. \tag{8.11}$$

This last equation represents the so-called *linear approximation* of the terrain correction. Similar approximation formulas for terrain correction as before can be

derived by using Molodensky's operator as it is discussed in Chap. 10 (Sect. 10.2.2) in Part II of the book. It is worth noticing that in the practical evaluation of the above mentioned approximation formulas (8.10 and 8.11), the series converges only for terrain inclination smaller than 45° and that, even for smaller inclinations, numerical instabilities might occur in the computation of higher order terms and especially when computations are based on high resolution DTMs (e.g., Tziavos et al. 1988, 1992; Sideris 1990; Peng 1994). Nevertheless, the computation of higher order terms, mainly in numerical tests in roughed terrains, is still significant in order to obtain more accurate results due to the better modeling of the high frequency part of the topography. To overcome the problem of numerical instabilities mentioned before, either the rigorous rectangular integration or combined computational schemes based on numerical integration for an inner zone and discrete FFT for the rest of the area can be used (see, e.g., Sun 2002; Jekeli and Serpas 2003; Jekeli and Zhu 2006; see also Part I, Chap. 5 of the book).

For the complete derivation and the assumptions made to derive (8.11), as well as a detailed discussion for its numerical evaluation, see, e.g., Forsberg (1984), Sideris (1984, 1985), Tziavos (1993), Tsoulis (1999). It should be noted that the double integral of this equation is a convolution integral and can be efficiently evaluated by FFT, as it is extensively discussed later on in Sect. 8.5.2 and in Chap. 10 (Part II).

When density values are available on a regular grid (DDM) of the same resolution with that of heights, the integral of (8.11) can easily be modified to account for different densities. Various studies have been conducted where lateral mass density variations have been considered in the computation of terrain reductions (see, e.g., Tziavos et al. 1996; Pagiatakis et al. 1999; Huang et al. 2001; Tziavos and Featherstone 2001; Kuhn 2000, 2003). In this case and when the FFT technique is implemented, extra computational effort is needed for the calculation of density spectra (see, e.g., Li 1993; Tziavos et al. 1996). The same problem can be also treated by the rigorous 3D FFT, but this is a time-consuming technique (Peng 1994; Peng et al. 1995).

Let consider now the effect of the bathymetric masses at a point lying on the geoid surface, i.e., on the ocean surface, that is the case of marine gravimetry. The methodology is similar to that of the terrain correction before and only the integration interval differs in (8.5), i.e.,

$$c_b(x_p, y_p, H_P) = G \iint_E \int_{-H}^{0} \frac{\Delta\rho(x, y, z)(H_P - z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{3/2}} \, dx dy dz, \tag{8.12}$$

where H are used as a function of depths, the term c_b denotes the effect of bathymetry and it is used to distinguish it from the terrain effect c. This effect is named sometimes in the geodesy and geophysical literature as *density contrast effect* or *bathymetry correction* (e.g., Tsoulis 1999), since it expresses the vertical

component attraction of the mass deficiencies over an oceanic area. This effect should be added to the gravity data on the geoid, which makes this effect always positive, like the corresponding terrain effect of the continental masses. The density contrast effect is evidently much smaller than the corresponding terrain effect; in (8.12), $\Delta\rho$ is the density contrast between the upper crust and bathymetry masses ($\approx 1.67\,\mathrm{gcm}^{-3}$). The above mentioned consideration of the bathymetry masses is not combined with the removal of a Bouguer plate, since the point is already located on the geoid and the bathymetry itself represents the relief of the bottom of the sea. By integrating with respect to z in (8.12), expanding in a binomial series and substituting $H_P = 0$, the following series of convolution integrals results (see, e.g., Parker 1995, 1996; Tsoulis 1999):

$$c_b(x_P, y_P) = \frac{1}{2}G \iint\limits_E \frac{\Delta\rho(x,y)H^2}{l_0^3}\,dxdy - \frac{3}{8}G \iint\limits_E \frac{\Delta\rho(x,y)H^4}{l_0^5}\,dxdy + \ldots. \tag{8.13}$$

The FFT representation of the last equation is given in Sect. 8.5.2.

8.3.2.2 Bouguer and Free-Air Gravity Anomalies

The attraction of the Bouguer plate expressed by (8.6) is the direct topographic effect of the Bouguer reduction on gravity. The gravity anomalies according to the Bouguer reduction scheme, i.e., the *incomplete* or *simple Bouguer gravity anomalies*, can be expressed as:

$$\Delta g_B = g - \gamma_o + F - B, \tag{8.14}$$

where g is the measured gravity at point P on the Earth's surface, γ_0 is the normal gravity computed on the reference ellipsoid and F is the free-air reduction. Taking into account the thin plate Bouguer reduction along with the terrain correction, i.e., the attraction for the complete Bouguer reduction (8.3) the *complete* or *refined Bouguer anomalies* are derived and expressed by the following formula:

$$\Delta g_B = g - \gamma_o + F - B + c. \tag{8.15}$$

The free-air reduction F constitutes a part of the topographic reduction procedure and it is used to transfer a gravity measurement from a point P on the Earth's surface to point P_0 on the geoid (see Fig. 8.2). The gravity change expressed by this reduction is given by the actual gravity gradient

$$F = -\frac{\partial g}{\partial H}H, \tag{8.16}$$

which is replaced in practice by the normal gradient of gravity

$$F = -\frac{\partial \gamma}{\partial H_P} H_P = 0.3086 H_P, \tag{8.17}$$

a sufficient approximation for flat or moderate terrains. Then the free-air gravity anomalies are given by the formula:

$$\Delta g_{FA} = g - \gamma_o + F. \tag{8.18}$$

It is evident that the Bouguer and free-air gravity anomalies are related through the formula:

$$\Delta g_B = \Delta g_{FA} - B + c. \tag{8.19}$$

The free-air gravity anomalies are the usual available input data in gravimetric geoid determination and a number of applications in other branches of geosciences, so that the different topographic reduction schemes are applied to these quantities. The free-air anomalies are referred to the geoid boundary surface in Stokes's BVP and to the topographic surface in Molodensky's BVP and more details on the solution of these problems through Δg_{FA} are given in Sect. 8.3.5.

Regarding the use of Bouguer anomalies in geosciences, it should be noticed that Bouguer gravity anomalies are frequently used in geophysics to infer geological information from gravity data and in geodetic applications to obtain boundary values on the geoid after the complete removal of all masses above the geoid. The planar approximation of the Bouguer gravity anomalies previously discussed, was used in the past in conjunction with a number of additional corrections (e.g., *Bullard B correction*) in order to account for a more realistic spherical Earth shape (see, e.g., Nowell 1999). In recent studies oriented to geodetic applications, Bouguer gravity anomalies in spherical form, either simple or complete ones, were computed over large regions in order to eliminate distortions from the use of an infinitely planar Bouguer plate (e.g., Kuhn et al. 2009). Special attention should be paid in spherical Bouguer computations with respect to the distance to which the computations have been carried out (Forsberg 1984). Moreover, a disadvantage of the spherical approach, from the computational point of view, is that terrain corrections have to be computed for the global topography, whereas they need to be computed over a restricted area in the planar case (Kuhn et al. 2009). Further discussions on this subject as well as numerical comparison results can be found in Featherstone and Dentith (1997), Kirby and Featherstone (1999), Featherstone and Kirby (2000), Novák et al. (2001), Vaníček et al. (2001, 2004), Kuhn et al. (2009).

8.3.2.3 Isostatic Reduction

In the general concept of isostasy, the topographic mass excesses (mountains) and deficiencies (waters) are compensated, to a large part, by a corresponding mass distribution in the interior of the Earth (e.g., Torge 2001). Two main theories

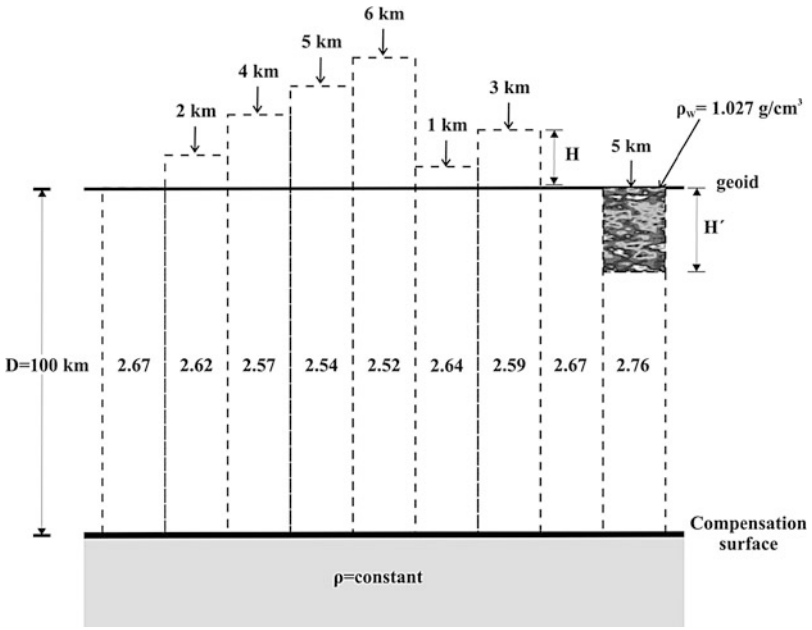


Fig. 8.3 The Pratt-Hayford model of isostatic compensation

have been developed in order to explain isostatic compensation, one following the Airy-Heiskanen (AH) model and another following the Pratt-Hayford (PH) model (Heiskanen and Moritz 1967). These two models are widely used in applications in geosciences, but the AH model has become a standard in geodetic research.

Pratt-Hayford Isostatic Model

According to the PH isostatic reduction scheme the topographic masses are distributed between the compensation surface and sea level. Furthermore, it is assumed that the density beneath the compensation level is constant, while the masses above that level for each column of cross-section are equal (see Fig. 8.3). Within that reduction scheme, the topographic masses are removed along with their isostatic compensation so that what remains is a homogeneous crust layer with constant density and constant depth of compensation.

The PH isostatic reduction considers that the level of compensation has a constant and uniform depth D assumed equal to 100 km measured from sea level. The topographic masses are delineated into columns of cross-section with height D that allows lateral changes in density in order to obtain isostatic equilibrium. Considering that a normal column ($H = 0$) has constant density ρ_o, the continental columns generate densities smaller than ρ_o while the oceanic columns are denser. The equilibrium conditions for the continental and oceanic areas are expressed as:

$$(D + H)\rho_{cont} = D\rho_o, \tag{8.20}$$

$$\left(D - H'\right)\rho_{oc} + H'\rho_w = D\rho_o, \tag{8.21}$$

with $\rho_o = 2.67\,\mathrm{gcm}^{-3}$, $\rho_w = 1.027\,\mathrm{gcm}^{-3}$. Then, the densities of the continental and oceanic columns are given as:

$$\rho_{cont} = 2.67\frac{D}{D + H}, \quad \rho_{oc} = \frac{2.67D - 1.027H'}{D - H'}. \tag{8.22}$$

For the condition represented by (8.20) to be satisfied (continental case) the actual density of the column $(D + H)$ is smaller than the normal constant value ρ_0 that implies that there is a density constant or mass deficiency. On the other hand for (8.21) the actual density of the column $(D - H')$ exceeds the normal constant value ρ_0 so that there is a density constant or mass surplus. The above mentioned density contrasts are expressed as

$$\Delta\rho_{cont} = \rho_0 - \rho_{cont} = \rho_0\frac{D}{D + H}, \quad \Delta\rho_{oc} = \rho_{oc} - \rho_0 = (\rho_0 - \rho_w)\frac{H'}{D - H'}. \tag{8.23}$$

The topographic effect due to this PH topographic isostatic scheme at a point P at the surface of the Earth and the corresponding PH reduction is the difference in the attraction between the topographic masses as described by the available DTM and the compensated masses within the depth of the root:

$$\Delta A_{PH} = A_{top/PH} - A_{comp/PH}. \tag{8.24}$$

The first term in (8.24) represents the attraction of the topographic masses and can be expressed in accordance to (8.2) as

$$A_{top/PH} = G\iiint_{E\ 0}^{H} \frac{\rho(x, y, z)(H_P - z)}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{3/2}}dxdydz, \tag{8.25}$$

and the second term that represents the attraction of the compensated masses is given as

$$A_{comp/PH} = G\iint_{E}\int_{-D-H_P}^{-H_P} \frac{\Delta\rho(x, y, z)(H_P - z)}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{3/2}}dxdydz, \tag{8.26}$$

where ρ and $\Delta\rho$ are given by (8.22) and (8.23) depending on the area of interest, i.e., continental or oceanic, respectively. Given the PH isostatic reduction the isostatic gravity anomalies can be computed as:

$$\Delta g_{PH} = g - \gamma_o + F - \Delta A_{PH}. \tag{8.27}$$

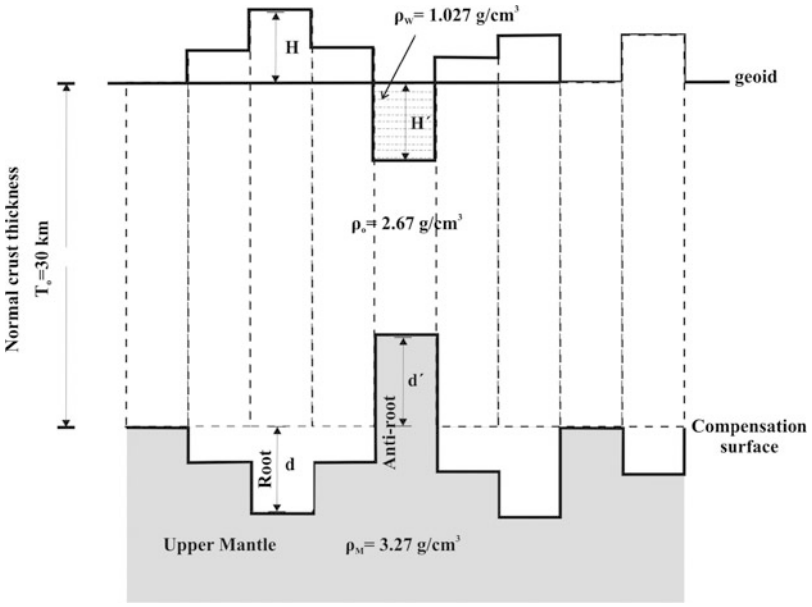


Fig. 8.4 The Airy-Heiskanen model of isostatic compensation

Airy-Heiskanen Isostatic Model

The AH model is based on the principle that the mountains are floating on some kind of higher density fluid meaning that there is mass deficit (roots) below mountains and mass surpluses (anti-roots) below the oceans. The AH model (see Fig. 8.4) is based on the assumptions that the isostatic compensation is complete and local, the density of the mountains is constant and equal to ($\rho_o = 2.67\,\mathrm{gcm}^{-3}$), the density of Earth's mantle is equal to ($\rho_M = 3.27\,\mathrm{gcm}^{-3}$) and the normal crust thickness T_0 is equal to 30 km (Heiskanen and Moritz 1967). Assuming a constant density of ($\rho_w = 1.027\,\mathrm{gcm}^{-3}$) for the ocean water, then the condition of floating equilibrium can be written as

$$(\rho_M - \rho_o)d = \rho_o H \tag{8.28}$$

for the continental cases, and

$$(\rho_M - \rho_o)d' = (\rho_o - \rho_w)H' \tag{8.29}$$

for the oceanic areas. In (8.28) and (8.29) d is the thickness of the root, d' is the thickness of the anti-root, H is the height of the topography and H' is the height of the ocean, i.e., the depth. Given the above mentioned density values for the crust, the mantle and ocean water, (8.28) and (8.29) can be written as:

$$d = 4.45H, \quad d' = 2.73H', \tag{8.30}$$

The topographic effect due to this AH topographic isostatic scheme at a point P at the surface of the Earth and the corresponding AH reduction is the difference in the attraction between the topographic masses as described by the available DTM and the compensated masses within the depth of the root:

$$\Delta A_{AH} = A_{top/AH} - A_{comp/AH}. \tag{8.31}$$

The first term in (8.31) represents the attraction of the topographic masses and can be expressed in accordance to (8.2) as

$$A_{top/AH} = G \iint_{E} \int_{0}^{H} \frac{\rho(x,y,z)(H_P - z)}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{3/2}} dxdydz, \tag{8.32}$$

while the second term that represents the attraction of the compensated masses is given as

$$A_{comp/AH} = G \iint_{E} \int_{-T_o - d - H_P}^{-T_o - H_P} \frac{\Delta\rho(x,y,z)(H_P - z)}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{3/2}} dxdydz, \tag{8.33}$$

where the values of ρ and $\Delta\rho$ depend on the area of interest (continental or oceanic). Given the AH isostatic reduction, the isostatic gravity anomalies can be computed as:

$$\Delta g_{AH} = g - \gamma_o + F - \Delta A_{AH}. \tag{8.34}$$

8.3.2.4 Interpolation and Gridding Through the Topographic Reductions

As it was mentioned in the introduction of this section, the topographic reduction methods should produce a smooth residual gravity field, most suitable for interpolation and gridding processes through LSC or other conventional techniques (splines, weighted means, etc.). This approach may sufficiently result in the creation of a high-resolution gravity database in a grid format or the densification of a test area with scarce gravity coverage, after the restoration of the effect of the topography in a second step through the employed reduction scheme.

Regarding the topographic reduction schemes discussed in this section, the complete Bouguer reduction removes all topographic masses above the geoid thus producing smooth residual gravity anomalies. The topographic isostatic reductions of PH and AH models remove the effects of the masses according to the isostatic compensation principle of each model and also produce smooth gravity residuals. Both Bouguer and isostatic reductions have physical meaning and present the proper characteristics for geophysical applications. Their disadvantage lies in the large indirect effect on the geoid that prohibits their use in geoid determination

(see Sect. 8.3.5). Nevertheless, these reduction schemes have a significant impact to gravity interpolation and thus can contribute to the creation of a, e.g., gridded free-air gravity anomaly field that can be used in geoid determination. Such a procedure can be realized according to the following scheme:

- Pre-processing and cleaning of the original point free-air gravity anomalies for gross errors and outliers.
- Removal of the topographic effect through the refined Bouguer reduction or the compensating masses (PH or AH isostatic model).
- Interpolation (prediction) on a selected grid using, e.g., collocation.
- Restoration of the topography through the Bouguer or isostatic reduction scheme.

The above mentioned procedure can be combined with another operation during the remove step, i.e., the removal of the contribution of a GGM, which further reduces regional trends and makes the reduced field less irregular. More details on this combination procedure are given in Sect. 8.4.5 and in Chap. 7 (Part II of the book).

8.3.3 *Topographic/Isostatic Effects on Gravity and Airborne Gravity and Gradiometry*

The effects of topographic and compensated masses according to the PH and AH models presented before, are given by (8.25–8.26) and (8.3–8.33), respectively. These equations can be rigorously evaluated by the 3D FFT method as it is shown in Sect. 10.4.2 (Chap. 10, Part II), but this is a time-consuming computational procedure. For this reason the above mentioned integrals can be simplified in a 2D form, as it is explained below, in order to be efficiently evaluated by 2D FFT (see Sect. 8.5.2). Further derivations can be found in several research papers (see, e.g., Forsberg 1984; Li 1993; Peng et al. 1995).

The effect of the compensated masses at sea level (see Fig. 8.5) is given by (8.33) for the AH model, which is written in the following simplified form by substituting ($H_P = 0$)

$$A_{comp/AH} = G \iint_E \int_{-T_o-d}^{-T_o} \Delta\rho \frac{z}{l^3} dx dy dz. \tag{8.35}$$

The kernel function (z/l^3) may be evaluated in a power series around a suitable reference level d_0 of the compensated masses represented by the function $d(x, y)$ (see Fig. 8.5), as it is shown in Forsberg (1984)

$$\frac{z}{l^3} = \frac{d_0}{l_0^3} + \frac{l_0 - 3d_0}{l_0^3}(z - d_0) + \ldots, \tag{8.36}$$

where l_0 in this case is $l_0 = \sqrt{(x_P - x)^2 + (y_P - x)^2 + d_0^2}$. For more details Forsberg (1984, 1985) should be consulted. Substituting (8.36) to (8.35) and

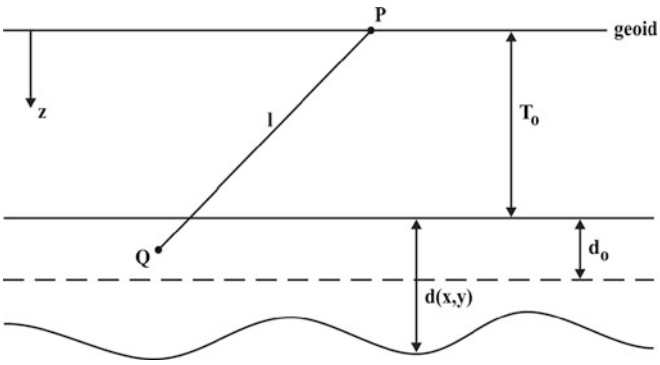


Fig. 8.5 The geometry of the isostatic effect (After Forsberg 1985, 2010)

integrating with respect to z, the following formula derives for the effect of compensated masses on gravity data located to the geoid (Forsberg 1984, 2010; Kuhn 2000):

$$A_{comp/AH} = G\left[\left[\Delta\rho\, d\,\left(\frac{d_0}{l_0^3} + \frac{l_0^2 - 3d_0^2}{l_0^5}(T_0 - d_0)\right)\right] + \frac{1}{2}\left(\Delta\rho\, d^2\,\frac{l_0^2 - 3d_0^2}{l_0^5}\right)\right], \tag{8.37}$$

which can be easily written in convolution form and then evaluated by 2D FFT as it is presented in Sect. 8.5.2.

The topographic and/or isostatic effects are also of main importance to airborne gravity and gradiometry data, since they contribute to account for the effect of the *topographic noise* of these data. In geophysical exploration this facilitates the interpretation of the subsurface density anomalies (e.g., Tziavos et al. 1988).

The gravitational potential at a point $P_0(x_P, y_P, z_0)$ due to the topography $H(x, y)$ in an area E is given by an equation analogous to (8.1) as

$$T(x_P, y_p, z_0) = G \iint_E \int_0^H \frac{\rho(x, y, z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (z_0 - z)^2\right]^{1/2}}\, dx dy dz, \tag{8.38}$$

where z_0 is the constant flight height (see Fig. 8.6). The vertical component of $T(x_P, y_p, z_0)$ (8.38) gives the topographic effect for airborne gravity measurements at the point $P_0(x_P, y_P, z_0)$ by

$$T_z(x_p, y_p, z_0) = G \iint_E \int_0^H \frac{\rho(x, y, z)(z_0 - z)}{\left[(x_P - x)^2 + (y_P - y)^2 + (z_0 - z)^2\right]^{3/2}}\, dx dy dz, \tag{8.39}$$

which can be derived directly from (8.2) for ($H_P = z_0$).

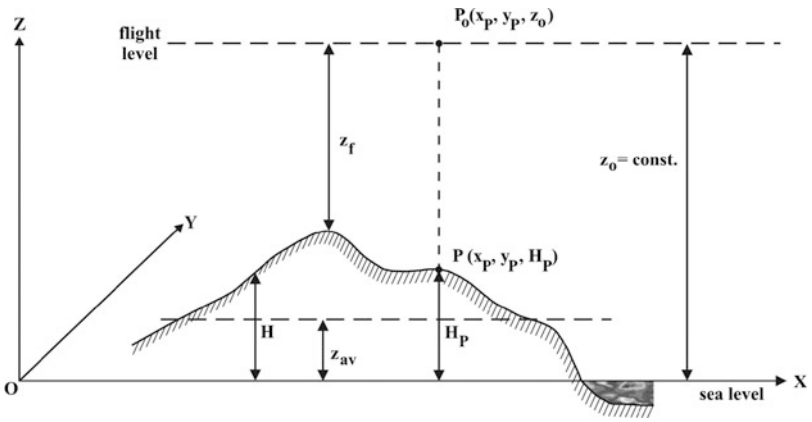


Fig. 8.6 The geometry of airborne gravity and gradiometry (After Tziavos et al. 1988)

In practice a 2D approximated formula is frequently used that results from the expansion of (8.39) around $(z = 0)$ and working in a similar way with that of Sect. 8.3.2 (see 8.5–8.11). It finally yields (see, e.g., Tziavos et al. 1988; Peng 1994)

$$T_z(x_P, y_P, z_0) = G \iint_E \rho \frac{z_0}{l_0^3} H(x, y) dx dy - \frac{G}{2} \iint_E \rho \left(\frac{1}{l_0^3} - \frac{3z_0^2}{l_0^5} \right) H^2(x, y) dx dy, \tag{8.40}$$

where in this case $l_0 = \sqrt{(x_P - x)^2 + (y_P - x)^2 + z_0^2}$.

Note that in the truncated development (8.40) the kernels l_0^{-3}, l_0^{-5} never become singular, which is a distinctive characteristic of airborne gravimetry.

The topographic effect on airborne gradiometry is realized by the second-order derivative of (8.2) for $(H_P = z_P = z_0)$. In order to obtain the 2D approximated formula as before, this derivative is expanded into a series around $(z = 0)$ and only first order terms are finally kept. Integrating with respect to z the six components of the topographic effects on airborne gradiometry can be derived. The interested reader should consult, e.g., Tziavos et al. (1988), Peng (1994), Peng et al. (1995) for analytical derivations. As an example the T_{zz} component at flight level z_0 is given here, expressed as:

$$T_{zz}(x_P, y_P, z_0) = G \iint_E \rho \frac{(l_0^2 - 3z_0^2)}{l_0^5} H(x, y) dx dy + \frac{G}{2} \iint_E \rho \left(\frac{9z_0}{l_0^5} - \frac{15z_0^2}{l_0^7} \right) H^2(x, y) dx dy. \tag{8.41}$$

The 2D convolution integrals expressed by (8.40) and (8.41) are efficiently evaluated by the FFT method, which is an obvious step for the computation of topographic

effects on airborne gravity and gradiometry, because large numbers of data are generated in grid like pattern. It has to be noticed that the expansions of (8.39) into a series around ($z = 0$) is evaluated around ($z = z_{av}$), that is around the mean average of the heights of the topography in order to make series converge faster (Tziavos et al. 1988).

The isostatic effects on airborne gravity and gradiometry can be evaluated in a similar way as before. Considering AH model, the isostatic effect of the compensated masses is represented by (8.33) by simply applying ($z_P = z_0$). First, an expansion of (8.36) is carried out into a series around ($z = -T_0$), then the first order terms of the series expansion are kept and finally the 2D linear formulas for the isostatic effects are derived for airborne gravity and gradiometry, respectively, similar to (8.40) and (8.41).

8.3.4 Terrain Reductions and Physical Heights

The Poincaré and Prey reduction, usually abbreviated as Prey reduction, refers to the need of determining gravity inside the earth, where gravity cannot be measured, but it can be computed by surface gravity (see Fig. 8.7). The purpose of this reduction is different of that of the other gravity reductions mentioned before, which give gravity values, or better gravity anomalies, at a boundary surface. Consequently, the Prey reduction cannot be used directly for geoid determination, but can be employed for obtaining orthometric heights (Heiskanen and Moritz 1967; Torge 1989, 2001).

A direct way of computing g_Q is by using the formula

$$g_Q = g_P - \int_Q^P \frac{\partial g}{\partial H}\, dH, \tag{8.42}$$

where P and Q are situated at the plumb line and under the assumption that the gravity gradient ($\partial g/\partial H$) inside the earth is known (see also discussion in Sect. 8.3.5). Using Poisson's equation and the normal free-air gradient, (8.42) reduces to

$$g_Q = g_P + 0.0848(H_P - H_Q), \tag{8.43}$$

where g in mGal and H in km (see for details Heiskanen and Moritz 1967; Hofmann-Wellenhof and Moritz 2005). Another practical way of computing g_Q is a kind of a remove-restore procedure, provided that between P and Q only the

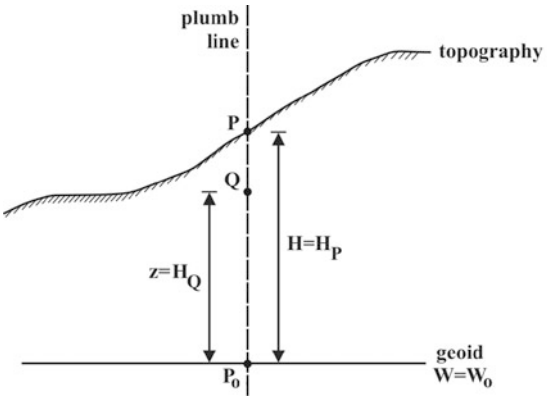


Fig. 8.7 The geometry of Poincaré and Prey reduction (After Heiskanen and Moritz 1967)

infinite Bouguer plate is used, neglecting the terrain correction. This procedure is carried out in three steps as follows (Heiskanen and Moritz 1967):

Gravity at P	g_P
• Remove Bouguer plate	$-0.1119(H_P - H_Q)$
• Free-air reduction (P to Q)	$+0.3086(H_P - H_Q)$
• Restore Bouguer plate	$-0.1119(H_P - H_Q)$
Gravity at Q	$g_Q = g_P + 0.0848(H_P - H_Q)$

Although the meaning of this reduction is different from that of the other gravity reductions, the above-mentioned three-step procedure is combined with some kind of downward continuation methodology in several applications.

8.3.5 *The Treatment of the Topography in Geoid and Quasi-geoid Determination*

The different mass reduction schemes are connected with the solution of Stokes's and Molodensky's BVPs which are extensively discussed in Chaps. 14 and 15 in Part III of the book.

The solution of the Stokes problem is based on gravity anomalies reduced onto the equipotential surface of the geoid and it is given in the form of geoid heights by the Stokes integral (see 3.98, Chap. 3 in Part I). It is immediately evident that extra computational effort is required to reduce the gravity anomalies Δg, measured on the Earth's surface to the boundary surface of the geoid. The density ρ of the topographic masses and the othometric heights H are needed for this reduction. In most cases, a sufficient spatial coverage in Δg and H exists, but density information is rather limited and thus assumptions about ρ and its variations have to be made. In the conventional application of the problem, i.e. the Stokes one, the topographic masses are condensed to a mass layer on the geoid and the outcome of

this process is either the free-air or Faye gravity anomalies, so that the entire process may be viewed as the result of a mass reduction and a downward continuation assumption (Forsberg 1984, 2010).

Based on the computational fashion of the above described BVP, a direct and an indirect effect on gravity anomalies should be first taken into account. The former is the difference in the attraction between the masses above the geoid and the masses condensed on the geoid. The latter is due to change of the potential as well of the above mentioned masses (see for complete formulation in Chap. 10, Part II). These two effects constitute the mass reduction step in the Stokes BVP. Then, the such derived gravity anomalies (Δg_c) enter into the Stokes's integral and co-geoid heights (N_c) are determined, that is

$$N_c = \frac{R}{4\pi\gamma} \iint_{\sigma} \Delta g_c S(\psi) d\sigma, \tag{8.44}$$

where $S(\psi)$ is the Stokes function (see 3.23, Chap. 3.3, Part I of this book). Finally, the co-geoid heights N_c are transformed into geoid heights N by restoring the effect of the condensed masses on the geoid (*indirect effect on the geoid* δN), that is

$$N = N_c + \delta N \tag{8.45}$$

In the solution of Molodensky's BVP in its scalar version, the Earth's surface and its external gravity field are determined from the gravity anomalies and the potential given everywhere on the topography (Molodensky et al. 1962). Since the Earth's surface is unknown, this BVP is free and it is closer to physical reality than Stokes's BVP, since all measurements are taken on the Earth's surface. Moreover, this problem is basically a non-linear problem, but it can however be linearized by the selection of a specific surface, called telluroid, to approximate the actual Earth's surface and a potential U to approximate the Earth's gravity potential W (see Fig. 8.8). Then, the Earth's surface is represented by its deviation from the telluroid, called height anomalies ζ, which are a function of the difference between W and U, i.e., the disturbing potential T.

Molodensky's theory handles the problem of gravity anomalies referring to a non-level surface, i.e., the telluroid (see Fig. 8.8) without any removal of mass effects. In the basic form of this theory and using harmonic continuation, the gravity anomalies Δg_0 at a level surface passing through point P of the telluroid (see Fig. 8.8) can be derived by a sum of terms (see Moritz 1980; Sideris 1987a, and the detailed discussion in Part I of this book)

$$\Delta g_0 = \sum_{n=0}^{\infty} G_n, \tag{8.46}$$

where

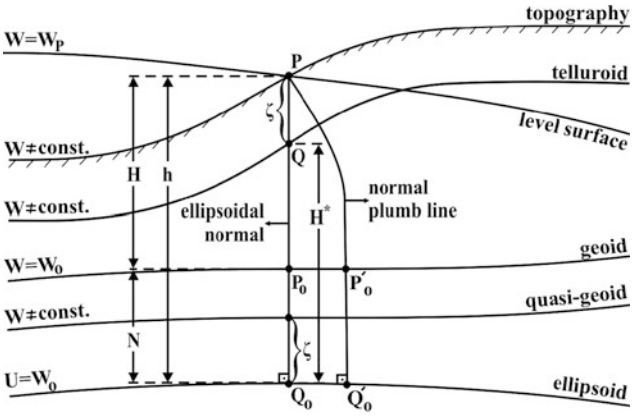


Fig. 8.8 The geometry of the Stokes and the Molodensky BVP

$$G_n = -\sum_{m=1}^{n} z^m L_m G_{n-m} \tag{8.47}$$

and ($G_0 = \Delta g,\ z = H_Q - H_P$, point Q at the telluroid). It should be stressed that the series giving the gravity anomalies at point level P (Fig. 8.8) consists of the free-air anomalies at ground level as the first term plus correcting terms dependent on the heights and the free-air anomalies. The first-order L-operator is given in planar approximation as

$$Lf = \frac{\partial f}{\partial z} = \frac{R^2}{2\pi} \iint\limits_{\sigma} \frac{f - f_P}{l_0^3} d\sigma, \tag{8.48}$$

where $l_0 = 2R\sin(\psi/2)$, ψ is the spherical distance between the running and the computation point and σ is the surface of the sphere of radius R.

After the computation of gravity anomalies Δg_0 on the level surface through point P the disturbing potential T can be obtained from Stokes's equation, and consequently the height anomaly ζ from Bruns's formula, as follows:

$$\begin{aligned} T_P &= \sum_{n=0}^{\infty} T_{nP} = \frac{R}{4\pi\gamma} \iint\limits_{\sigma} \Delta g_0 S(\psi)\, d\sigma = \sum_{n=0}^{\infty} \frac{R}{4\pi} \iint\limits_{\sigma} G_n S(\psi)\, d\sigma \\ &= \frac{R}{4\pi} \iint\limits_{\sigma} \Delta g S(\psi)\, d\sigma + \sum_{n=1}^{\infty} \frac{R}{4\pi} \iint\limits_{\sigma} G_n S(\psi)\, d\sigma, \end{aligned} \tag{8.49}$$

$$\zeta_P = \frac{T_P}{\gamma} = \frac{R}{4\pi\gamma} \iint\limits_{\sigma} \Delta g S(\psi)\, d\sigma + \sum_{n=1}^{\infty} \frac{R}{4\pi\gamma} \iint\limits_{\sigma} G_n S(\psi)\, d\sigma. \tag{8.50}$$

The solution for T in (8.49) consists of the classical Stokes formula plus the correcting terms computed from free-air gravity anomalies and heights. Equations 8.49 and 8.50, that give by analytical continuation the solution of T and ζ, contain 2D convolutions on the sphere and can be therefore evaluated by numerical integration, which is a time-consuming procedure. Since these formulas can be projected on to a plane, they become appropriate for computation by FFT techniques (see examples in Sect. 8.5.2). Limiting to the first two terms in (8.50), the height anomaly is given by

$$\zeta = \frac{R}{4\pi\gamma} \iint_{\sigma} (\Delta g + G_1)\, S(\psi) d\sigma, \tag{8.51}$$

where

$$G_1 = \frac{R^2}{2\pi} \iint_{\sigma} \frac{(H - H_P)}{l_0^3} \Delta g d\sigma. \tag{8.52}$$

Assuming linear correlation of gravity anomalies with height, G_1 can be approximated by the gravimetric (linear) terrain correction formula (8.11).

It is important to clarify that the classical free-air gravity anomalies used in (8.44) are not identical to the Molodensky-type free-air gravity anomalies used in (8.49). In Stokes's BVP the geoid heights N are computed from gravity anomalies on the geoid defined as (see Fig. 8.8)

$$\Delta g = g_{P_0} - \gamma_{Q_0}, \tag{8.53}$$

where g_{P_0} is usually computed from the measured g_P by using the gradient of normal gravity (see 8.17) that approximates the actual gradient of gravity $(\partial g/\partial H)$. The resulting anomalies are the classical free-air gravity anomalies, which, according to Helmert's condensation reduction (see Sect. 8.4.1), are a sufficient approximation of boundary Δg on the geoid. Then, the geoid heights are computed using (8.44).

In Molodensky's BVP the height anomalies ζ are computed at ground level (8.46) and the free-air gravity anomalies are obviously different from the previously defined classical ones and they are defined as (see also Fig. 8.7)

$$\Delta g = g_P - \gamma_Q \tag{8.54}$$

The gravity anomalies defined by (8.52) represent the data known at any point Q on the telluroid and they are used in (8.50) to compute height anomalies at point P, through which the level surface is passing (Fig. 8.8). The quasi-geoid determination realized by (8.50) breaks down to the following computational steps (Forsberg 2010):

- Downward continue gravity anomalies at the level surface.
- Apply Stokes's operator to gravity anomalies and compute height anomalies.
- Upward continue height anomalies to ground level.

This scheme is more stable if mass reduced gravity anomalies are used, although Molodensky's theory can be applied as well to the original free-air gravity anomalies. Therefore, Molodensky's theory and mass reductions are complementary and can be applied for geoid/quasi-geoid determination for optimal results (Forsberg 1984).

From Fig. 8.8 the following relationship is obtained that connects the different height systems used in the two BVPs:

$$h = \zeta + H^* = N + H, \tag{8.55}$$

and consequently

$$N = \zeta + (H^* - H) = \zeta + \delta\zeta \tag{8.56}$$

Given the definitions of the orthometric height H and normal height $H^*$and (1.90 and 1.198 in Part I, respectively) the following expression for $\delta\zeta$ is derived:

$$\delta\zeta = \frac{\bar{g} - \bar{\gamma}}{\bar{\gamma}} H \approx \frac{\Delta g_B}{\gamma} H, \tag{8.57}$$

where Δg_B is the Bouguer anomaly at point P, while for the mean gravity $\bar{g}$ and mean normal gravity $\bar{\gamma}$ see the definitions in Sect. 1.11 of Chap. 1 in Part I of the book. In (8.57) $\delta\zeta$ can be interpreted as a correcting term for the upward continuation of geoid heights N from, e.g., sea level to ground level (see 8.49). For more details about this continuation procedure the reader should consult Sideris (1987a).

Whenever masses are moved, compensated and condensed, a change in the gravity potential is caused. Therefore, when applying the mass reductions to gravity anomalies in order to compute geoid heights, this change in potential should be accounted for (see Sect. 8.3.2 before and Sect. 10.2.3 in Chap. 10 of Part II). In order to compute the actual geoid surface we need to restore to the heights computed this change in gravity potential that was caused by the topographic reduction. This effect is called the indirect effect (see Chap. 10) and can be computed by Bruns's formula if we denote by δT the change of the gravity potential at the geoid

$$\delta N = \frac{\delta T}{\gamma}, \tag{8.58}$$

where

$$\delta T = T - T_{cond/comp}, \tag{8.59}$$

and T is the gravity potential of the actual topographic masses and $T_{cond/comp}$ the potential of the masses condensed (Helmert's method, Sect. 8.4.1) or compensated. The potential of the topographic masses T and that of the compensated masses according to the PH and AH isostatic models are given as follows (see, Forsberg 1984; Peng 1994; Bajracharya 2003):

$$T = G \iint_E \int_0^H \frac{\rho}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{1/2}} dx dy dz, \tag{8.60}$$

$$T_{comp/PH} = G \iint_E \int_{-D}^0 \frac{\Delta\rho}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{1/2}} dx dy dz, \tag{8.61}$$

$$T_{comp/AH} = G \iint_E \int_{-T-d}^{-T} \frac{\Delta\rho}{\left[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2\right]^{1/2}} dx dy dz. \tag{8.62}$$

The integrals in (8.60)–(8.62) can be numerically integrated by the prism method (see Sect. 8.5.1) or expressed in general convolution form and evaluated by 3D or 2D FFT (see Sect. 8.5.2). Additionally, by substituting the previous integrals in (8.59) the indirect effect of the geoid for the different reduction schemes can be derived.

A simplified scheme for the effect of the compensated masses to geoid heights or height anomalies, i.e., the indirect effect ΔN_{comp}, can be obtained after an expansion of the kernel function $(1/l)$. It holds in an approximate form (Forsberg 1984, 2010)

$$\frac{1}{l} \approx \frac{1}{l_0} - \frac{d_0}{l_0^3}(z - d_0) \tag{8.63}$$

where d_0 is a suitable reference level of the depth of compensated masses (see Sect. 8.3.3 and Fig. 8.5). It finally holds (Forsberg 1984, 2010)

$$\Delta N_{comp/AH} = \frac{G}{\gamma}\left(\Delta\rho\, f_1 * d + \Delta\rho\, f_2 * d^2\right), \tag{8.64}$$

which is the abstract form of a set of two convolutions in d and d^2 that can be evaluated by 2D FFT; the functions f_1 and f_2 in this case are (Forsberg 2010):

$$f_1 = \frac{1}{l_0} - \frac{d_0\,(T_0 - d_0)}{l_0^3}, \quad f_2 = -\frac{d_0}{2l_0^3}. \tag{8.65}$$

Regarding the different reduction methods presented in Sect. 8.3 and the above formulas as well, we can claim that the Bouguer reduction has larger indirect effect on the geoid compared to isostatic models and Helmert's scheme (Sect. 8.4.1). The latter has generally small indirect effects although it leads to a rough gravity anomaly field.

For a detailed discussion on the indirect effect Wichiencharoen (1982) can be consulted. Numerical results, especially in areas characterized by terrain roughness, where this effect reaches the level of tens of centimeter or even more, can be found in different studies (see, e.g., Tziavos et al. 1992; Bajracharya 2003).

8.4 Terrain Effects in Geoid and Quasi-geoid Determination

In this section the Helmert's second method of condensation and the RTM reduction scheme are discussed, which are more representative than the terrain reductions given in Sect. 8.3, towards the direct treatment of topographic masses in geoid and quasi-geoid modeling. Moreover, Rudzki's reduction is briefly commented separately, since it has practically no indirect effect on the geoid. Finally, the terrain effects on geoid and quasi-geoid heights computed from high-resolution DTMs, especially in rugged terrains, are discussed in connection with the use of high-resolution GGMs, towards the estimation and reduction of the geoid signal omission error.

8.4.1 Helmert's Second Method of Condensation

As it was mentioned in Sects. 8.3.2 and 8.3.5 before, the free-air downward continuation ignores the masses between the Earth's surface and the geoid, and consequently gravity is reduced from the topographic surface to the geoid using the vertical component of the gravity gradient. It was also commented that the Bouguer reduction removes completely the topographic masses and the contribution of the topography is taken into account through a remove-restore scheme. In Helmert's second method of condensation, which is one of the most common reduction schemes used in gravimetric geoid computation in local and regional applications, the masses are shifted and condensed to a layer on the geoid (see, e.g., Heiskanen and Moritz 1967; Wichiencharoen 1982; Wang and Rapp 1990; Martinec and Vaníček 1994; Martinec 1998; Tenzer et al. 2003; Forsberg 2010; Sideris 2010). More specifically, the topographic masses of volume density ρ are shifted and condensed to a surface layer of surface density σ ($\sigma = \rho H$) along the plumb line (see also Fig. 10.1 in Chap. 10 of Part II). There is additionally another method of Helmert's condensation defined in a different way, named the first method of Helmert's condensation (see, e.g., Heiskanen and Moritz 1967; Heck 2003b). In this first method the masses are condensed on a surface parallel to the geoid and located 21 km below the geoid, contrary to Helmert's second method.

The gravity anomalies reduced to the geoid by the aforementioned reduction are directly associated with the attraction change from the surface of the Earth to the geoid surface and this change is given by (8.59), where the first term is represented by (8.3) and the second one approximately by (8.6). In this sense, a gravity anomaly on the geoid obtained by the second method of Helmert's condensation differs from the free-air gravity anomaly Δg_{FA} by the amount of terrain correction and this kind of gravity anomaly is called Faye anomaly (Δg_{Faye}) and some times Helmert anomaly, i.e.,

$$\Delta g_{Faye} = \Delta g_{FA} + c. \tag{8.66}$$

Faye anomalies are additionally subject of a further correction, since due to the shifting of masses the potential changes as well (indirect effect of the potential). Due to this potential change, when using Δg_{Faye} the so-called co-geoid is primarily computed. Thus, before applying Stokes's equation, the gravity anomalies must be transformed from the geoid to co-geoid by applying a correction $\delta\Delta g$ called the indirect effect on gravity or the secondary indirect effect (see, e.g., Wichiencharoen 1982)

$$\delta\Delta g = 0.3086\delta N, \tag{8.67}$$

where δN is the separation between the geoid and co-geoid (indirect effect on the geoid, see also Sect. 8.3.5). In (8.67) $\delta\Delta g$ is given in mGal when δN is given in meters. Analytical expressions for δN are given in different researches (see, e.g., Wichiencharoen 1982; Tziavos et al. 1992) and in Sect. 10.2.2 in Part II of this book, in conjunction with a gravimetric geoid computation scheme based on Stokes's integral.

Finally, in a more complete way and taking into account both the corrections due to the changes of the attraction and the potential, the Faye gravity anomalies are given by the formula:

$$\Delta g_{Faye} = \Delta g_{FA} + c + \delta\Delta g. \tag{8.68}$$

Generally, Faye anomalies are not smooth at all although they produce a very small indirect effect on the geoid, present a perfect correlation with height, higher than that of free-air gravity anomalies (Forsberg 1984, 2010) and are used frequently as input data in gravimetric geoid determination through the Stokes's integral.

8.4.2 *Rudzki's Inversion Scheme*

Rudzki's reduction or Rudzki's inversion is the only gravimetric reduction which does not change by definition the equipotential surface and thus has zero indirect effect in geoid height computations (Heiskanen and Moritz 1967; Bajracharya 2003). The masses above the geoid are inverted below the geoid and the masses produced in this way are called *mirrored masses*. The geometry of this effect in planar approximation is shown in Fig. 8.9. Although the potential of the topography and the mirrored topography are equal (zero indirect effect on the geoid heights), the attractions of the topography and the inverted topography are not equal.

In an analogous way to the Bouguer and RTM reduction schemes, the gravitational attraction of all topographic masses above the geoid in Rudzki's reduction is represented by the sum of the attraction of the regular and the irregular part of the topography and the following equation is valid for a point P (see, Bajracharya 2003):

$$A_{top/R} = 2\pi G\rho H_P - G \iint_E \rho \left[\frac{1}{s_0} - \frac{1}{\left[s_0^2 + (H - H_P)^2 \right]^{1/2}} \right] dE, \tag{8.69}$$

where the different quantities are shown in Fig. 8.9. The corresponding attraction at point P due to the inverted (mirrored) masses of density ρ' is given as follows:

$$A_{inv/R} = 2\pi G\rho' H'_P - G \iint_E \rho' \left[\frac{1}{\left[s_0^2 + (H_P + H')^2 \right]^{1/2}} - \frac{1}{\left[s_0^2 + (H_P + H'_P)^2 \right]^{1/2}} \right] dE. \tag{8.70}$$

Then, the direct topographic effect on gravity of all masses, those above the geoid and the inverted ones, is given as (Bajracharya 2003):

$$\Delta A_R = A_{top/R} - A_{inv/R} = G \iint_E \rho \left[\frac{1}{s_0} - \frac{1}{\left[s_0^2 + (H - H_P)^2 \right]^{1/2}} + \frac{1}{\left[s_0^2 + (H_P + H')^2 \right]^{1/2}} - \frac{1}{\left[s_0^2 + (2H_P)^2 \right]^{1/2}} \right] dE. \tag{8.71}$$

It is evident from this last equation that the attractions due to the regular part of the topographic masses and the mirrored topographic masses are equal and cancel out (see Bajracharya and Sideris 2004). Finally, Rudzki's anomalies can be computed as:

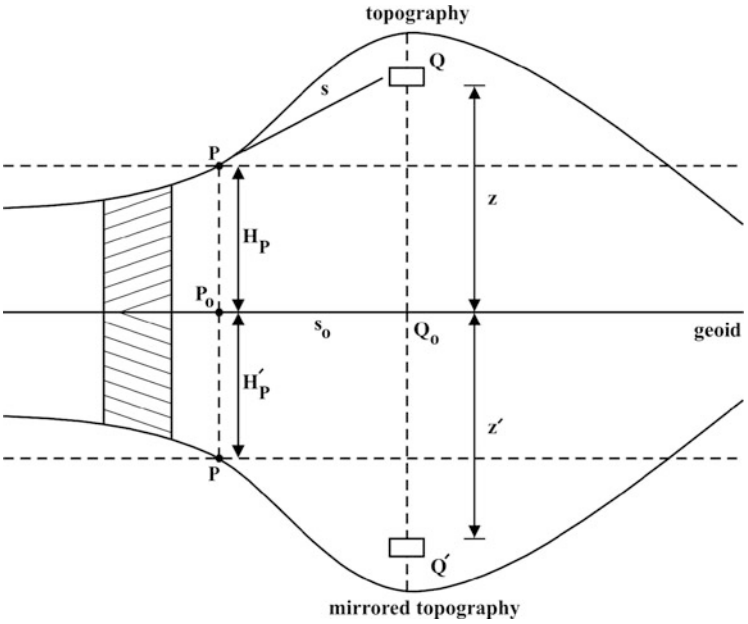


Fig. 8.9 The geometry of Rudzki's reduction in planar approximation (After Bajracharya 2003)

$$\Delta g_R = g - \gamma_o + F - \Delta A_R. \tag{8.72}$$

Using the Rudzki reduction scheme in gravimetric geoid computation in the Canadian Rockies, comparable results were reported with those derived by Helmert and RTM reductions, while Rudzki's geoid had smaller bias than Helmert and RTM corresponding geoid models (Bajracharya 2003; Bajracharya and Sideris 2004).

8.4.3 Residual Terrain Model (RTM)

The Residual Terrain Model (RTM) is one of the most common mass reduction methods used mainly in quasi-geoid determination. Within this scheme the contribution of the topography is removed and restored using a model of the topography equal to the difference between the true topography and a reference, smooth but varying, elevation surface (see Fig. 8.10). Therefore, the topographic masses above this reference surface are removed and masses fill up the deficits below this reference surface. The reference surface can be constructed by averaging the fine (detailed) resolution topography grid and then low-pass filtering the average grid generated by taking moving averages of an appropriate number of adjacent blocks. In practical applications the detailed topographic grid is used out to a maximum distance and the coarse grid is used for the remaining topography of the area under study (Forsberg 1984, 2010). The radius of the inner zone depends on the resolution of the

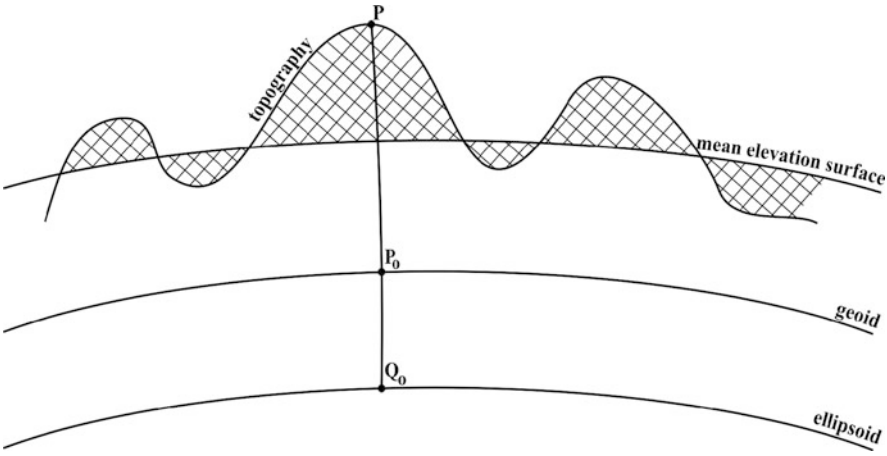


Fig. 8.10 The geometry of the RTM reduction

available detailed DTM. For points outside this inner area, a lower-resolution grid can be used derived from the fine one by averaging or from a spherical harmonics expansion model of the Earth's topography (e.g., Sünkel 1986; Abd-Elmotaal 1995; Tsoulis et al. 2007; Hirt et al. 2010).

The consideration of the coarse/detailed grid system increases the computation speed and this scheme is fully implemented in the GRAVSOFT software, especially within the TC program (Tscherning et al. 1992; Forsberg 2010; Tscherning 2010) widely used in applications of gravity field modeling. Based on this software, and in a small inner zone around the computation point, a further densification of the topographic data is made by a bicubic spline interpolation technique in order to integrate the often large effects of the inner zone (Forsberg 2010). This densification is necessary in cases where the topography is approximated by prisms and it is possible some times for the computation point to be located at the edge of a prism and consequently the computed terrain effects to be unrealistic. It is to be noticed that in numerical applications in a restricted region, the detailed (fine) grid can be used for the entire test area, since the numerical burden is not significant taking also into account nowadays computer facilities.

Following Forsberg (1984) the topographic effect on gravity of the RTM reduction is computed as

$$\Delta A_{RTM} = 2\pi G\rho\left(H - H_{ref}\right) - G\iint_{E}\int_{H_{ref}}^{H}\rho\frac{H_P - z}{l^3}\,dxdydz, \tag{8.73}$$

where H_{ref} represents the height of the reference surface used, H the height of the topographic masses according to the fine resolution DTM, while the integral term is the terrain correction c (see, e.g., 8.5 and 8.11). The first term in (8.73) is the difference of two Bouguer plates with different thickness. The thickness of the first

one is realized by the height of the computation point and that of the second plate by the height of the reference surface. This scheme implies that the masses above the geoid are first removed by the complete Bouguer reduction and then are restored with the reference Bouguer plate. Given the topographic effect on gravity through the RTM reduction (ΔA_{RTM}), the RTM gravity anomalies are given by the following formula:

$$\Delta g_{RTM} = g - \gamma_o - \delta A_{RTM}. \tag{8.74}$$

It is to be noticed that if RTM gravity anomalies are used in (8.50) then the quasi-geoid is obtained instead of the geoid. A correction term should be applied to convert the quasi-geoid to the geoid (see 8.57). The main advantage of the RTM reduction is that the reduced gravity anomalies are generally smoother than those resulting from other reduction methods. Depending on whether the topography of an area is above or below the reference elevation surface, the topographic RTM density anomalies will create a set of positive and negative anomalies (Forsberg 1984, 1985). As a result, the topographic effect contributions are computed up to a specific distance from the computation point, thus minimizing the effects of the far topographic masses, since the RTM density anomalies will in general cancel out at large distances from the computation point (Forsberg and Tscherning 1981). Additionally, another advantage of the RTM reduction is that the quantity to be restored, in the restore step of the computation of height anomalies, is very small compared to the indirect effect on the geoid from other methods (e.g., Helmert) and no assumption about isostatic compensation is needed as in the isostatic reductions. The main disadvantage of the RTM method is that the gravity potential will no longer be a harmonic function at those stations below the mean elevation surface. If, for instance, a station is located in a valley, then, after applying the RTM method it will be located inside the smooth topographic surface determined by the boundaries of the reference elevation surface used (Forsberg 1984). In that way, the final value of the reduced observation will be inside the topographic masses, where the potential is no longer a harmonic function. To answer that question, Forsberg (1984) indicated that the density above a plane through that station could be condensed in a mass plane layer immediately below the station. In that way, geoid heights and deflections of the vertical will be almost unchanged due to the small slope of the smooth reference surface. On the other hand the gravity anomalies will be changed and for the modern statistical methods of gravity field approximation to be used, which request the harmonicity of the field, a special correction known as *harmonic correction* should be applied. For more details about the RTM reduction method Forsberg (1984, 1985) and Forsberg and Tscherning (1981) should be consulted.

The potential of the topographic masses in the RTM reduction is given as (compare with 8.60–8.62):

$$T_{RTM} = G \iint_{E} \int_{Href}^{H} \frac{\rho}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{1/2}} dxdydz. \quad (8.75)$$

The RTM reduction method gives primarily height anomalies and the quasi-geoid and the restored terrain effect on the quasi-geoid is given by the following equation:

$$\Delta\zeta_{RTM} = \frac{G}{\gamma} \iint_{E} \int_{Href}^{H} \frac{\rho}{[(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{1/2}} dxdydz \quad (8.76)$$

In an approximated convolution form (8.70) reads as (Forsberg 1984, 2010):

$$\Delta\zeta_{RTM} \approx \frac{G\rho}{\gamma}(H - H_{ref}) * \frac{1}{l_0}, \quad (8.77)$$

where l_0 denotes planar distance. It should be noted also that the RTM effects are usually computed over large regions, so that spherical FFT methods offer advantages in numerical evaluations (see, e.g., Vergos et al. 2005a; Barzaghi et al. 2009; Denker et al. 2009; Forsberg 2010; Sansò et al. 2008; Featherstone et al. 2011). Further derivation on the evaluation of the previous integrals either by NIM or FFT are reported in Sect. 8.5.

8.4.4 *Terrain Effects and High-Resolution Global Geopotential Models*

The theory of the different mass reduction methods is applied in practice relative to a reference field represented by a GGM series expansion. The computation of the anomalous potential T by a set of spherical harmonic coefficients, as, e.g., that of the recent Earth Gravitational Model 2008 – EGM2008 (Pavlis et al. 2008) is carried out as

$$T_{EGM} = \frac{GM}{r} \sum_{n=2}^{n_{\max}} \left(\frac{a}{r}\right)^n \sum_{m=0}^{n} \left(\overline{\delta C}_{nm} \cos m\lambda + \overline{S}_{nm} \sin m\lambda\right) \overline{P}_{nm}(\cos\theta), \quad (8.78)$$

where n and m are the maximum degree and order of the harmonic expansion (2,160 for EGM2008 with some additional harmonic coefficients up to degree 2,190), GM is the geocentric gravitational constant and a the semi major axis (scaling parameters of EGM2008), $\overline{P}_{nm}(\cos\theta)$ are the fully normalized associated Legendre functions and the term $\overline{\delta C}_{nm}$ denotes that the zonal harmonics of the reference ellipsoid have been removed from the $\overline{C}_{nm}$ coefficients of the EGM2008; see for details Chap. 6, Part II of this book. The expansion of (8.78) is a function in space and the computation is primarily carried out at an elevation ($r = R + H_P$). In practice the reference effects are computed in grids at a constant elevation and

the height anomalies ζ_{EGM} can be derived from the above mentioned expansion in the required grids by applying Bruns's equation. The entire numerical approach is completed through a remove-restore procedure (Forsberg 1984, 2010).

Even though ultra high degree and order GGMs, like EGM2008 (spatial resolution $5'$ or approximately 10 km), are available nowadays for the computation of quasi-geoid heights and other components of the Earth's gravity field, these models are not capable to represent the high-frequency band of the spectrum apparent at scales finer than the spatial resolution of the GGM used. This problem is known as the *omission error* in gravity field modeling. Hence, quasi-geoid heights, i.e., height anomalies ζ, computed from a GGM are affected by this error that may exceed 10 cm in mountainous terrains even when EGM2008 is used, since the high-frequency gravity signal cannot be represented by a truncated spherical harmonic series expansion (Hirt et al. 2010). The problem is more serious in rugged terrains without sufficient gravity data coverage. In such cases a high resolution RTM data set may be used for quasi-geoid omission error estimates as it is proposed, e.g., by Hirt et al. (2010). This RTM-based omission error estimates may improve significantly the EGM2008 height anomalies at a level of approximately 50% as it has been realized from numerical researches (e.g., Hirt et al. 2010). The RTM data set in the above mentioned methodology can be constructed as the difference between a high-resolution DTM, e.g., a $3''$ SRTM-based elevation model (see Sect. 8.3.1) and a high-degree and order spherical harmonic expansion of the Earth's topography, as, e.g., the DTM2006.0 model (Pavlis et al. 2007a) that serves as a high-pass filter removing the long-wavelength features from the SRTM data (e.g., Hirt et al. 2010; Tziavos et al. 2010). The elevations from DTM2006.0 can be computed by a spherical harmonic expansion of the following form (Pavlis et al. 2007a)

$$H_{DTM2006.0} = \sum_{n=2}^{n_{\max}} \sum_{m=0}^{n} \left(\overline{HC}_{nm} \cos m\lambda + \overline{HS}_{nm} \sin m\lambda\right) \overline{P}_{nm}(\cos\theta), \qquad (8.79)$$

with the maximum degree of the expansion being set to 2,160. The RTM height data derived from the combination of SRTM and DTM2006.0 models ($H_{RTM} = H_{SRTM} - H_{DTM2006.0}$) are then transformed to RTM-based height anomalies (ζ_{RTM}) using, e.g., the prism integration method (Forsberg 1985; Nagy et al. 2000). These height anomalies contain additional spectral power beyond the band of the spectrum covered by the EGM2008 height anomaly and thus reduce the omission error affecting the finally computed height anomalies. This RTM-based methodology, applying RTM omission error estimates to EGM height anomalies can be a promising alternative for the improvement of quasi-geoid models in areas characterized by rugged terrains and the lack of gravity data.

The standard remove-restore concept for quasi-geoid determination based on the regular use of the RTM method should be still exploited even in areas with rugged topography but with sufficient gravity anomaly coverage. Moreover, the regional quasi-geoid modeling described before by applying RTM omission error estimates to EGM2008 height anomalies could be employed in a modified form by replacing

the global expansion of the Earth's topography with a coarser height grid in the wider area of interest either based on SRTM or national DTMs.

8.4.5 *The Remove-Restore Methodology and the Different Reduction Schemes*

In gravity field modeling in general and in geoid prediction height in particular, the mass or terrain reductions are applied to gravity anomalies in a remove-restore fashion, as it was already discussed in previous chapters. First, the mass effects are removed from the available observations, then predictions are carried out and in the final step the mass effects are restored. This procedure is usually combined in practice with a simultaneous remove-restore procedure of the contribution of a GGM (e.g., EGM2008) that forms the reference field within the entire process.

In a gravimetric geoid model (see 10.14 in Chap. 10, Part II), the general idea of the remove-restore technique is to use:

(a) The global geopotential model for the recovery of long-wavelength structures,
(b) The topography through a mass model or DTM to represent the short-wavelength components (in order to smooth the data and avoid aliasing effects), and,
(c) The terrestrial gravity data for the computation of medium to short-wavelength features of the gravity field.

In the following five theoretical examples are given on the use of different reduction schemes in the frame of the remove-restore technique. The original ideas of these examples are given in Sideris (1987), Omang and Forsberg (2000), Vergos et al. (2005b), Forsberg (2010), Tziavos et al. (2010) and have been properly modified in the present work. In the first four examples the remove-restore effect of the geopotential model is omitted, since they primarily focus on how to handle the mass effects in different computational schemes. Additional numerical examples are given in Sect. 7.3 (Chap. 7 in Part II) where the geoid/quasi geoid heights are computed by the LSC method.

A. Geoid/quasi-geoid height prediction from gravity anomalies – direct use of mass reductions

The theory handles gravity anomalies without any removal of topographic masses towards the smoothing of the gravity field. It is based on the harmonic continuation since the data primarily refer to a non-level surface.

(a) Apply mass reductions to Δg_{obs} ($\Delta g_{red} = \Delta g_{obs} - \Delta g_H$).
(b) Gridding of the reduced Δg_{red} ($\Delta g_{red} \rightarrow \Delta g_{grid}$).
(c) Predict reduced geoid/quasi-geoid heights (e.g., $N_{grid} = S\ (\Delta g_{grid})$ or ζ_{grid}).
(d) Restore mass effect ($N = N_{grid} + N_H, \zeta = \zeta_{grid} + \zeta_H$).

The advantage of this mass remove-restore scheme is that the reduced gravity anomalies are smooth enough, with low variability and easy to grid. Thus, in step

(c) the errors in the geoid prediction are minimized. Usually, within this procedure either the classical Helmert mass reduction scheme can be used or the RTM model, but it can easily handle all the mass reductions presented in Sects. 8.3 and 8.4.

B. Geoid/quasi-geoid height prediction from gravity anomalies – indirect use of gravity anomalies

(a) Apply mass reductions to Δg_{obs} ($\Delta g_{red} = \Delta g_{obs} - \Delta g_H$).
(b) Gridding of the reduced Δg_{red} ($\Delta g_{red} \rightarrow \Delta g_{grid}$).
(c) Restore mass reduction effect (e.g., simple Bouguer anomaly term), and produce free-air gravity anomalies or Faye anomalies ($\Delta g_{Faye,\,grid} = \Delta g_{grid} + \Delta g_H$).
(d) Predict final geoid/quasi-geoid (e.g., $N_{grid} = S(\Delta g_{Faye,\,grid})$ or ζ_{grid}).

The drawback of this procedure is that in the prediction of the final geoid/quasi-geoid, the full variability of the gravity field should be handled after step (c) and probably large errors can be propagated into the computed heights. In the case of Molodensky's approach higher order terms can be used in order to reduce the errors. It should be noticed that the classical Helmert/Stokes theory or Molodenky's one are applied in step (d). The terms *direct* and *indirect* use of the terrain reduction in the previous two examples has been introduced by Forsberg (2010).

C. Quasi-geoid height prediction from gravity anomalies using Molodensky's theory

The Molodenky approach can handle gravity anomalies without any removal of the topographic masses and is based on the harmonic continuation of gravity observations which finally refer to the level surface by a sum of harmonic terms (see 8.47). The harmonic continuation of the gravity observations (downward/upward scheme) can be realized by a second-order gradient T_{zz} that can be computed from, e.g., Δg_{obs} (Forsberg 1984, 2010). This computational step is implemented in GRAVSOFT software (Tscherning et al. 1992, Forsberg 2010; Tscherning 2010) widely used in gravity field applications.

(a) Predict vertical gravity gradient T_{zz} from Δg_{obs}.
(b) Downward continue at level surface $\Delta g_0 = \Delta g_{obs} - T_{zz} H$.
(c) Compute quasi-geoid heights form Δg_0 at level surface (e.g., Stokes' operator).
(d) Upward continue the height anomalies ζ_0 by the T_{zz} and compute ζ at the topography.

The downward continuation procedure is realized through the L surface operator (see 8.48). The upward continuation of the surface height anomalies ζ_0 through T_{zz} is carried out using (8.49) and (8.50) in a slightly modified form (Forsberg 2010):

$$\zeta = \zeta_0 + \sum_{n=1}^{\infty} \frac{1}{n!} z^n L_n \zeta_0. \tag{8.80}$$

The scheme can be more stable if mass reductions are used for gravity anomalies and the vertical gradient, which will be presented clearer in the analysis of the next example. Using Molodensky's theory in combination with mass reductions, optimal results can be obtained. In case the computations stop at the first-order term without mass reductions, the scheme results in an integral over the heights squared which is similar to the terrain correction integral (compare 8.11 and 8.52). When reduced gravity data are used and higher-order terms of the Molodensky series are employed, the procedure is more flexible when it is combined with RTM as it is shown in example *E* below.

D. Geoid height prediction from Helmert (Faye) anomalies

In this example, the complete removal of the masses is carried out using the full topographic effect (8.15), or in other words the masses are condensed by shifting them to a mass layer on the geoid (Sect. 8.3.5). The different steps of this procedure are summarized as follows:

(a) Remove the complete topographic effect (8.15) and produce complete Bouguer gravity anomalies (Δg_B).
(b) Downward continue Δg_B at the level surface, although T_{zz} is very small now and this step can be eliminated.
(c) Restore condensed topography and compute Faye anomalies.
(d) Predict geoid heights by applying the Stokes's operator.

The produced Faye gravity anomaly field is not smooth at all, although this scheme can be regarded as composed by a terrain reduction part and a downward continuation, but the effect of the latter assumption is negligible, since T_{zz} is much smoother now than in example *C* before.

E. Gravity database generation and geoid/quasi-geoid computation using a combined reduction procedure (e.g., RTM/Helmert)

The methodology of this example has been applied in the past for the creation of high-resolution gravity databases and the subsequent computation of various geoid solutions not only using FFT techniques (Omang and Forsberg 2000) but also spatial methods like LSC (Vergos et al. 2005a, b). The topography and bathymetry effects are of importance in the various stages of the procedure, either to smooth the original gravity field or to compute the final geoid/quasi-geoid model in the restore step.

The methodological scheme of this example can be developed in alternative ways, but as outlined below it contains all the stages for a complete geoid modeling, as it was computed in a $5^0 x\ 8^0$ test area in south Aegean Sea, eastern Mediterranean (Vergos et al. 2005b).

(a) *Free-air gravity data base processing and validation*

First, all the available gravity anomalies data, i.e., free-air gravity anomalies are collected and merged for both sea and land areas. The data are then reduced

by a GGM (EGM96 in this case) and the full topographic effect is subtracted. Blunders and gross errors in the so-reduced data are indentified and removed using least-squares collocation (Tscherning 1991; Vergos et al. 2005b). Finally, the GGM contribution and mass effects are restored in order to construct a "clean" free-air gravity data base.

(b) *Construction of a free-air gravity database*

The free-air gravity anomalies from step (a) are referenced to EGM96 and RTM reduced. The residual gravity anomalies are then gridded using LSC to a regular grid ($1'$ in the example given by Vergos et al. (2005b)). Finally, the EGM96 and RTM effects are restored to the gridded residual gravity anomalies in order to construct the final gravity anomaly database.

(c) *Geoid/quasi-geoid modeling*

The reduced free-air gravity anomalies of step (b) are used to estimate residual geoid heights both by FFT and LSC. Finally, the EGM96 and RTM effects are restored to the gridded residual geoid heights in order to construct the final $1'$ geoid model for the test area. For the transformation between geoid and quasi-geoid models the corresponding correction (8.57) is applied. The contribution of the RTM reduction scheme to the gravity anomalies already reduced to the GGM (EGM96) used is tabulated below. From these results are evident the advantages of the RTM method related to (a) the optimal computation of mass effects in a spherical cap around the computation point through a properly selected reference field ($5'$ in this case), (b) the significant reduction of the remote residual topography and (c) the considerable smoothing of the range of the residual gravity field.

$\Delta g_{red/EGM96}$		Before	After
		RTM reduction	
Variance	[mGal^2]	717	264
Mean value	[mGal]	−3.4	−0.5
Range (min/max)	[mGal]	314	162

8.5 Methods for the Numerical Estimation of Direct and Indirect Topographic Effects

In the past, due mainly to the lack of computer availablity, mass reductions and their effects to gravity were computed by the aid of overlays on maps, subdivided in concentric circles and radial sectors forming zones around the computation point, where mean elevations were read or digitized from maps (Forsberg 1984, 2010). Then, the mass effects, e.g., terrain corrections, were summed up using either tables

or simple calculations based on closed gravitational formulas corresponding to regular geometric representations (e.g., prisms, cilinders). Similar overhead zone systems with different names (e.g., *Hammer zones*, *Hayford zones*) were used to compute terrain corrections in a number of rings ranging from the computation point and they were extensively used in geophysical prospecting in the past as well as in local and regional gravity field modeling (Forsberg 1984, 2010). According to the above systems, mass effects were theoretically computed globally, although in practice the calculations were extended to a smaller distance from the computation point (e.g., 167 km in Hayford system).

The above mentioned traditional and more or less approximative techniques were progressively replaced by rigorous mathematical formulations implemented either by NIM or through efficient computational algorithms (e.g., Fourier transforms, Hartley transforms). The above alternatives take also advantage from the available nowadays high resolution models of topography and bathymetry and the enormous capabilities of modern computer systems. Generally, there are two types of methods used in mass reduction computations, i.e., the space (spatial domain) methods and the spectral (frequency domain) ones. The former contain different prism-based mass representations through, e.g., flat-top or inclined-top prisms, tesseroids, cylinders, and they are evaluated by NIM, which are generally rigorous but very time-consuming methods, especially in cases that large amounts of gridded data are handled (see, e.g., Nagy 1966; Forsberg 1984; Nagy et al. 2000; Smith 2000; Biagi and Sansò 2001; Heck and Seitz 2007; Wild-Pfeiffer 2008). The latter are mainly represented in gravity field modeling applications by FFT and FHT methods, which are computationally very efficient mainly in cases of large grids of height data. Usually, FFT algorithms are employed in 2D approximation form including second or higher-order terms and from the numerical point of view they give identical results with those by NIM, when they are combined with zero-padding techniques for the elimination of edge effects or circular convolution (see, e.g., Schwarz et al. 1990; Li 1993; Tziavos 1993; Li and Sideris 1994; Peng 1994). In early nineties, the FHT has been introduced in physical geodesy applications that works with real operations, contrary to FFT that is based on complex operations (e.g., Li 1993; Li and Sideris 1994; Sideris 2010). The FHT technique presents several computational advantages over the classical FFT mainly in terms of computer memory and CPU time needed in computations.

Given the fact that the rigorous formulas for the different reduction schemes (e.g., 8.5, 8.26, 8.33 and 8.73) and the indirect effects (e.g., 8.60–8.62) are all 3D convolution integrals, the 3D FFT or FHT (see Chap. 10 in Part II) methods have been also used in gravity field modeling to compute them effectively without assumptions and approximations. Additionally, through these rigorous spectral techniques, we can overcome some limitations and problems unique in 2D spectral methods, as, e.g., numerical instabilities, using only some of the terms in a series expansion of the rigorous formulas and avoiding 3D density information when available (Peng 1994). It is worth mentioning that the 3D spectral approach is a time-consuming method and several authors in the past developed different strategies (e.g., linear approximation formulas, filtering) to solve 3D convolutions by 2D

FFT/FHT techniques (e.g., Tziavos et al. 1988; Harrison and Dickinson 1989; Sideris 1990; Vermeer and Forsberg 1992). More details and formulation about 3D FFT are given in Chap. 10, Part II of this book.

In the next two sections we mainly focus on the formulation of the mass prism evaluation by NIM and FFT along with some examples and applications, as, e.g., the terrain effects on airborne gravity gradiometry data by the 2D spectral approximation.

8.5.1 The Mass Prism Topographic Model and the Numerical Integration Method (NIM)

The rectangular prism representation of topographic and bathymetric masses has been widely used during the last decades for numerical integration of the mass reductions schemes presented in the previous sections. Closed formulas for the potential and the attraction of the topographic and bathymetric masses can be derived from the original rigorous integral equations and transformed to a series expansion as it is shown below.

In practical applications the topography/bathymetry is usually represented by a 2D DTM/DBM, where the height/depth of each cell is represented by a prism with a mean height and a mean density. For a point P coinciding with the origin of the coordinate system and assuming constant density ρ, the potential of a prism defined by the intervals $(x_1 - x_2,\ y_1 - y_2,\ z_1 - z_2)$ will be given by the formula (Nagy 1966; Forsberg 1984; Nagy et al. 2000)

$$T(x,y,z) = G\rho \left|\left| xy\ln(z+r') + xz\ln(y+r') + yz\ln(x+r') - \frac{x^2}{2}\arctan\left(\frac{yz}{xr'}\right) - \frac{y^2}{2}\arctan\left(\frac{xz}{yr'}\right) - \frac{z^2}{2}\arctan\left(\frac{xy}{zr'}\right) \right|_{x_2}^{x_1} \Bigg|_{y_2}^{y_1} \Bigg|_{z_2}^{z_1}, \tag{8.81}$$

where $r'(x,y,z)$ in this case is the distance kernel defined as $r' = (x^2 + y^2 + z^2)^{1/2}$.

The rigorous terrain correction formula (8.5), when the NIM is applied and in case the masses are represented by a 2D (NxM) digital grid, can be evaluated as (Peng 1994)

$$c\,(x_P, y_P, z_P) = G\rho \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} \left\{ (x_P - x)\,\ln\left[(y_P - y) + r\right] + (y_P - y)\ln\left[(x_P - x) + r\right] - (z_P - z)\arctan\left[\frac{(x_P - x)\,(y_P - y)}{(z_P - z)\,r}\right] \right\} \Bigg|_{x-\Delta x/2}^{x+\Delta x/2} \Bigg|_{y-\Delta y/2}^{y+\Delta y/2} \Bigg|_{H_P}^{H_{nm}}, \tag{8.82}$$

where $r = [(x_P - x)^2 + (y_P - y)^2 + (H_P - z)^2]^{1/2}$. This is a very time-consuming method and usually is evaluated by 3D FFT as it is shown in Sect. 10.4.2 (Chap. 10 in Part II of this book). Equation 8.82 can be used for the computation of the topographic effect on airborne gravity data at an altitude z_0 by simply replacing ($z_P = H_P$) by z_0. In order to reduce the computational burden, (8.5) can be re-written in a series form (Li 1993)

$$c\,(x_P, y_P, z_P) = G\rho \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \int_{x_n-\Delta x/2}^{x_n+\Delta x/2} \int_{y_n-\Delta y/2}^{y_n+\Delta y/2} \int_{H_P}^{H_{nm}} \frac{(H_P - z)}{r^3\,(x_P - x,\ y_P - y,\ H_P - z)}\, dxdydz, \tag{8.83}$$

and in an equivalent way, the terrain correction formula takes the following form for its implementation in 2D by NIM (Li 1993):

$$c\,(x_P, y_P) = G\rho \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \int_{x_n-\Delta x/2}^{x_n+\Delta x/2} \int_{y_n-\Delta y/2}^{y_n+\Delta y/2} \left(\frac{1}{r(x_P - x,\ y_p - y,\ 0)} - \frac{1}{r(x_P - x,\ y_p - y,\ H_P - H_{nm})} \right) dxdydz \tag{8.84}$$

In prism representation and in the case of a 2D DTM as before (8.78) can be implemented by the following double summation formula (e.g., Li and Sideris 1994),

$$c(x_P, y_P) = G\rho \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \Big[x \ln(y + r) + y \ln(x + r) - z \arctan\left(\frac{xy}{zr}\right) \Big] \Big|_{x_n-\Delta x/2}^{x_n+\Delta x/2} \Big|_{y_m-\Delta y/2}^{y_m+\Delta y/2} \Big|_{H_{nm}}^{H_P}, \tag{8.85}$$

which can be efficiently evaluated by 2D FFT as it is shown in the next Sect. 8.5.2.

The prism topographic model realized by the aforementioned equations can be simplified, if the mass of the prism is mathematically concentrated along its vertical symmetric axis and the prism is represented as a line. Then, the expression of the terrain correction formula, instead of carrying out the double integration in (8.84) can be derived by the following expression (Li 1993; Li and Sideris 1994):

$$c(x_P, y_P) = G\rho\Delta x\Delta y \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \left[\left(\frac{1}{r(x_P - x_n,\ y_p - y_m,\ 0)} - \frac{1}{r(x_P - x_n,\ y_p - y_m,\ H_P - H_{nm})} \right) \right] \tag{8.86}$$

This last approximation is known in the literature as the *mass line representation* of topographic masses (e.g., Li 1993; Li and Sideris 1994).

In a similar way as before the isostatic effect can be computed for gravity data on the geoid. As an example, the isostatic effect on gravity data on the geoid according to the AH model (8.33) is given as

$$\begin{aligned} &A_{comp/AH}(x_p, y_p, z_p) \\ &= G\Delta\rho \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \Big\{ (x_P - x) \ln[(y_P - y) + r] + (y_P - y) \ln[(x_P - x) + r] \\ &\quad - (z_P - z) \arctan\left[\frac{(x_P - x)(y_P - y)}{(z_P - z)\, r} \right] \Big\} \Big|_{x_n - \Delta x/2}^{x_n + \Delta x/2} \Big|_{y_m - \Delta y/2}^{y_m + \Delta y/2} \Big|_{-T_0 - H}^{-T_0 - d - H}. \end{aligned} \tag{8.87}$$

Equation 8.87 can be used also for the computation of the isostatic effect on airborne gravity data by simply replacing z_P by z_0.

Another application of NIM is the computation of the topographic effect on airborne gradiometry data. This is based on the differentiation of the kernel function f_z (see details in Sect. 10.4.2)

$$f_z(x, y, z) = \iiint_{\Delta_{xyz}} \frac{z}{(x^2 + y^2 + z^2)^{3/2}}\, dx dy dz \tag{8.88}$$

with respect to z that results in the second-order gradient of the kernel function expressed as

$$\begin{aligned} f_{zz}(x, y, z) = &\left[\left(\frac{x}{y + r'} + \frac{y}{x + r'} \right) \frac{z}{r'} \right. \\ &\left. - \arctan\left(\frac{xy}{zr'} \right) + \frac{xyz(r'^2 + z^2)}{z^2 r'^3 + r' x^2 y^2} \right] \Big|_{x - \Delta x/2}^{x + \Delta x/2} \Big|_{y - \Delta y/2}^{y + \Delta y/2} \Big|_{z - \Delta z/2}^{z + \Delta z/2}. \end{aligned} \tag{8.89}$$

Δ_{xyz} in (8.88) is the volume of each grid element. The topographic effect on airborne gradiometry data is finally given as (Peng 1994)

$$\begin{aligned} T_{zz}(x_P, y_P, z_0) = G\rho \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} &\left\{ \left(\frac{x_P - x}{(y_P - y) + r} + \frac{x_P - x}{(y_P - y) + r} \right) \frac{z_0 - z}{r} \right. \\ &- \arctan \frac{(x_P - x)(y_P - y)}{(z_0 - z) + r} \\ &\left. + \frac{(x_P - x)(y_P - y)\,(z_0 - z)\,[r^2 + (z_0 - z)^2]}{(z_0 - z)^2\, r^3 + r(x_P - x)^2 (y_P - y)^2} \right\} \Big|_{x_n - \Delta x/2}^{x_n + \Delta x/2} \Big|_{y_m - \Delta y/2}^{y_m + \Delta y/2} \Big|_{0}^{H(x,y)}. \end{aligned} \tag{8.90}$$

A last example is the evaluation of the indirect effect of topographic/isostatic reductions on the geoid (see 8.60–8.62). The kernel function in this case for both topographic and isostatic reduction is expressed as

$$f_{in}(x,y,z) = \iiint_{\Delta_{xyz}} \frac{dxdydz}{(x^2+y^2+z^2)^{3/2}} \tag{8.91}$$

and can be re-written as follows (see also 8.81):

$$\begin{aligned} & f_{in}(x,y,z) \\ & = G\rho\left[xy\ln(z+r') + xz\ln(y+r') + yz\ln(x+r') - \frac{x^2}{2}\arctan\left(\frac{yz}{xr'}\right)\right. \\ & \left. - \frac{y^2}{2}\arctan\left(\frac{xz}{yr'}\right) - \frac{z^2}{2}\arctan\left(\frac{xy}{zr'}\right)\right]\Bigg|_{x-\Delta x/2}^{x+\Delta x/2}\Bigg|_{y-\Delta y/2}^{y+\Delta y/2}\Bigg|_{z-\Delta z/2}^{z+\Delta z/2}. \end{aligned} \tag{8.92}$$

The indirect effect due to the topographic reduction is expressed as (Peng 1994)

$$\begin{aligned} & \Delta N(x_P, y_P, z_P) \\ & = \frac{G\rho}{\bar{\gamma}}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\left\{(x_P-x)(y_P-y)\ln(z+r)\right. \\ & + (y_P-y)z\ln[(x_P-x)+r] + (x_P-x)z\ln[(y_P-y)+r] \\ & - \frac{1}{2}(x_P-x)^2\arctan\frac{(y_P-y)z}{(x_P-x)r} - \frac{1}{2}(y_P-y)^2\arctan\frac{(x_P-x)z}{(y_P-y)r} \\ & \left. - \frac{1}{2}z^2\arctan\frac{(x_P-x)(y_P-y)}{zr}\right\}\Bigg|_{x_n-\Delta x/2}^{x_n+\Delta x/2}\Bigg|_{y_m-\Delta y/2}^{y_m+\Delta y/2}\Bigg|_0^H \end{aligned} \tag{8.93}$$

and a similar formula gives the indirect effect of the isostatic reduction, where in this case ρ is replaced by $\Delta\rho$ and the integration interval with respect to z is now $(-T_0 - H,\ -T_0 - d - H)$.

It should be stressed once again that the numerical implementation of prism formulas is a time-consuming procedure and it needs advanced computer resources, especially when dense DTMs/DBMs are employed. This drawback can be considerably reduced by using approximated formulas at larger distances without decreasing the accuracy of the computations and obtaining reasonable computational speed (Forsberg 1984, 2010). In practical applications, this methodology is based on the use of a detailed grid around the computation point and a coarser grid for the remaining test area, as it was already mentioned in Sect. 8.4.4.

8.5.2 The Fast Fourier Transform (FFT) Method

The different mass reduction methods discussed in the previous sections are mainly formulated through convolution integrals, either directly or after an expansion into a series in case of non-linear integrals (e.g., Tziavos et al. 1988; Schwarz et al. 1990). In the latter case, two or three terms, are typically sufficient to meet, in terms of accuracy, the recent demands of the majority of applications related to gravity field modeling. Since the nowadays digital models of topography and bathymetry are available in regular grids and the convolutions in practice are carried out with gridded data, the finally derived mass reduction convolution integrals can be efficiently evaluated by means of FFT. In the sequel, a few theoretical examples will be illustrated for different reduction schemes, based on equations already presented in the previous sections of this chapter and making appropriate reference to equations given also in Chap. 10 (Part II of this book) either in space or frequency domain.

8.5.2.1 Terrain Corrections

The rigorous terrain correction formula (8.5) can be evaluated by 3D FFT as it is shown in Sect. 10.4.2 in Chap. 10 (Part II of this book) and it is not repeated here. In practical applications the *linear approximation* of the terrain correction of the terrain correction is often used (8.11). This convolution integral can be evaluated by means of FFT provided the resolution of the available DTM and DDM is the same and the following formula can be finally derived

$$c(x_P, y_P) = \frac{1}{2} G \left[\mathbf{F^{-1}}\{PH_2(u,v) L_c(u,v)\} - 2H(x_P, y_P)\mathbf{F^{-1}}\{PH(u,v) L_c(u,v)\} \right.$$
$$\left. + H^2(x_P, y_P)\mathbf{F^{-1}}\{P(u,v) L_c(u,v)\} \right], \tag{8.94}$$

where $P(u,v) = \mathbf{F}\{\rho(u,v)\}, PH_2(u,v) = \mathbf{F}\{\rho(x,y) H^2(x,y)\}, PH(u,v) = \mathbf{F}\{\rho(x,y) H(x,y)\}$, $L_c(u,v)$ is the spectrum of the kernel function (planar distance), $\mathbf{F}$ and $\mathbf{F^{-1}}$ denote the direct and inverse 2D Fourier transform, respectively, and u, v are the frequencies corresponding to x, y, respectively. It is to be noticed that the terms spectrum and Fourier transform are used synonymously within this section. In the frequency domain formulas given below the argument (u, v) of the different spectra is omitted for simplicity reasons. When horizontally varying density values are available through a 2D DDM and comparing with the case of constant density value (10.53a in Chap. 10), it is evident from (8.94) that some additional computational effort is needed for the computation of the spectra of the density values and their products with heights.

A refined 2D FFT-based expression for c has been investigated by Li (1993) and numerically tested in several studies (e.g., Li and Sideris 1994; Tziavos et al. 1996; Bajracharya 2003). In (8.85), the expansion of terms containing $(z = H_P - H_{nm})$ into a series results in the terrain-correction formula for the mass prism topographic model (see Sect. 8.5.1), that can be expressed as follows (Li and Sideris 1994):

$$c(x_p, y_P) = c_0(x_p, y_P) + c_1(x_p, y_P) + c_2(x_p, y_P) + c_3(x_p, y_P) + \ldots.. \quad (8.95)$$

The term $c_0(x_p, y_P)$ can be evaluated directly by (8.86) and the other terms efficiently by FFT. As an example the $c_1(x_p, y_P)$ term is evaluated by means of the Fourier transforms as

$$c_1(x_p, y_P) = \frac{G\rho}{2}\left[(h_P^2 - \alpha^2)\mathbf{F}^{-1}\{H_0 F_1\} - 2h(x_P, y_P)\mathbf{F}^{-1}\{H_1 F_1\} + \mathbf{F}^{-1}\{H_2 F_1\}\right], \quad (8.96)$$

where

$$H_k = \mathbf{F}\{H^k\}, k = 0, 1, 2, 3, \ldots.., \quad (8.97)$$

$$F_1 = \mathbf{F}\left\{f_{11}(x, y, \alpha) + f_{11}(y, x, \alpha) - f_{12}(x, y, \alpha)\right\}, \quad (8.98)$$

$$f_{11}(x, y, \alpha) = \frac{-x}{[y + r(x, y, \alpha)]\, r(x, y, \alpha)} \Bigg|_{x-\Delta x/2}^{x+\Delta x/2} \Bigg|_{y-\Delta y/2}^{y+\Delta y/2}, \quad (8.99)$$

$$f_{12}(x, y, \alpha) = \frac{xy\left(r^2 + \alpha^2\right)}{(x^2 y^2 + \alpha^2 r^2) r} - \frac{1}{\alpha} \arctan\left(\frac{xy}{\alpha r}\right) \Bigg|_{x-\Delta x/2}^{x+\Delta x/2} \Bigg|_{y-\Delta y/2}^{y+\Delta y/2}. \quad (8.100)$$

In the above equations the role of parameter α is to speed up the convergence of the series in (8.95) and its optimal value in the mass prism topographic model is the *std* of the heights, i.e., $\alpha = \sigma_h$ (e.g., Li 1993). In the case that horizontally varying density values are available through a 2D DDM, (8.96) can be transformed as follows:

$$c_1(x_p, y_P) = \frac{G}{2}\left[(h_P^2 - \alpha^2)\mathbf{F}^{-1}\{P\, F_1\} - 2H(x_P, y_P)\mathbf{F}^{-1}\{PH_1\, F_1\} + \mathbf{F}^{-1}\{PH_2\, F_1\}\right], \quad (8.101)$$

where *P, PH, PH*$_2$ are the spectra of the same quantities as in (8.94) before. Analytical derivations for the mass prism and mass line topographic models along with expressions for the higher order terms in (8.95) can be found in (Li and Sideris 1994; Tziavos et al. 1996).

8.5.2.2 Effect of Bathymetry

In Sect. 8.3.2 the effect of bathymetric masses on gravity data located on the geoid was realized by (8.12), which is written in convolution form as follows:

$$c_b(x_p, y_P, 0) = \frac{1}{2}G\left[\left((\Delta\rho H^2) * \frac{1}{l_0^3}\right)\right] - \frac{3}{8}G\left[\left((\Delta\rho H^4) * \frac{1}{l_0^5}\right)\right] + \ldots.. \tag{8.102}$$

It is noticing once again here that H represent depths and consequently they provide the relief of the bottom of the sea. This last equation, when employing FFT, becomes

$$c_b(x_p, y_P, 0) = \frac{1}{2}G\left[\mathbf{F}^{-1}\{\Delta PH_2\, L_3\}\right] - \frac{3}{8}G\left[\mathbf{F}^{-1}\{\Delta PH_4\, L_5\}\right] + \ldots., \tag{8.103}$$

where $\Delta PH_2 = \mathbf{F}\{\Delta\rho H^2\}, \Delta PH_4 = \mathbf{F}\{\Delta\rho H^4\}$ and L_3, L_5 denote the spectra of the different powers of the kernel function (planar distance). As it has been mentioned already in Sect. 8.3.2, (8.103) and the corresponding equation for the classical terrain correction (8.94) do not converge satisfactorily and numerical instabilities occur in the computations. These problems are more evident in cases that rough and high resolution digital terrain or bathymetry models are used and can be overcome or at least significantly reduced by different ways, as, e.g., by: (a) Introducing an appropriate parameter α equal to the *std* of the heights/depths in the FFT-based formulas, as was indicated above in (8.96–8.101). (b) Using directly the time-consuming NIM for the entire test area. (c) Using the NIM in an area around the computation point and the FFT technique for the remaining area. All methods present advantages and drawbacks and the final choice is directly connected with the required accuracy, the topographic/bathymetric features of the test area and the available computer resources.

8.5.2.3 Topographic/Isostatic Effects on Gravity and Geoid

As it was mentioned in Sect. 8.3.2, the effects of topographic and compensated masses in the PH and AH isostatic models are given by (8.25–8.26) and (8.32–8.33), respectively. These equations can be efficiently evaluated by 3D FFT as it is shown also in Sect. 10.4.2 in Part II of this book. In the AH model for example, the effect of the isostatic masses on the gravity vector is expressed by 3D FFT as

$$A_{comp/AH}(x_p, y_P, H_P) = G\mathbf{F}^{-1}\{\Delta P\, F_z\}, \tag{8.104}$$

where ΔP and F_z are the spectra of the density contrast values $\Delta\rho$ and the kernel function f_z (8.91). Even though the 3D FFT method is used for the computation of topographic/isostatic effects, this is a time-consuming procedure. For this reason simpler formulas have been derived, as it is shown in Sect. 8.3.3, which can be efficiently evaluated by 2D FFT. For the AH model for example the isostatic effect on gravity can be computed by (8.37), which is written in convolution form as

$$A_{comp/AH}(x_P, y_P, H_P) = G\left[\left[(\Delta\rho d) * \left(\frac{d_0}{l_0^3} + \frac{l_0^2 - d_0^2}{l_0^5}(T_0 - d_0)\right)\right] + \frac{1}{2}\left((\Delta\rho d^2) * \frac{l_0^2 - 3d_0^2}{l_0^5}\right)\right] \tag{8.105}$$

Using the substitutions

$$f_1 = \frac{d_0}{l_0^3} + \frac{l_0^2 - 3d_0^2}{l_0^5}(T_0 - d_0) \quad f_2 = \frac{l_0^2 - 3d_0^2}{l_0^5}, \tag{8.106}$$

(8.105) becomes in the frequency domain:

$$A_{comp/AH}(x_P, y_P, H_P) = G\left[\mathbf{F^{-1}}\{\Delta PD\ F_1\} + \frac{1}{2}\mathbf{F^{-1}}\{\Delta PD_2\ F_2\}\right]. \tag{8.107}$$

8.5.2.4 Topographic/Isostatic Effects on Airborne Gravity and Gradiometry

The topographic effect on airborne gravity data (8.40) can be transformed in convolution form as

$$T_z(x, y, z_0) = G\left[(\rho H) * f_1\right] - \frac{G}{2}\left[(\rho H^2) * f_2\right], \tag{8.108}$$

where in this case

$$f_1(x, y) = \frac{z_0}{l_0^3}, \quad f_2(x, y) = \frac{1}{l_0^3} - \frac{3z_0}{l_0^5}, \quad l_0 = \sqrt{(x_P - x)^2 + (y_P - y)^2 + z_0^2}. \tag{8.109}$$

Equation 8.108 can be evaluated by 2D FFT method and becomes in the frequency domain:

$$T_z(x, y, z_0) = G\left[\mathbf{F^{-1}}\{PH\ F_1\} - \frac{1}{2}\mathbf{F^{-1}}\{PH_2\ F_2\}\right]. \tag{8.110}$$

The topographic effect on airborne gradiometry (gradient) data (8.41) is expressed in convolution form as

$$T_{zz}(x, y, z_0) = G\left[(\rho H) * f_1\right] + \frac{G}{2}\left[(\rho H^2) * f_2\right], \tag{8.111}$$

where in this case

$$f_1(x, y) = \frac{l_0^3 - 3z_0^2}{l_0^5}, \quad f_2(x, y) = \frac{9z_0}{l_0^5} - \frac{15z_0^3}{l_0^7}, \quad l_0 = \sqrt{(x_P - x)^2 + (y_P - y)^2 + z_0^2}. \tag{8.112}$$

Equation 8.111 can be evaluated by 2D FFT and takes the following form:

$$T_{zz}(x, y, z_0) = G\left[\mathbf{F^{-1}}\{PH\ F_1\} + \frac{1}{2}\mathbf{F^{-1}}\{PH_2\ F_2\}\right]. \tag{8.113}$$

It should be mentioned also that the effects of topographic masses on airborne gravity and gradiometry data can be evaluated by the rigorous 3D FFT method as in (8.104) before (see also details in Chap. 10, Sect. 10.4.2, Part II of this book), but this is a time-consuming method that needs extended computer resources.

In a similar way as above, the isostatic effects on airborne gravity and gradiometry can be evaluated by 2D FFT, as it was mentioned already in Sect. 8.3.3. Thus, the isostatic effects on airborne gravity by the AH model can be evaluated by (8.108)–(8.110) by replacing ρ by $\Delta\rho$, z_0 by $(z_0 + T_0)$ and H by d (see Figs. 8.5 and 8.10). In an analogous way, the isostatic effects for airborne gradiometry by the AH model are computed using 2D FFT through (8.111)–(8.113) using the same replacements as before.

8.5.2.5 Indirect Effects of Topographic/Isostatic Reductions on the Geoid

A last example is the representation of the indirect effect of the topographic/isostatic reductions (AH model) on the geoid implemented by 3D FFT. It is based on (8.93) used for evaluation by NIM. The total indirect effect on the geoid is given as follows:

$$\Delta N(x, y, 0) = \Delta N_{top/AH}(x, y, 0) + \Delta N_{comp/AH}(x, y, 0). \tag{8.114}$$

where the first term covers all topographic masses and the second one the space from the compensated masses to the geoid. The two constituents of the indirect effect are written generally in convolution form as

$$\Delta N_{top/AH}(x, y, 0) = \frac{G}{\gamma}\left[\rho(x, y, z) * f_{in}(x, y, z)\right], \tag{8.115}$$

$$\Delta N_{comp/AH}(x, y, 0) = \frac{G}{\gamma}\left[\Delta\rho(x, y, z) * f_{in}(x, y, z)\right], \tag{8.116}$$

where the kernel function f_{in} in both equations has the same form given by (8.91). When (8.115)and (8.116) are implemented by 3D FFT, the total indirect effect on the geoid is expressed in the frequency domain as

$$\Delta N(x, y, 0) = \frac{G}{\gamma}\mathbf{F^{-1}}\{P\ F_{in} + \Delta P\ F_{in}\}. \tag{8.117}$$

where P, ΔP and F_{in} are the spectra of $\rho, \Delta\rho$ and the kernel (distance) function (8.91), respectively. Approximated formulas for the indirect effect by 2D convolution integrals implemented by 2D FFT are presented in Chap. 10 (Sect. 10.4.1) in Part II of the book.

8.6 Numerical Examples

The numerical tests given below are based on the various mass reduction methods outlined in the previous sections. The Shuttle Radar Topographic Mission 3″ (SRTM3″) DTM (e.g., Farr et al. 2007) formed the original height data for the test area located in the central and north part of Greece covering also a region of the neighbouring countries to the north and bounded between $38° \leq \phi \leq 42.5°$ and $20° \leq \lambda \leq 23.5°$. Given that the SRTM mission contains voids and gaps the original 3″ DTM was corrected over Greece through comparisons with a national DTM, so that a complete SRTM3″ DTM was first constructed (see, Tziavos et al. 2010). In order to investigate the dependence of the estimated topographic effects by the resolution of the DTM, from this fine-resolution DTM two coarser models of 15″ and 1′ have been determined and finally used in the computations of the topographic effects in an inner area bounded between $39° \leq \phi \leq 41.5°$ and $21° \leq \lambda \leq 22.5°$ (Fig. 8.11). It should be noted that the available height data cover a wider area than that of the computations (one degree wider in all directions) in order to eliminate aliasing effects. For the generation of the coarser SRTM DTMs simple averages have been taken from the heights of the fine 3″ SRTM model, while no additional filtering or smoothing has been applied. The area under study is composed by land and sea. The land part is mostly mountainous with very few lowlands and a smooth transition from high to low elevations, while in the marine part the variation of depths is quite significant (Fig. 8.11). However, bathymetry effects have not taken into account since the different mass reductions have been computed for an inner zone as it has been mentioned before. Table 8.1 summarizes the statistics of the two finally constructed DTMs for the continental part of the area under study and used in the numerical tests.

In continuation of the theoretical examples given in Sect. 8.4.5 two numerical tests are presented in the following sections. In the first one the results of different terrain reduction schemes on gravity and geoid are computed using the above mentioned SRTM-based DTMs for the area under study. In the second test gravimetric geoid solutions are presented based on the remove-restore method, employing land and marine gravity data, GOCO02s[1] (Goiginger et al. 2011) and EGM2008 (Pavlis et al. 2008) as reference geopotential models and the 15″ DTM for the computation of the different topographic effects. Finally, an evaluation of the different computed geoid models is carried out over a network of GPS/leveling benchmarks.

[1] *The GOCO02s global gravity field model is based on SLR, CHAMP, GRACE and GOCE data and its expansion is complete to degree and order 250.*

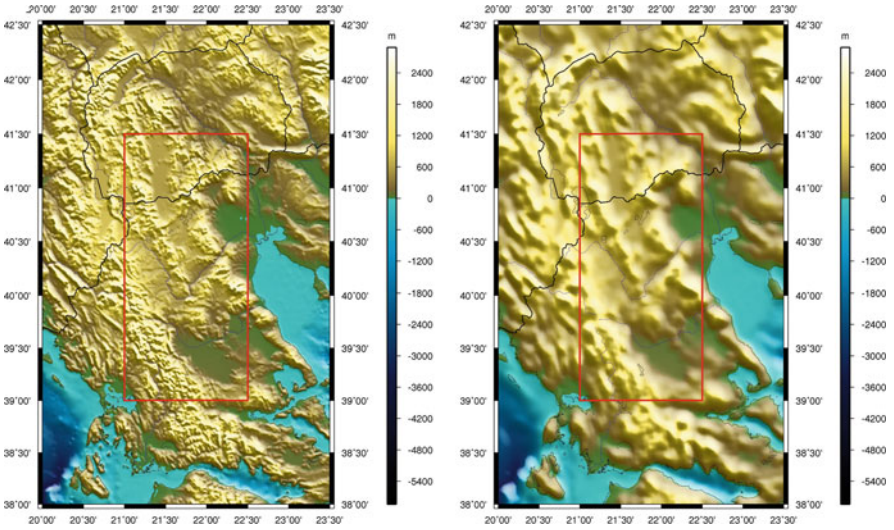


Fig. 8.11 The corrected (**a**) SRTM 15″ DTM (*left*), (**b**) SRTM 1′ DTM (*right*) in the area under study and the inner working area denoted by the *red* frame

Table 8.1 Statistics of the DTMs and their differences in the area under study (Unit: m)

DTMs	Max	Min	Mean	std
SRTM 15″	2657.2	0.0	859.9	±691.9
SRTM 1′	2442.5	0.0	840.5	±637.4

8.6.1 Effects of Terrain Reductions on Gravity Anomalies and Geoid Heights

The topographic effects on gravity and the geoid computed were (a) full topographic effects, i.e., the combined effect of the Bouguer and terrain corrections used to construct refined Bouguer anomalies, (b) isostatic effects using the AH model, (c) terrain correction (TC) effects and (d) RTM effects. Furthermore, indirect effects on the geoid have been computed estimating all three terms of (10.58) in Chap. 10 of Part II of the book. For the results computed to be representative for the two SRTM models, the effects were estimated and then compared on a $1' \times 1'$ regular grid, which corresponds to cases that a geoid and/or gravity field model of that resolution is needed. Such a high resolution $1' \times 1'$ geoid model is clearly within reach today in the presence of new gravity-field related data and ultra-high resolution GGMs like EGM2008. The latter GGM has a maximum degree and order of expansion equal to 2190, which corresponds to a spatial resolution of ∼10 km half wavelength (∼5′); therefore with the aid of local gravity, altimetry and topography data a geoid model of 1′ is feasible.

Table 8.2 Full topographic, AH isostatic, TC and RTM effects on gravity anomalies in the area under study (Unit: mGal)

DTMs	Max	Min	Mean	std
Full topographic (Bouguer + TC)				
SRTM 15″	211.94	−1.52	45.82	±45.18
SRTM 1′	192.27	−1.96	42.39	±41.48
AH isostatic				
SRTM 15″	149.09	−42.44	0.54	±34.82
SRTM 1′	133.71	−39.37	3.86	±32.88
TC				
SRTM 15″	19.55	0.12	2.02	±2.15
SRTM 1′	16.08	0.07	1.22	±1.27
RTM				
SRTM 15″	140.12	−95.66	−20.39	±34.88
SRTM 1′	127.01	−92.80	−17.17	±31.41

One further characteristic of the topographic reductions that is sought for is their correlation with height. For all topographic reductions computed for the test area their correlation coefficient with height has been determined. Appropriate formulas for the correlation coefficient computation can be found, e.g., in Bendat and Piersol (2000).

Table 8.2 presents the statistics of the estimated full topographic effects, terrain corrections, RTM and isostatic effects (AH model) on gravity from the two available SRTM models. From this Table, the magnitude of the computed effects on gravity can be seen, indicating that the full topographic (Bouguer and TC) effect has a large range as does the AH isostatic one, with a variation of 214 mGal and 192 mGal, respectively. The TC effect, which is always positive as shown in Table 8.2, has a smaller variation of 19 mGal only, while the RTM effect has the largest variation of 236 mGal. The difference in the RMT effects compared to the full-topographic and AH ones can be viewed in the mean values, which are at the −20 mGal, 46 mGal and 0.5 mGal level, respectively.

As for the evaluation of the topographic effects on gravity anomalies for the lower resolution SRTM model, it can be seen that its differences to the 15″ are generally noticeable. The std of the differences between the 15″ and the 1′ models is at the ±1.0 mGal level for the TC, ±3.5 mGal for the RTM, and ±2.0 mGal for the AH isostatic and ±4.0 mGal for the full topographic effects. The variation of the range of the differences is more significant since it is as at the 3 mGal for the TC, 16 mGal for the RTM, and 19 mGal for the AH isostatic and full topographic effects. Figures 8.12 and 8.13 present the full-topographic, AH isostatic, TC and RTM effects on gravity anomalies for the 15″ DTM used.

Figure 8.14 presents the correlation with height of the full-topographic, TC, RTM and AH reductions in the area under study, where the values of each reduction have been plotted against heights. From Fig. 8.14 it is clear that the refined Bouguer reduction has the highest correlation with height reaching 99.9% which is almost

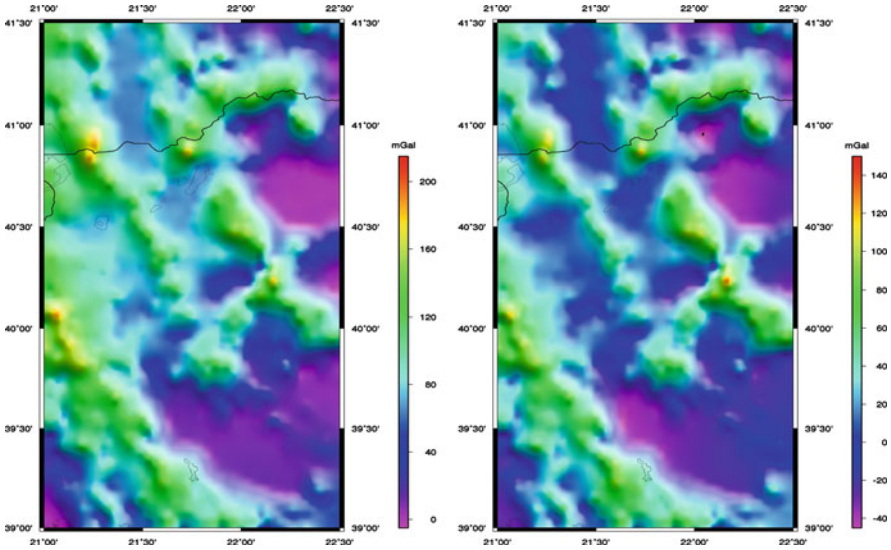


Fig. 8.12 Full topographic (*left*) and AH isostatic (*right*) effects on gravity anomalies in the inner working area based on the SRTM 15″ model (Unit: mGal)

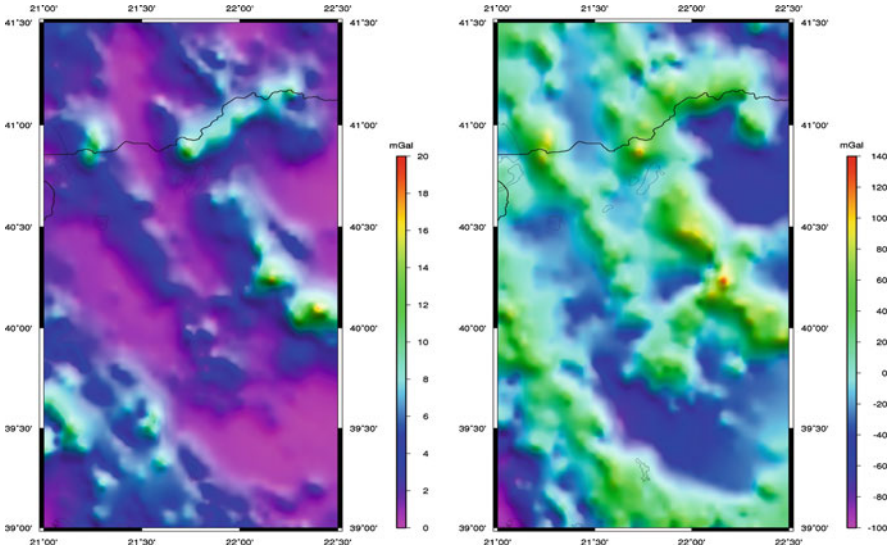


Fig. 8.13 TC (*left*) and RTM (*right*) effects on gravity anomalies in the inner working area based on the SRTM 15″ model (Unit: mGal)

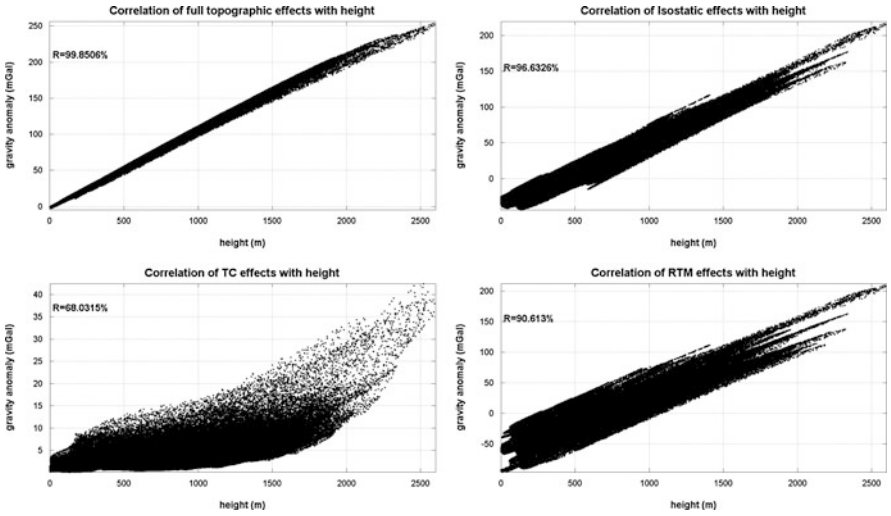


Fig. 8.14 Correlation of the various reduction schemes with height based on the SRTM 15″ model

Table 8.3 Full topographic, AH isostatic, TC and RTM effects on geoid heights in the test area (Unit: m)

DTMs	Max	Min	Mean	std
Full topographic (Bouguer + TC)				
SRTM 15″	12.876	3.516	8.938	±2.252
SRTM 1′	12.811	3.382	8.930	±2.230
AH isostatic				
SRTM 15″	4.035	0.605	2.427	±0.833
SRTM 1′	3.931	0.578	1.822	±0.774
TC				
SRTM 15″	0.530	0.188	0.328	±0.057
SRTM 1′	0.489	0.210	0.288	±0.041
RTM				
SRTM 15″	1.613	−0.748	−0.316	±1.105
SRTM 1′	1.560	−0.681	0.300	±1.081

complete correlation. The second best result is achieved for the AH model where a 96.6% correlation is achieved. These high correlations with height denote one of the main reasons for their use in geophysics, since they manage to produce almost uncorrelated with height reduced gravity anomalies. The TC reduction has a correlation with height at the 68% level, while the RTM reaches 91% correlation with the topography.

Table 8.3 presents the statistics of the estimated full topographic effects, terrain corrections, RTM and isostatic effects on geoid heights from the two available SRTM models. As expected, the magnitude of the effects on geoid heights is significant, with the full topographic effect having a range at the level of 9 m, the AH

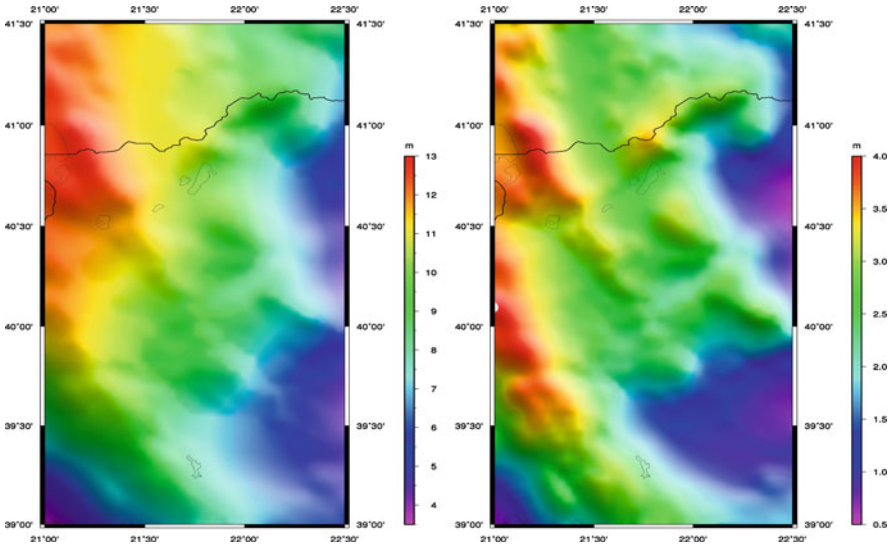


Fig. 8.15 Full topographic (*left*) and AH isostatic (*right*) effects on geoid heights in the inner working area based on the SRTM 15″ model (Unit: m)

Table 8.4 Indirect effects on geoid heights (Helemert's second method of condensation) (Unit: m)

DTMs	Max	Min	Mean	std
Indirect effects on geoid heights				
SRTM 15″	0.000	−0.315	−0.052	±0.049
SRTM 1′	0.000	−0.221	−0.050	±0.041

isostatic 3 m, the TC effect having a smaller variation of 0.3 m only and the RTM effect reaching the 2 m. As for the evaluation of the topographic effects on geoid heights for the two SRTM models, it can be seen that the results generally decrease in terms of the different statistical parameters with an increase in grid spacing. The std of the differences between the 15″ and the 1′ models varies between ±0.02 m (full topographic, TC, RTM) to ±0.06 m (AH), while the range of the differences is more significant since it is as at the 0.08 m for the TC, 0.12 m for the RTM, 0.07 m for the AH isostatic and 0.06 m for the full-topographic effects. Figures 8.15 and 8.16 present the full-topographic, AH isostatic, TC and RTM effects on geoid heights for the 15″ DTM used.

The final topographic effect computed was the indirect effect on geoid heights as described by (10.58) in Chap. 10 (Part II of this book). Table 8.4 summarizes the statistics of the indirect effect for both DTMs. It is important to stress once again that the indirect effect on the geoid corresponding to Helmert's second method of condensation has a small magnitude (std of ±0.049 m). The same holds for the RTM effects on the geoid presented in Table 8.3 before, which have a larger

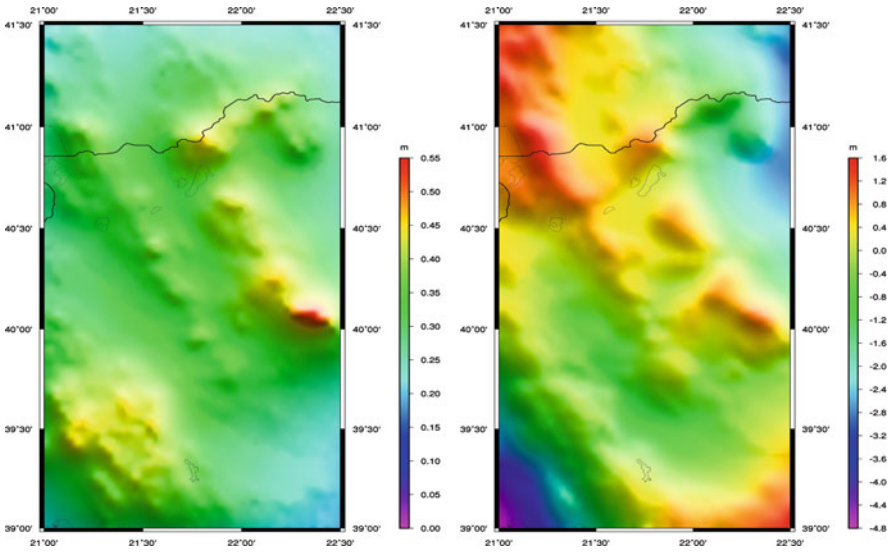


Fig. 8.16 TC (*left*) and RTM effects (*right*) on geoid heights in the inner working area based on the SRTM 15″ model (Unit: m)

contribution to geoid heights (± 0.105 m), but, generally, these effects are very small compared to the several meters of indirect effect resulting from the full-topographic and AH models. These characteristics of Helmert's second method of condensation and RTM reduction schemes along with those outlined in Sect. 8.4, make them the dominant schemes to compute the topographic attraction in different geodetic research applications. The differences between the estimated indirect effects from the two SRTM models reach the ± 0.008 m level in terms of the std and the 0.094 m in terms of the range, signaling that even a small deterioration in the resolution of the available DTM can have a significant impact on the estimated indirect effect to the geoid.

8.6.2 Determination and Evaluation of Gravimetric Geoid Models

A second numerical test has been carried out in the same test area as before and its objectives are summarized as follows. The first goal is to investigate the effect of the terrain on a gravimetric geoid computation in conjunction with surface gravity data and a GGM derived only from satellite data (GOCO02s), taking into account the topography through the four mass reduction schemes employed in the first

numerical test presented in Sect. 8.6.1. The second purpose is to identify the most appropriate terrain reduction method to be used in the next computational step. The criteria used for this procedure are the smoothness of the produced gravity field and the assessment results of the four derived geoid models over a network of GPS/leveling benchmarks available in the test area. The final outcome of this numerical example is the optimal combination of gravity and terrain data with reference to EGM2008 for the determination of a high-accuracy and resolution geoid model in the area under study, investigating the impact of the topographic information and the local gravity data as well in such a combination procedure.

8.6.2.1 Data Sources and Reductions

A number of 2,053 point free-air gravity anomalies (Δg_{FA}) are irregularly distributed in the test area, belonging to a gravity database for the Greek territory and are referred to IGSN71/GRS80. There are also available 64 GPS/leveling stations in the area under study measured on triangulation pillars of the national horizontal geodetic network. The locations of the gravity data are pictured in Fig. 8.17 along with the network of the GPS/leveling benchmarks. The computation of the topographic effects was performed through the four reduction schemes (full topographic, AH isostatic, TC, RTM), used already in the first numerical experiment (Sect. 8.6.1), while the SRTM-based 15″ DTM was used to provide the height information for these calculations (see Fig. 8.11). The contribution of the GOCO02s GGM (Δg_{EGM}) as well as the contribution of the terrain effects (Δg_H) were subtracted from the free-air gravity anomalies in the first step of this numerical test and the reduced (Δg_{red}) and residual (Δg_{res}) gravity anomalies were derived, respectively. These reductions followed the remove-restore technique according to the formulas

$$\Delta g_{red} = \Delta g_{FA} - \Delta g_{EGM}, \Delta g_{res} = \Delta g_{red} - \Delta g_H. \tag{8.118}$$

The contribution of the geopotential model was restored at the output in the computed geoid heights along with the terrain contribution. The statistical results of all the above mentioned reductions are summarized in Table 8.5. From the results acquired it is evident that the RTM reduction in conjunction with the GGM produce a smooth residual field with respect to the range of the gravity anomalies and std and outperforms all other schemes in terms of the mean value. This means that both high and low frequencies are sufficiently blocked in this combination approach and a residual gravity data field close to a normal distribution is obtained. The AH isostatic and full topographic reductions, although they present similar range with that of RTM, have a considerably large mean value.

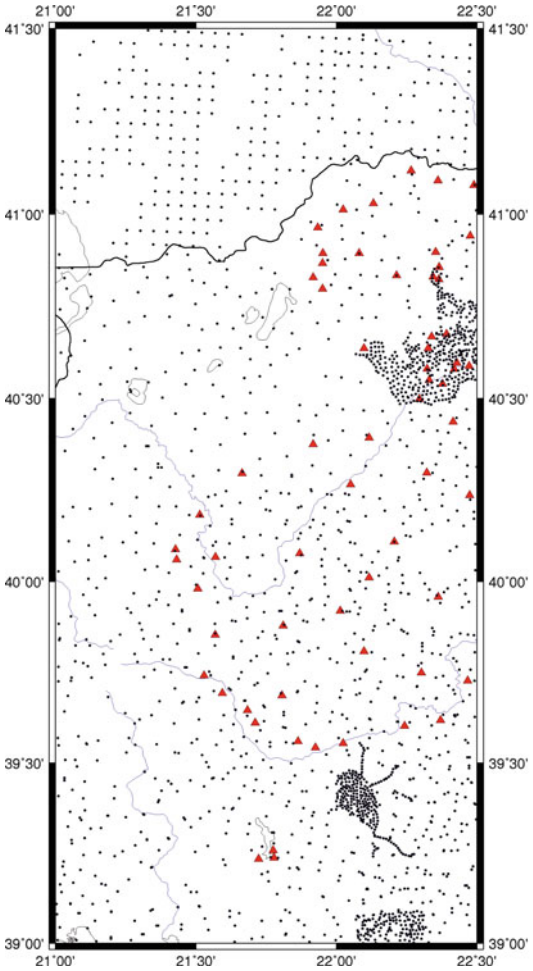


Fig. 8.17 Free-air gravity anomaly data (denoted by *dots*) and GPS/leveling stations (denoted by *red triangles*)

Table 8.5 Free-air, reduced to GOCO02s and residual gravity anomalies in the test area (Unit: mGal)

Gravity anomalies	Max	Min	Mean	std
Δg_{FAA}	192.13	109.24	11.30	±40.62
Δg_{red} GOCO02s	150.88	−85.30	−22.39	±30.74
Δg_{res} full topographic	−12.13	−156.68	−68.21	±27.60
Δg_{res} AH isostatic	24.62	−95.83	−22.94	±17.63
Δg_{res} TC	133.19	−93.35	−24.41	±29.71
Δg_{res} RTM	67.55	−58.83	−2.01	±20.18

8.6.2.2 Gravity Data Gridding

To construct the final residual gravity fields and in order the gridding procedure to be rigorous the LSC algorithm was chosen. This method is obviously more time consuming compared to other conventional techniques (e.g., spline interpolation, weighted means), but provides statistically optimal results. To grid the data using LSC the variance and the correlation length of the residual field had to be computed, thus the empirical covariance function of the data has been computed and fitted to the Tscherning and Rapp model (see Chap. 7 of Part II). In this way the final residual free-air gravity anomaly grids with a 2′ resolution (corresponding to about 3.4 km spatial resolution) have been estimated. Figure 8.18 presents the empirical covariance functions of the gravity data reduced to GOCO02s (Δg_{red}) along with the corresponding covariance functions of the four residual gravity fields (Δg_{res}) after the additional removal of the different topographic effects. In Table 8.6 the statistical results of the variance and the correlation length for the above mentioned gravity anomaly fields are tabulated. The aforementioned results strengthen the previous conclusion that the data after the RTM and AH isostatic reductions are indeed smooth, since the variance of the data reduced and the correlation length of the field increased. The RTM residual field is even smoother taking into account the mean value of the residual gravity anomalies (Table 8.5).

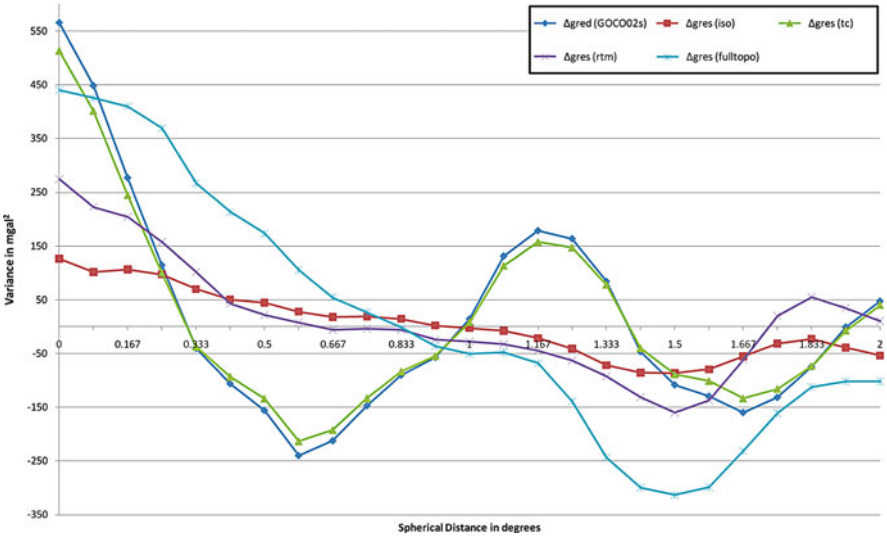


Fig. 8.18 Reduced empirical covariance functions for the different reduction schemes

Table 8.6 Variance and correlation length of the empirical covariance functions

	Δg_{red} (GOCO02s)	Δg_{res} (Full topo)	Δg_{res} (AH iso)	Δg_{res} (TC)	Δg_{res} (RTM)
Variance [mGal^2]	565.7	440.14	126.21	513.29	274.64
Corr. Length [deg]	0.1631	0.4059	0.3628	0.1596	0.2804

8.6.2.3 Geoid Computation and Validation

From the 2′ gridded residual gravity data four geoid models of the same resolution have been computed using the efficient 1D FFT method based on the spherical Stokes convolution (Haagmans et al. 1993). Then, adding back the effect of the topography through the reduction schemes employed and that of the geopotential model (GOCO02s) resulted in the final (2′ × 2′) geoid models. This restore step of the remove-restore method followed for the geoid computation is represented by the formula

$$N_{grav} = N_{res} + N_{EGM} + N_H. \tag{8.119}$$

The N_H term in the last equation can be interpreted as the restored topographic effect on the geoid and it represents the indirect effect of the corresponding mass reduction method used in the computations. In Fig. 8.19 a representation of the 2′ gravimetric geoid solution is shown with reference to GOCO02 and in conjunction with a RTM terrain effect estimation.

The evaluation of the estimated geoid models was performed through comparisons over the network of the 64 GPS/leveling benchmarks available in the area under study (see Fig. 8.16). Table 8.7 presents the statistics of the four estimated geoid models along with the differences between GPS/leveling and geoid heights at the 64 control points of the test area. From the results presented in Table 8.7 it is clear that the local gravimetric solution referenced to GOCO02s in conjunction with the construction of a RTM for the terrain effect estimation outperforms all other solutions, since the std of the differences that it provides is approximately 10–30 cm better than that of the other AH isostatic and Full-topographic geoid models, while it is also better by ∼1 cm compared to the TC-based geoid. An additional note refers to the significant biases detected between the GPS/leveling heights and the gravimetric ones in all geoid models. This can be mainly attributed to un-modeled datum shifts existing in the national datum, the datum shifts between the GPS/leveling vertical datum and that of the geoid models, while a small part of these bias values can be regarded as random errors of the vertical datum in the area under study.

The final numerical test performed, incorporated the EGM2008 global gravity field model in a new 2′ geoid determination for the test area in combination with the same gravity and height data as in the previous geoid computations and employing again the remove-restore technique and the 1D FFT for the residual geoid part. The EGM2008 was used in its full expansion, i.e., complete to degree and order 2,159. The objectives of this numerical example were the following: (a) To investigate

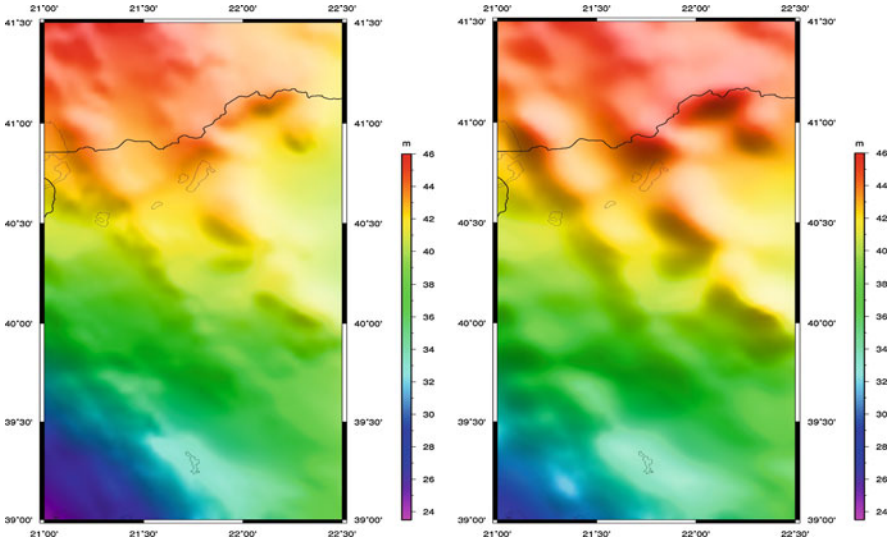


Fig. 8.19 Local gravimetric geoid model referenced to GOCO02s (*left*) and EGM2008 (*right*) using RTM for the effect of the topographic masses (Unit: m)

Table 8.7 Statistics of (**a**) the local gravimetric geoid models and (**b**) the differences between GPS/levelling and gravimetric geoid heights from the local models at 64 GPS/levelling benchmarks (Unit: m)

	Max	Min	Mean	std
N_{grav} (Full topographic)	51.638	27.364	41.103	±5.269
$N_{GPS} - N_{grav}$	0.840	-2.075	−0.630	±0.693
N_{grav} (AH isostatic)	49.108	29.636	42.833	±4.432
$N_{GPS} - N_{grav}$	−1.918	−4.038	−2.913	±0.511
N_{grav} (TC)	45.064	27.829	38.984	±4.415
$N_{GPS} - N_{grav}$	1.077	−1.030	0.317	±0.403
N_{grav} (RTM)	45.957	23.992	39.243	±4.613
$N_{GPS} - N_{grav}$	0.976	−0.583	0.231	±0.392

and assess the improvement that EGM2008 brings compared to the satellite only GOCO02s GGM used in the four geoid models before. (b) To determine the performance of EGM2008 with respect to a local geoid model. (c) To estimate the effect of topographic masses to geoid heights through the RTM reduction method even an ultra-high degree and order GGM as EGM2008 is used in geoid modeling.

The statistical results of the reduced to EGM2008 gravity anomaly field and the residual field after the removal of RTM-effects from the reduced values are summarized in Table 8.8. For comparison purposes the statistics of the original free-air anomaly field is repeated in this Table. From the results of Table 8.8 it is evident that the reduced to EGM2008 free-air gravity anomaly field is considerably smoother since the range is reduced by 47% (140.2 mGal), the mean value by 62%

Table 8.8 Free-air gravity anomalies, reduced to EGM2008 and residual gravity anomalies in the test area (Unit: mGal)

Gravity anomalies	Max	Min	Mean	std
Δg_{FAA}	192.13	−109.24	11.30	±40.62
Δg_{red} (EGM2008)	57.13	−104.06	−4.33	±16.35
Δg_{res} (RTM)	30.87	−63.49	−0.76	±10.31

Table 8.9 Statistics of (**a**) the local gravimetric geoid models and (**b**) the differences between GPS/levelling and gravimetric geoid heights from the local models at 64 GPS/levelling benchmarks (Unit: m)

	Max	Min	Mean	std
N_{res}	0.355	−0.259	0.088	±0.108
$N_{EGM2008}$	45.919	28.247	39.937	±4.331
N_{RTM}	0.060	−0.047	0.001	±0.014
$N_{grav} = N_{res} + N_{EGM2008} + N_{RTM}$	45.784	28.427	40.026	±4.296
$N_{GPS} - N_{EGM2008}$	−0.171	−1.344	−0.540	±0.210
$N_{GPS} - N_{grav}$	−0.249	−1.277	0.616	±0.160

(6.97 mGal), the std by 60% (24.27 mGal). An even smoother gravity anomaly field results after the removal of the RTM-effects from the EGM2008 reduced gravity anomalies (see statistics in the last line of Table 8.8). Then, the construction of the final 2′ grid was carried out using LSC as in the previous numerical test. From the residual gravity field geoid heights were computed on the same 2′ grid by the 1D FFT method in spherical approximation. Finally, the contribution of the EGM2008 and the effect of the topography were added back and the complete gravimetric geoid model with reference to EGM2008 has been derived.

In Table 8.9 the statistics of the final gravimetric geoid model is given along with the statistics of the residual geoid field and the contribution of EGM2008 and that of the topography through RTM to geoid heights. In Fig. 8.19 a representation of the 2′ gravimetric geoid solution is shown with reference to EGM2008 and in conjunction with a RTM terrain effect estimation.

A first note refers to the contribution of the topography to geoid heights through RTM which has a range of approximately 11 cm and a std of ±1 cm. This contribution is still significant since an absolute accuracy for a geoid determination at the cm level is the basic requirement nowadays for a wide number of applications in geodesy and geosciences. The validation of the final gravimetric geoid model of this test is also performed over the network of the 64 GPS/leveling benchmarks in the area under study. The statistics of the differences between the GPS/leveling heights and the gravimetric geoid heights are also presented in Table 8.9. From these results it is obvious the significant improvement that EGM2008 brings to geoid modeling even in local and regional applications. A clear indication of the superior performance of EGM2008 geoid model is the std of the above mentioned differences compared with that of the combined geoid models derived from GOCO02s in

conjunction with gravity and topography (see Table 8.7). This std is better at a level of about $\pm 20 - \pm 50$. Comparing the performance of the local gravimetric geoid model to EGM2008, it can be concluded that it gives better results than EGM2008 by ± 5 cm in terms of std. Moreover, the range of the differences for the local gravimetric geoid model is smaller by ~ 15 cm. All this numerical assessment is a good indication that even in the presence of high-resolution and high-accuracy GGMs, like EGM2008, local and regional gravimetric geoid models have still to offer and need not to be abandoned.

8.7 Summary and Concluding Remarks

The topographic, bathymetric and compensated masses have a significant contribution to gravity field modeling in general and gravimetric geoid determination in particular, since they provide the high-frequency content of the gravity spectrum through the available reduction methods. The mass reductions have a two-fold effect to gravity field constituents, i.e., the direct effect and the indirect one. The former is of main importance towards the smoothness of the gravity observations by removing mainly local and regional mass effects and thus improving the accuracy of prediction, gridding and interpolation operations. The latter is also crucial, since this indirect effect should be restored to the reduced gravity field observables, e.g., geoid or quasi-geoid heights, following the remove-restore methodology which has been widely used in physical geodesy applications. Larger indirect effects can probably result in larger prediction errors, which will be propagated to the estimation of geoid heights and other gravity field parameters. It should be noticed that an additional operation in the remove-restore scheme is the use of a GGM, like EGM2008, which further reduces irregularities in the long-wavelength band of the gravity spectrum.

The high resolution and accuracy DTMs/DBMs are used to compute mass effects to gravity and indirect effects to geoid/quasi-geoid heights. Recent satellite missions like SRTM and ASTER have improved considerably the knowledge of Earth's topography in a global scale with homogeneous coverage. The combination of these models with national DTMs towards the elimination of problems mainly affecting the global models, as, e.g., roof-top effects, gaps in mountainous areas, resulted in the production of digital elevation models suitable for gravity field modeling applications in local and regional scale. A corresponding improvement has been also made with respect to bathymetric models in marine areas, taking advantage from the huge amount of the nowadays available multi-satellite altimety data, through which an inverse computation of bathymetry is performed from sea surface heights. The so resulted DBMs are appropriate for the computation of different kinds of mass reductions either individually or in combination with DTMs, especially in mixed, both land and marine, regions.

Some of the reduction methods discussed in the previous sections, as the RTM, Helmert and TC reduction schemes are extensively used in geoid/quasi-geoid determination. RTM is mainly connected with quasi-geoid determination, produces

smooth gravity field and the restored terrain/bathymetry effect on geoid heights or height anomalies is generally small. It is important to stress that in local geoid determination, by subtracting a global gravity model we also eliminate a large part of the terrain correction and its isostatic compensation; so the only logically related correction method is RTM. It has to be noticed that the problem left open with RTM is that the reference topographic surface used for it might not be exactly equal to the one that has been used to compute the global model.

Rudzki's reduction scheme has indeed zero-indirect effect in geoid heights, but it produces a rough residual gravity field and as such it has been used in a restricted number of local or regional geoid estimates. The Poincaré and Prey reduction refers to the need of computing gravity inside the Earth and it is contained as an intermediate step in other reduction schemes, downward continuation techniques and the computation of orthometric heights. The Bouguer and isostatic models (PH, AH) produce smooth residual gravity anomalies, but their indirect effect on the geoid is generally large. These models are widely used in geophysical studies since the produced residual field has a physical meaning towards the understanding of the compensation and condensation principles of the Earth's masses. Given the above, the appropriate choice of the mass reduction scheme is directly connected with the requirements posed by the specific application either oriented to geodesy or geophysics.

The direct computation of the different mass reductions to gravity (topographic, bathymetric, isostatic effects) is primarily carried out by (a) NIM using the mass prism or mass line representation, (b) FFT algorithms and (c) a combination approach based on NIM and FFT techniques. The same methods are used for the estimation of indirect effects on the geoid, unless an approximation is followed and the higher order terms are not taken into account in the computations. The NIM is a very time-consuming procedure when handling large amounts of digital height or depth models, which is the case of modern day geodetic research. On the contrary, the 2D FFT methods are computationally very efficient in the treatment of large amounts of gridded data and, in connection with some appropriate techniques (e.g., zero padding) towards the elimination of circular convolution and edge effects, give almost identical results with those obtained by rigorous discrete NIM. Furthermore, some well-known shortcomings of the spectral methods (singularity problems, numerical instabilities due to series convergence problems in cases of dense terrains with inclinations higher than 45°) can be overcome using different alternatives. Firstly, the FFT method can be combined with NIM; the latter is applied in an inner zone around each computation point and the former is evaluated in the rest of the test area for the sake of computational efficiency. Secondly, the rigorous 3D FFT can be employed, but this is also a time-consuming method that needs additional spectra computations and, generally, extensive computer resources.

In case that surface 2D density values are available along with a DTM of the same resolution, the reduction schemes outlined before can be easily evaluated by simply inserting the density values under the convolution integral and employing again one of the above mentioned methods.

As a general comment, it is evident that a denser digital model reflects better the topography or bathymetry masses than a coarser model, but special attention should be paid in the computations either using an analytical or an FFT-based method.

Ongoing research is focused on the refinement of mass reduction formulas and gravity reductions as well, in conjunction with the optimal combination of geopotential model and mass contributions in the frame of a remove-restore scheme. Further optimal remove-restore operations with reference to GGMs derived from a combination of, e.g., EGM2008 and GOCE-based or pure satellite models, may pose the requirements of even higher spatial resolution DTMs/DBMs/DDMs in order to better infer the frequency information of the gravity spectrum. The latter can form a crucial step towards the geoid determination with an absolute accuracy at the cm level and a relative accuracy better than 1 ppm.

Chapter 9
Marine Gravity and Geoid from Satellite Altimetry

Ole B. Andersen

Abbreviations

AVISO	Archivage, Validation et Interprétation des données des Satellites
CHAMP	CHAllenging Minisatellite Payload (German satellite)
CRYOSAT	Cryosphere Satellite
DNSC	Danish National Space Centre
DOV	Deflection of the vertical
ECMWF	European Centre for Medium-range Weather Forecasts
EIGEN	European Improved Gravity model of the Earth
EGM	Earth Gravity Model
ERM	Exact Repeat Mission
ERS	European Remote-sensing Satellite
ESA	European Space Agency
Envisat	Environmental Satellite
EUMETSAT	Eur. Organ. for the Exploitation of Meteorological Satellites
FFT	Fast Fourier Technique
GM	Geodetic Mission
GGM	Global Geopotential Model
ICESat	Ice, Cloud and Elevation Satellite
JPL	Jet Propulsion Laboratory
Geosat	Geodetic Satellite
GFO	Geosat Follow-On
GOCE	Goddard Ocean Tide model
GPS	Global positioning system
GRACE	Gravity Recovery and Climate Experiment
GRAVSOFT	Gravity prediction Software
GSFC	Goddard Space Flight Center

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_9,

KMS	Kort- og Matrikelstyrelsen (National Survey and Cadastre, Denmark)
LSC	Least Squares Collocation
MDT	Mean Dynamic Topography
MSS	Mean Sea Surface
NCTU	National Chaotung University (Taiwan)
NOAA	National Oceanic and Atmospheric Administration
OSU	Ohio State University
PO.DAAC	Physical Oceangraphy Distributed Active Archiving Center
RADS	Radar Altimeter Database System (http://rads.tudelft.nl/)
RMS	Root Mean Square
SAR	Synthetic Apertur Radar
SRAL/SIRAL	Syntetic Interferometric Radar Altimeter
SS	Sandwell and Smith
SWH	Significant Wave Height
TOPEX	Topography Experiment

9.1 Outline of the Chapter

Two thirds of the globe is covered with water, and large parts of the ocean are not covered with marine gravity observations. In large parts of the Southern Pacific Ocean the distance between surveys lines are several hundred kilometres thus only resolving signals of twice that distance. Satellite altimetry can provide information of the height of the oceans over nearly 60% of the Earth surface. These data can be used to derive a high resolution global marine gravity field with an accuracy ranging between 2 and 4 mGal.

In this chapter satellite altimetric data are introduced and the importance to global geoid and gravity mapping is demonstrated. Individual satellite altimetry observations might not provide as accurate measure of the gravity field as those by marine gravity, but the ability to provide a near global uniform accurate gravity field makes satellite altimetry un-surpassed and essential for determining the high resolution global marine gravity field of the Earth.

Initially the altimetric sea surface height observations is described. Then the process of isolating residual geoid signal is covered. Subsequently, methods for converting altimetric sea surface height observations and/or sea surface slopes to global marine gravity are described. The accuracy of the global marine gravity field is presented along with methods for combining satellite altimetry with marine and airborne gravity using least squares collocation. Finally some of the current frontiers and trends in development of the next generation global marine gravity fields are covered.

9.2 Altimetry Data

Prior to the space age global marine geoid and gravity field mapping of the worlds ocean relied on sparse measurements from surveying ships and tide gauge stations located along irregular local coastline. During the last three decades, satellite radar altimetry has revolutionized marine geodesy and proven to be an essential tool for recovery of the global marine geoid and gravity field especially in areas of sparse ship coverage (Zlotnicki 1984). Individual satellite altimetry observations might not provide as accurate direct gravity field observations as marine gravity, but the ability to provide near global accurately gravity field makes satellite altimetry un-surpassed for determining the global marine gravity field of the Earth.

Altimeter observations of sea surface height offer a fundamentally different way to measure the local gravity than that provided by space gravity missions such as GRACE, CHAMP or GOCE. Space gravity missions measure the gravity field directly at an altitude of 250–700 km. However, due to upward continuation short wavelength scale features in the gravity field is attenuated. Consequently only long wavelength features can be obtained from space gravity field missions. In terms of space-borne instrumentation only altimeters can measure the high resolution gravity field from space (in the range of 5–100 km). This is because the satellite altimetry indirectly measures gravity via measuring the geoid height variations at the sea-surface (by measuring sea surface height variations). Hereby satellite altimetry provide observations directly at the sea surface which is far closer to the gravity field sources in the Earth's crust responsible for gravity field variations in the 5–100 km wavelength.

The height of the oceans closely assembles an equipotential surface of gravity and dense observations of the height of the ocean have become an increasing important supplement to traditional terrestrial, ship borne and airborne observations.

Satellite altimetry works conceptually by the satellite transmitting a short pulse of microwave radiation with known power towards the sea surface, where it interacts with it. Part of the signal is returned to the altimeter where the travel time is measured accurately using atomic clocks. Accurate determination of sea surface height from the altimeter range measurement involves a number of corrections: those expressing the behavior of the radar pulse through the atmosphere, and those correcting for sea state and other geophysical signals.

During the design of a satellite mission one of the first steps is to make a decision of how the satellite is flown and how the orbital parameters are defined (e.g., inclination with respect to the Equator and repeat time). This will define the observational pattern of the satellite given by the ground track distance in Table 9.1 and also shown in Fig. 9.1, where the denser ground tracks is preferred for recovering high resolution marine gravity. The inclination in Table 9.1 also determines the maximum latitude covered by the satellites. All altimetric satellites leave a polar gap of different size which stresses the importance of recovering gravity in the Polar Regions through project like the Arctic Gravity field (ArcGP) project (Kenyon and Forsberg 2008). An inspection of the different inclinations in

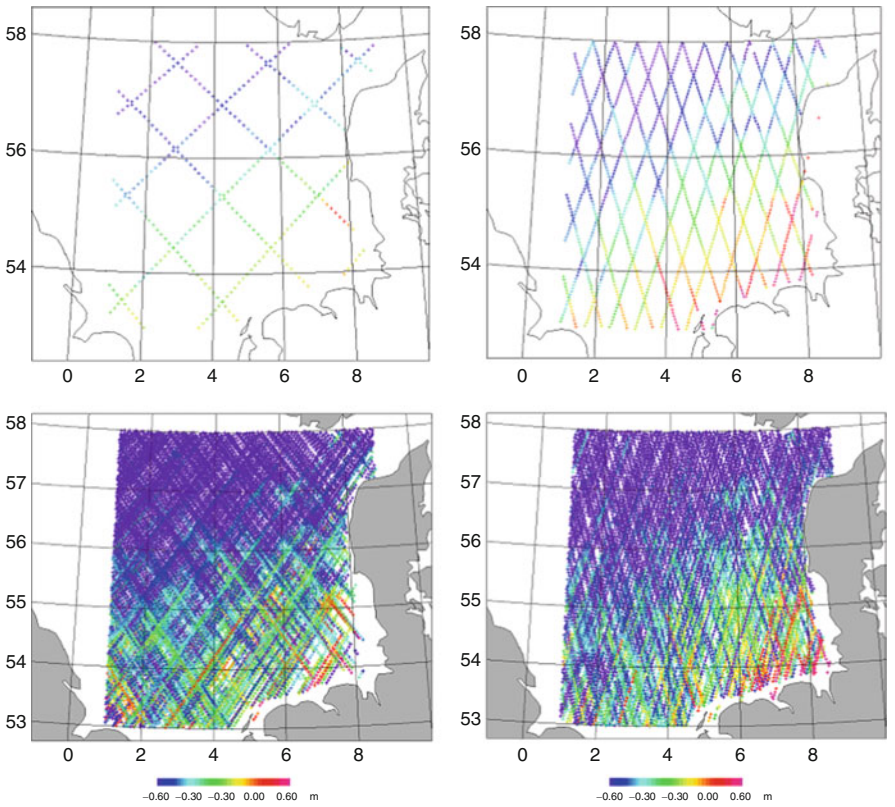


Fig. 9.1 Ground tracks patterns in the North Sea for Exact Repeat Missions (*upper*) versus Geodetic Missions (*lower*). The TOPEX/Poseidon and Jason satellites (*upper left*); ERS-1, ERS-2 and Envisat ERM (*upper right*); Geosat GM (*lower left*), and ERS-1 GM (*lower right*)

Table 9.1 reveals, that the ERS and Envisat satellites leave smaller polar gap than the TOPEX/Poseidon, Jason, Geosat and GFO. ICESat laser mission leaves a polar gap with a radius of 400 km and the newly launched Cryosat-2 leaves a polar gap with only 200 km.

Data from satellite altimeters are available as either exact repeat mission (ERM) in which the sea surface height observations are being repeated at regular intervals at a low spatial resolution. Such design is very important for oceanography and climate science, but not applicable for high resolution gravity field modelling at least for its stationary part. The geodetic mission (GM) data are far more interesting for geodesy. In the GM the satellite flies in a non-repeating orbit or an orbit with a very long repeat and hence, the sea surface height observations are only taken once at each location but at a much higher spatial density. Consequently, this configuration creates a much denser mesh of observations as shown in Fig. 9.1.

During the last 25 years, the eight satellites carrying altimeters (Geosat, GFO, ERS-1, ERS-2, Envisat, TOPEX/POSEIDON, Jason-1 and Jason-2) have recorded

Table 9.1 Specifications for recent and ongoing satellite missions carrying altimeters

Satellite	Duration	Inclination (degrees)	Repeat times (days)	Track distance at equator (km)	Noise (m)
Geosat	1984–1988	108	∼ 3, 17	4, 150	0.07
ERS-1	1991–1996	98	3,35,356	900, 75, 8	0.06
ERS-2	1995–2006	98	35	75	0.05
TOPEX/Poseidon	1992–2006	66	∼ 9.9156	315	0.04
Jason-1	2002–2008	66	∼ 9.9156	315	0.03
Jason-2	2008–>	66	∼ 9.9156	315	0.03
GFO	2001–2008	108	17	150	0.06
Envisat	2001–>	98	35	75	0.04
ICESat	2002–>	94	90	‘110’	0.04
Cryosat-2	2010–>	88	369	7	0.01

more than 60 years of ERM observations (over a period of 25 years), whereas less than 2.5 years of GM altimetry have been recorded. The only two geodetic missions were performed by the ERS-1 and Geosat satellites. During 199 the ERS-1 performed two interleaved repeats of 168 days resulting in a uniform global dataset having 7 km along track resolution and 8 km across track resolution at the Equator. The Geosat GM lasted 1.5 years during 1985 and 1986. However, these data were not declassified by the US navy until 1995. The Geosat GM did not have a constant across track distance, as Geosat was put in a drifting ∼3-day orbit during the GM. A total of 35 million altimetric sea surface height observations with an average track distance of 6 km at the Equator are available from this mission within the $+/-72°$ parallels. Examples of the ground track pattern measured by the TOPEX & Jason-1 ERM; ERS and Envisat ERM; Geosat ERM and ERS-1 ERM are shown in Fig. 9.1.

The major problem for the recovery of high resolution gravity is the fact that only the old and relatively in-accurate GM data (compared with present day altimeters) have adequate spatial resolution. Consequently the geodetic community has made every effort possible in order to enhance the quality and the resolution of the GM data (Yale et al., 1995). This is because the accuracy of the derived gravity field is directly proportional to the accuracy with which the sea surface height can be determined.

Sea surface height accuracy has been improved dramatically over the last decade through a reanalysis of the old data applying a technique called retracking. Retracking describes the way a mathematical model is fitted to the returned power from the sea surface also called the waveform. From the parameters derived to fit the chosen mathematical model the sea surface height is derived. Below retracking is briefly introduced for interested readers.

9.3 Retracking

Over the ocean the power returned from the sea surface has a characteristic waveform shape as a function of time which was mathematically described by Brown (1977), and this general form has since been called a Brown waveform.

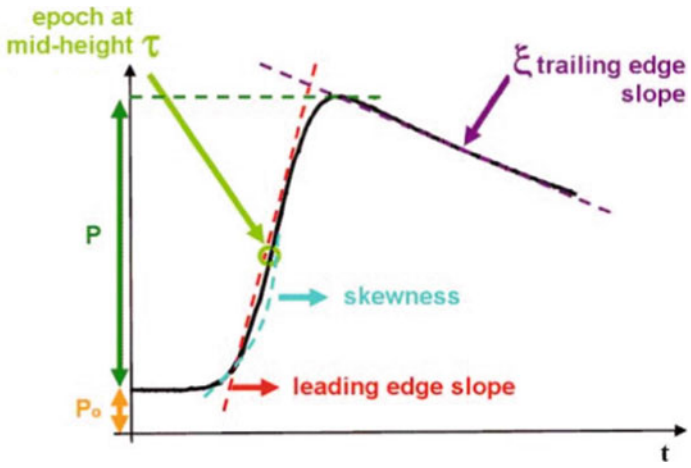


Fig. 9.2 The returned power as a function of recording time for a typical altimetric observation over the ocean modeled as a Brown waveform (The figure has been modified from ESA (www.esa.int))

A total of six parameters can be seen to determine the waveform as shown in Fig. 9.2. These are: the epoch time at mid height, the trailing edge slope, the leading edge slope, the skewness, the thermal noise (P_0) and the amplitude of the signal (P).

The epoch time at mid height where the waveform have risen to half its full power, is defined to determine the exact time of the return pulse defining the height of the sea surface (by multiplying with the speed of the radar pulse and dividing by 2 for the return of he signal). The 'leading edge slope' reflects the scattering of the radar signal by the sea surface. Higher waves will create more uniform distribution of the returned power and consequently, the 'leading edge slope' will be low. In the opposite situation where the surface is flat (acting like a mirror) the power will be returned instantly, and the leading edge and trailing edge slopes will be nearly vertical.

Maus et al. (1998) pointed out that in the least squares estimation of the six parameters defining the Brown waveform, the correlation between the 'leading edge slope' and the epoch time is very high. This leads to the development of a secondary re-analyzing of the waveform data through re-tracking (also called repicking) of each 18 Hz individual waveforms. In this secondary run the leading edge slope or equivalently the significant wave height is fixed from the first re-tracking run through smoothing and the estimation can be limited to a few parameters which result in a much more robust and smooth sea surface height estimation as shown in Fig. 9.2. This proved to be particularly important particularly for the ERS-1 data where the thermal noise (P_0) is suppressed.

The second important finding was the fact that between 6% and 9% of the (ERS-1) data are rejected globally by the Brown retracker applied by the space agencies, as their retracker proved to be too restrictive. This leads to the development

of a suite of more tolerant retrackers by Berry et al. (2005) to account for reflections from various surfaces. This later proved to be particularly important in coastal and polar region where the percentage of non-Brown waveforms increases dramatically (Andersen et al. 2010a). This increased the number of altimetric observations significantly as also shown later in Fig. 9.27. This development will be further described in Sect. 9.14.4, as a major contributor to the improvement of high resolution global marine gravity field modeling over the last 10 years can be directly associated with retracking and improved accuracy of sea surface height estimation.

9.4 Sea Surface Height Observations

Altimeter data are distributed through agencies like, EUMETSAT, AVISO, PO.DAAC and NOAA. In addition to these operational data centers, the Radar Altimeter Database System (RADS) delivers harmonized, validated and cross-calibrated sea level database from all altimetric missions (see Appendix A).

The altimeter measures the range to the sea surface and the (retracked) altimetric range observations are initially corrected for a number of range corrections to model the behavior of the speed of the radar pulse (speed of light) through the atmosphere. The range corrections also accounts for the interaction with the sea surface through the sea state correction (e.g., Andersen and Scharroot 2011; Fu and Cazenave 2001). The height of the spacecraft is determined relative to the reference ellipsoid through Precise Orbit Determination and more recently including GPS (Fu and Cazenave 2001). Combining the knowledge of the height of the spacecraft with the corrected range gives the sea surface height relative to the reference ellipsoid as also shown in Fig. 9.3. The sea surface height h can, in its most simple form, be described according to the following expression

$$h = N + \zeta + e \tag{9.1}$$

Where N is the geoid height above the reference ellipsoid, ζ is the time-variable sea surface topography, and e is the error.

In geodesy the geoid N (or the geoid slope) is the important signal. In oceanography the sea surface topography ζ is of prime interest.

The geoid N can be described in terms of a long wavelength reference geoid N_{REF}, and residuals ΔN to this. Similarly the sea surface topography can be described in terms of a mean dynamic topography (ζ_{MDT}) and a time varying sea surface topography ($\zeta(t)$) also called the dynamic ocean topography (DOT). Normally the largest contributors to the time varying sea surface topography ($\zeta(t)$) are removed as part of the standard set of geophysical corrections. These include the tidal correction and the dynamic atmosphere correction. The ocean tide correction is responsible for more than 75% of the total signal variance (Andersen, 1995, Ray 1991). The dynamic atmosphere account for less than 10% of the total signal variance and include a correction for the atmospheric pressure effect, as the sea level

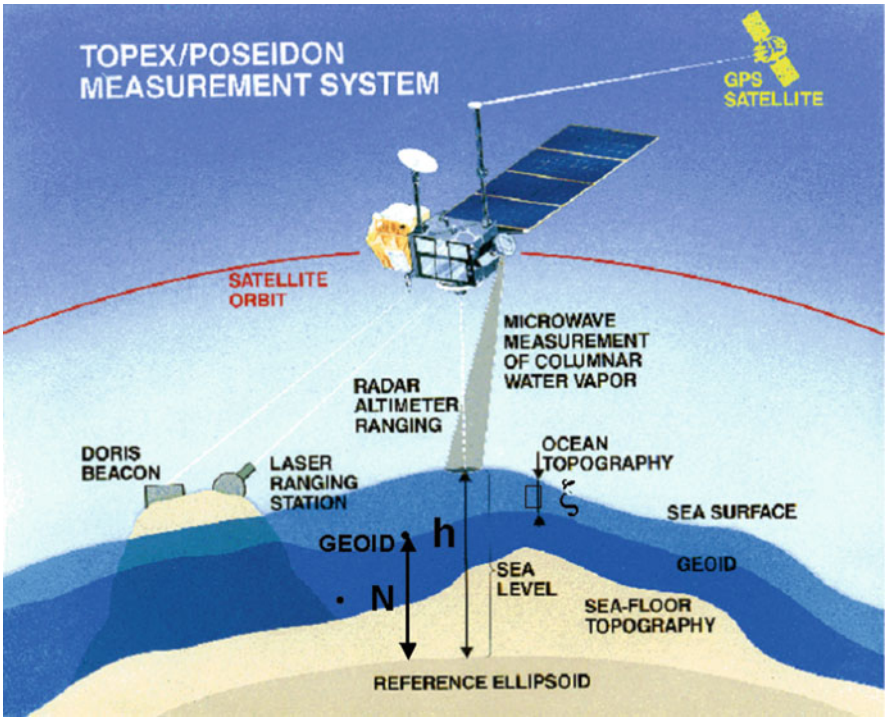


Fig. 9.3 Schematic illustration of the satellite altimeter measuring principle. The sea surface height (h) relative to the reference ellipsoid is the sum of the Geoid N and ocean topography (Figure modified from AVISO)

react as a huge inverted barometer coming up with the atmospheric pressure is low and going down when the pressure is high.

This way the time varying sea surface topography ($\zeta(t)$) will only contain contributions from primarily wind and other high frequency effects. Sea surface height can then be written like:

$$h = N_{REF} + \Delta N + \zeta_{MDT} + \zeta(t) + e \tag{9.2}$$

The interesting quantity for gravity field modeling is the residual geoid height ΔN. The accuracy with which this quantity can be determined is directly related to the accuracy with which the other contributors in (9.2) can be determined. Consequently, it is important to model and remove these as accurately as possible which is the focus of the subsequent section.

Assuming that N_{REF}, ζ_{MDT} and $\zeta(t)$ are all of long wavelength characters then these are almost identical between two neighbouring altimetric point (h_i,h_j) some kilometres apart. Consequently the difference becomes equal to the slope of the residual geoid signal along the altimeter track like

$$h_i - h_j \approx \Delta N_i - \Delta N_j + e \approx \partial N + e \tag{9.3}$$

The geoid slopes is closely related to the deflections of the vertical (DOV) in the north and east direction called (ξ, η) as defined in (2.101) and later in this chapter their use with altimetry will be described in detail.

The major argument of using DOV rather than geoid heights is the fact that DOV values are less contaminated by long-wavelength errors as will be demonstrated easier to process as the user does not need to go to model and remove long wavelength signals and particularly the time-variable dynamic sea surface topography (ζ(t)).

There are, however two drawbacks of using slopes compared with direct height observations. The first stems from the inclination of the satellites (108° and 98° for Geosat and ERS, respectively). This means, that at low latitudes the geoid slope in the north-south direction is derived much more accurately, than the east-west slope. Similarly the north-south-slope is less accurate derived at the maximum latitudes of the satellites (see Sandwell and Smith 1997 for details). The second drawback is the fact that in shallow water regions, the spatial scales of the time-variable dynamic topography (ζ(t)) is scaled down with the square root of the depth and also amplified and the assumption that this quantity is identical from one altimetric observations to the next becomes questionably and the noise e is increased.

To get from along track slopes in (9.3) to DOV in the north and east directions several possibilities exist. By definition, the along-track DOV called ∂h defined as the along-track gradient of the geoid (with opposite sign) is given like

$$\partial h = -\frac{\partial N}{\partial s} \tag{9.4}$$

with s being the along track distance. Consequently a gridded geoid surface is needed. This can be created from e.g., a cubic spline fit to the along track altimetric geoid height data. Then the along track derivative is obtained by differentiating the spline. This approach, however, gives noisy DOV due to the interpolation error of the spline.

A better result is obtained by approximating the along-track DOV by the slope of two successive geoid heights.

$$\partial h \cong -\frac{N_2 - N_1}{d} \tag{9.5}$$

where d is the along track point spacing and the location of ∂h is the mean location of the two points. In order to derive the northern and eastern DOV from the along track DOV the following equation system is set up to determine these using several points in a small cell (cf. Sect. 1.9)

$$\partial h_i + v_i = \xi \cos\alpha_i + \eta \sin\alpha_i \qquad i = 1, \ldots, n \tag{9.6}$$

where v_i is the residual, α_i is the azimuth of ∂h_i, n is the number of points and (ξ,η) is the north and east component of the DOV. It must be noticed that the

along-track DOV from different satellite mission and at different latitudes have different azimuth, which complicate the use of (9.6) for resolving gridded northern and eastern DOV from along-track DOV.

At crossover location where one north going track crosses a south going track, the northern and eastern components of the DOV can be directly derived from the two along track DOV observations. This gives far better determination of the slopes. However crossover locations are infrequently spaced. A thorough description of the individual steps in the method is given by Hwang et al. (2002)

9.4.1 Mean Sea Surface and Mean Dynamic Topography

In a perfect world altimetric observations would be available over infinite time. This would mean, that the dynamic topography $\zeta(t)$ average out from repeated observations along exact repeated ground tracks making ($\zeta(t) = 0$) in (9.2). The surface defined by the repeated satellite observations would then be the mean sea surface (h_{MSS}), which is the sum of the geoid height N and the mean dynamic topography (ζ_{MDT}). This way (9.2) reduces to

$$h_{MSS} = N + \zeta_{MDT} + e = N_{REF} + \Delta N + \zeta_{MDT} + e \tag{9.7}$$

In case a "perfect" MSS with adequate resolutions existed then ΔN could be determined directly from this model. Present day MSS models like DNSC08MSS (Andersen and Knudsen 2009) are derived using the most accurate filtering of the temporal sea surface variability with a limited time span and simultaneously obtaining the highest spatial resolution. This is normally achieved by combining data from the highly accurate exact repeat mission (ERM), with data from the older non-repeating geodetic mission (GM) like ERS-1 and Geosat.

This also means that in between the repeat tracks the mean sea surface is only determined from the GM data. In order to obtain the "best" high resolution marine gravity field experiments have shown that it is more accurate to use the remove-restore of the geoid signal and crossover adjustment on individual tracks as proposed in the subsequent sections and not use the MSS as reference.

The mean dynamic topography (MDT) is the quantity bridging the geoid and the MSS and the quantity constraining large scale ocean circulation. Equations (9.7) also state, that a better estimation of the geoid and altimetric MSS is, in particular, expected to improve the determination of the mean ocean circulation (Wunch 1993).

The MDT has long wavelength character and ranges between +/-1.8 m as shown in Fig. 9.4 with highest values around the Equator and lowest values towards the Poles which shows that a major part of the MDT is due to thermal expansion in the upper layer of the ocean.

In order to derive gravity from residual geoid signal it is important to remove the MDT contribution from the MSS as mentioned in (9.2). Failure to do this will introduce a false gravity signal in the altimetric derived gravity signal. The lower part of Fig. 9.4 shows exactly this effect from failure to account for the effect

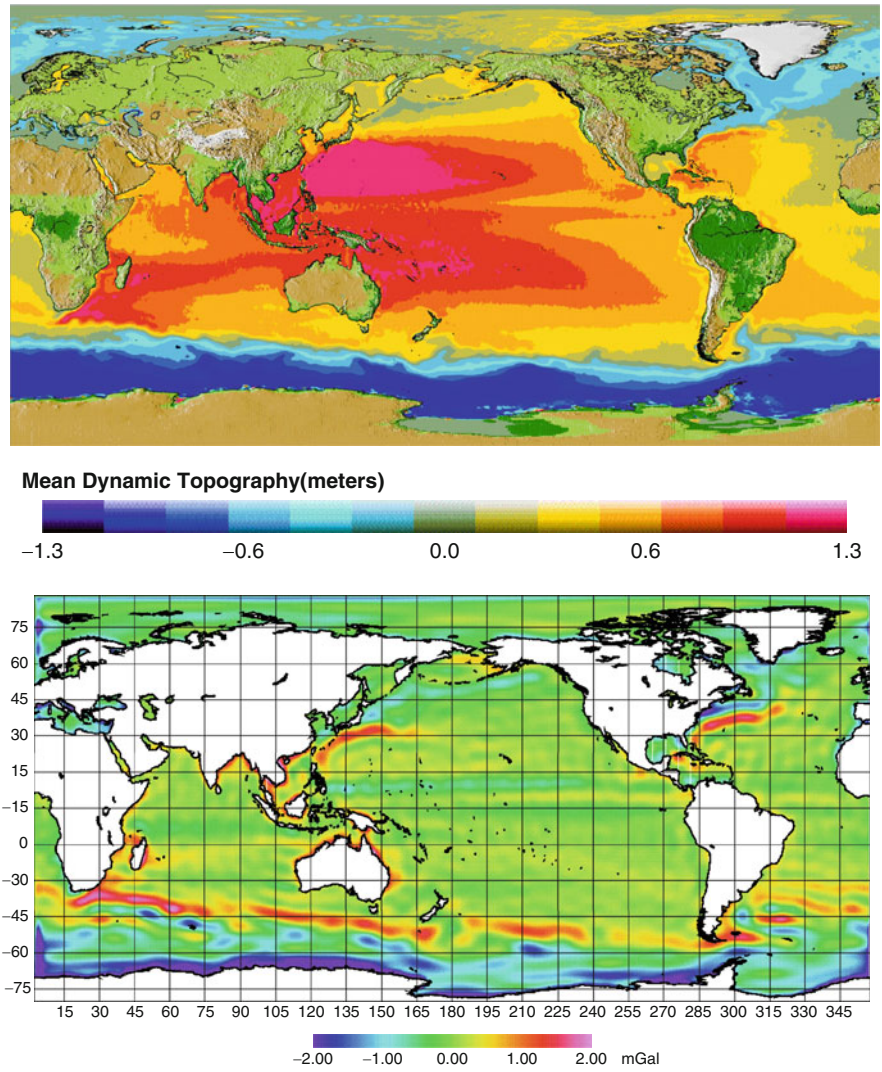


Fig. 9.4 The Mean dynamic topography (*upper figure*) and the false gravity signal (*lower figure*) caused by the mean dynamic topography (PGM07A) if this is not removed from the sea surface height observations prior to gravity field determination

of the MDT. The false gravity signal ranges up to 3–5 mGal in the large current regions even though the figure only shows the smoothed gravity effect ranging up to 2 mGal.

9.4.2 Remove-Restore for Satellite Altimetry

The use of the remove restore technique is extremely important for the efficient computation of short wavelength gravity field using altimetric sea surface height data. By removing a known reference geoid model (e.g. EGM96 or EGM2008) a residual geoid field is obtained, which is statistically more homogeneous and smoother than the total field. The removal of a reference field has the effect, that gravity field information outside the data-area is implicitly accounted for and the covariance functions will have smaller correlation distance (Part II, (7.82)). Therefore the computation can be carried out in smaller a region.

Along with the reference geoid the mean dynamic topography (ζ_{MDT}) must also be removed as described above. This gives the residual sea surface height h_{res} from (9.2) like

$$h_{res} \approx \Delta N + \zeta(t) + e \tag{9.8}$$

It is important to be aware of how much signal is removed along with the remove/restore of the geoid signal. This will be a function of the accuracy of the geoid as well as the degree and order used for the spherical harmonic expansion. An example of this is the new EGM2008 geoid (Pavlis, ibid.) which removes signal up to spherical harmonic degree and order 2,160. This is far more than most other geoid model like EGM96, GGM02, EIGEN-GL04, which only models geoid signal up to spherical harmonic degree and order 360, 200 and 150, respectively.

The residual signal can e.g. be evaluated using the Tscherning/Rapp degree variance model (Tscherning and Rapp 1974)

$$\sigma_i^{TT} = \begin{Bmatrix} \kappa_i & i = 2, ..., 2160 \\ \dfrac{A}{(i-1)(i-2)(i+4)} \left(\dfrac{R_B}{R}\right)^{i+1} & i = 2161..... \end{Bmatrix} \tag{9.9}$$

where the Bjerhammer radius $R_B = R - 7$ km and R is the radius of the Earth, $A = 1,571,496\ m^4/s^4$, i is the degree and κ_i is the error degree variance of EGM2008.

Evaluating (9.9) gives a residual geoid signal of 4–5 cm and a correlation length of 7–9 km once EGM2008 has been removed up to degree and order 2,160. This compares to a residual signal of 30–40 cm and correlation length of 20–25 km for the EGM96 geoid model complete up to degree and order 360.

9.4.3 Dynamic Sea Surface Topography

The amplitude of the dynamic sea surface topography $\zeta(t)$ – recalling that ocean tides and atmospheric pressure have been removed – will be largest in the major current systems such as the Gulf Stream, the Kuroshio Extension in the Pacific Ocean, the Antarctic Circumpolar Current, and in the coastal regions as seen in Fig. 9.5.

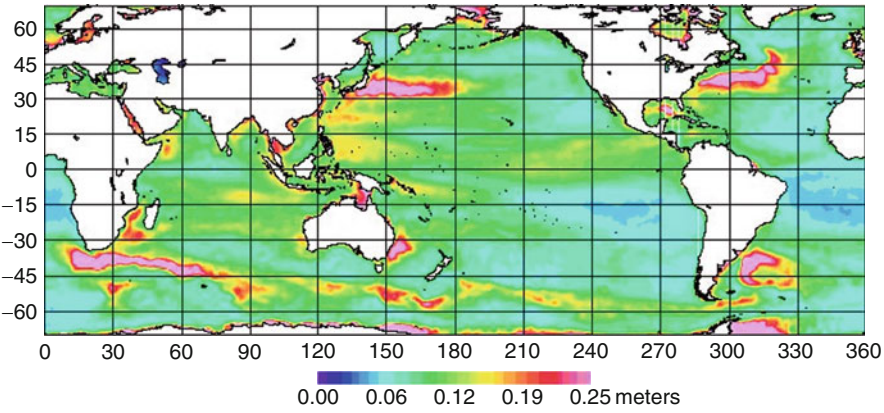


Fig. 9.5 Standard deviation of the timevarying dynamic sea surface topography from 6 years of Envisat altimetry

Data from repeated ERM like T/P, JASON1+2, GFO, ERS-2 and Envisat efficiently average out the dynamic sea surface topography through multiple observations at the same locations. However, the ground track spacing of these satellites (>75 km) does not enable adequate resolution for retrieving the high resolution gravity field. In non-repeating geodetic mission data the sea surface height observations are observed once and consequently measures must be taken to remove the dynamic sea surface topography that will otherwise contaminate the residual geoid height signal.

The dynamic sea surface topography $\xi(t)$ are mainly caused by wind, waves and pressure *and* generally has a long wavelength characters with wavelength longer than 100–200 km. Failure to remove this signal will create along track stripes in the derived gravity field known as the "orange skin" effect after the texture of an orange. The effect on one of the first altimetric mean sea surfaces is illustrated in Fig. 9.6.

Erroneous track related "orange skin" signal will result in large along track gravity field errors. One way of avoiding this is to use DOV values in stead of heights as stated previously (Olgiati et al., 1995).

Another way is to use sea surface height observations but to perform a cross over adjustment on the data. A cross-over adjustment uses the fact that the geoid residuals should be identical at all locations where ascending tracks cross descending tracks hereby mutually adjusting the tracks to limit track related errors. The cross-over adjustment is the subject of Sect. 9.5.

9.5 Crossover Adjustment

In order to remove the dynamic topography on particularly non-repeating geodetic mission tracks from the ERS-1 and Geosat missions a crossover adjustment is applied. The location where a descending track intersects an ascending track is

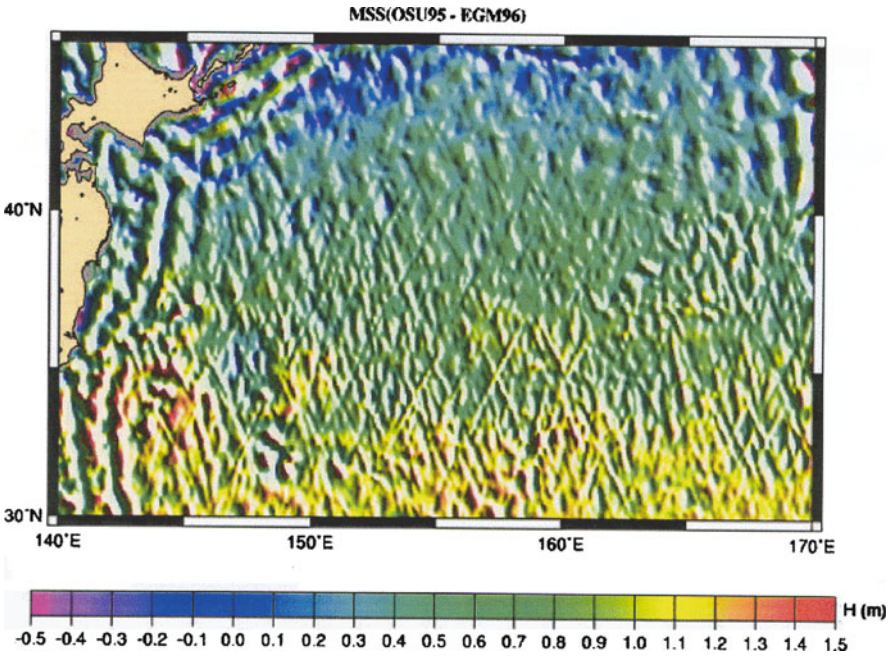


Fig. 9.6 The "orange skin" effect in the Kuroshio Extension in the Pacific Ocean from unmodelled dynamic sea surface topography. The picture shows the OSU95 mean sea surface (Yi 1995) relative to EGM96 (after Hernandez and Schaeffer 2000). The same orange skin effect will be visible if gravity is derived from these data

called the single satellite crossover location. Altimetric satellites are designed to create a fine interweaved net of tracks and diamonds for the use of orbit computation (see Fig. 9.1) and for GM missions this mesh is extremely fine. The crossover adjustment is carried out to limit track related errors and other long wavelength errors by minimizing height differences at crossover location between ascending and descending tracks.

The motivation for performing crossover adjustment is the assumption that the geoid signal is stationary at each location. With the launch of GRACE temporal geoid variations have been demonstrated (e.g., Andersen and Hinderer 2005; Andersen et al. 2005), but these are extremely small, and for the current investigation it can be assumed that the geoid is static.

Consequently the geoid height should be the same on ascending and descending tracks at crossing locations. On the contrary dynamic sea level signals should be different. Crossover discrepancies are computed as differences in sea surface heights between observations on north and south going tracks, like $d_{ij} = h_i - h_j$.

For very short arcs the track related errors can be modelled by a constant bias terms for each track, then

$$h_i - h_j = a_i - a_j + v_{ij} \tag{9.10}$$

where $(a_i, a_j,)$ are the unknown bias parameters related to the north and south-going track and v_{ij} are the residuals. On matrix form, this observation equation takes the form $\underline{d} = A\underline{x} + \underline{v}$, where $\underline{x}$ is a vector containing the unknown bias parameters. These are then estimated in a least squares adjustment (e.g. by minimizing the residuals, v_{ij}) like

$$\underline{x} = (A^T C_d^{-1} A + cc^T)^{-1} A^T C_c^{-1} \underline{d} \tag{9.11}$$

The equation system has a rank deficiency of one, so a constraint is needed. The constraint **c**, used is normally that the mean value of the biases should be zero, $\mathbf{c}^T \underline{x} = 0$ (Knudsen 1993).

For medium length arcs (e.g., shorter than 2,000 km) the track related errors can be modelled by bias and tilt terms. Then the residuals, v_{ij}, are minimized in a least squares adjustment of

$$h_i - h_j = (a_i + b_i \mu_j) - (a_j + b_j \mu_i) + v_{ij} \tag{9.12}$$

For longer arcs (e.g., longer than 2,000 km) the track related errors are not conveniently modelled using linear models (bias + tilt) but must be modelled using cosines and sine terms like

$$h_i - h_j = (a_i + c_i \sin \mu_j + d_i \cos \mu_j) - (a_j + c_j \sin \mu_i + d_j \cos \mu_i) + v_{ij} \tag{9.13}$$

where $(h_i - h_j)$ is a cross-over difference and $(a_i, b_i, c_i d_i, a_j, b_j, c_j . d_j)$ are the unknown bias, tilt and higher order parameters. μ_j and μ_i are the coordinates along the i'th and the j'th track of the cross-over points of the j'th and i'th track respectively. Here once could use orbital angles (true anomaly) times or longitudes coordinates but these are not exactly linear function of one another.

After remove the reference geoid (EGM96 or EGM2008) only relative short altimetry tracks needs to be investigated as shown in Sect. 9.4.2. Consequently, a crossover adjustment using bias and tilt is adequate. In this case the cross-over adjustment has a rank deficiency of four and the free or unknown surface is described by a bilinear function (Schrama 1989; Knudsen and Brovelli 1991):

$$D = s_1 + s_2 \mu_j + s_3 \mu_i + s_4 \mu_j \mu_i \tag{9.14}$$

The rank deficiency problem is illustrated in Fig. 9.7 for a free surface in a bias and tilt cross-over adjustment. The problem may be solved by fixing two parallel tracks (Rummel et al. 1993). Then two "master" tracks have to be selected, which can be difficult, since criteria for judging some tracks better than others are needed (here ERM tracks can be used). Normally, it is more attractive to do a "free cross-over adjustment" by applying a constraint that minimizes the free surface, (9.12), so the solution projected on the null space is zero. Such a constraint is given by Knudsen and Brovelli (1991) where a week minimum variance constraint is used.

Occasionally, the combined effect of removing the mean dynamic topography ξ_{MDT} and the free surface is that the altimetric surface does not have zero mean

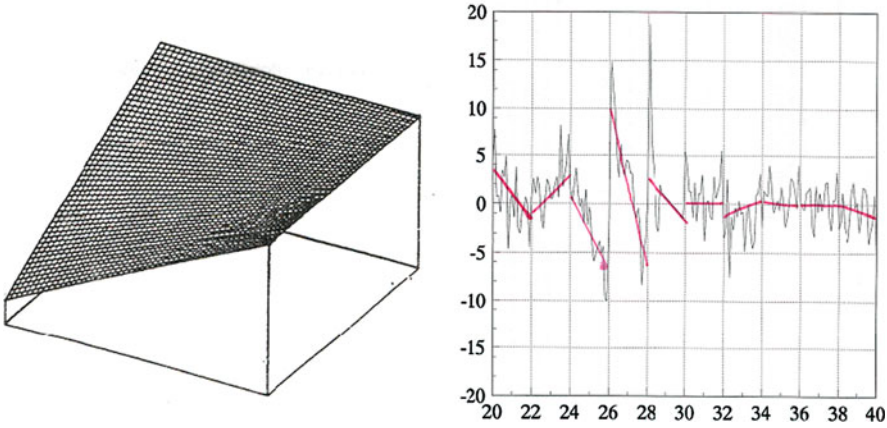


Fig. 9.7 *Left*. Illustration of the free surface in a bias and tilt crossover adjustment. *Right*: Gravity difference (mGal) along neighbouring areas of independent crossover adjustments for the KMS98 gravity field crossing the Hawaiian chain along 180°E. The block sizes used for the crossover adjustment is 2° latitude by 10° longitude with a 1° boundary in the computation

after the crossover adjustment, even if the area of computation is larger than the wavelengths included in the geoid model removed during the processing of the data. It may be corrected by re-estimating the parameter, s_1, s_2, s_3, and s_4, of the free surface, (9.14), and removing them from the data. The drawback is that some long wavelength parts of the residual geoid are removed. Hence, the altimeter observations will only represent the relatively short wavelength parts of the geoid residuals, δN, and the time-variable dynamic topography, $\delta\xi$, that is

$$h^c = \delta N + \delta\xi + v \tag{9.15}$$

The deviations between the altimeter data and the geoid model may alternatively be removed before the crossover adjustment by fitting each of the individual tracks to the geoid model. Again using a bias and a tilt for each track this may be carried out in a least squares adjustment minimizing the residuals, V_{ik}, along the ith track. That is

$$h_k = a_i^o + b_i^o \mu_k + V_{ik} \tag{9.16}$$

The residuals, V_{ik}, contain geoid and stationary SST of wavelengths shorter than the length of the ith track. For sufficiently long tracks the residuals may be used as geoid height observations. However, the cross-over discrepancies have not been minimized.

A joint fitting of the tracks to a geoid model and an adjustment of the cross-over discrepancies can be obtained by minimizing the residuals, v_{ij} and V_{ik} in (9.12) and (9.16), simultaneously (e.g. $a_i = a_i^o$ and $\mathrm{b}_i = \mathrm{b}_i^o$). In that case no rank deficiency and subsequent free surface problems exist, but relative weights of the residuals have to be determined in order to obtain satisfactory results. Hence, if a relative weight, w,

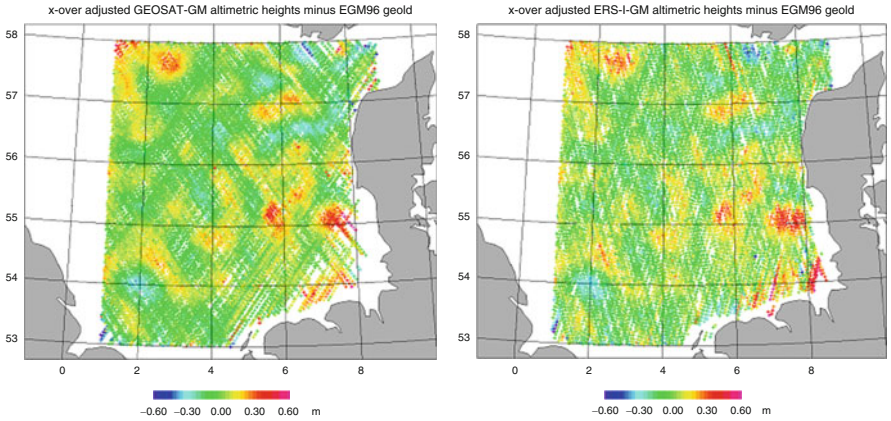


Fig. 9.8 The effect of a crossover adjustment on the Geosat (*right*) and ERS-1 (*left*) GM observations shown in Fig. 9.1

is applied on the residuals in (9.16), an adjustment of the following expression is carried out:

$$\sum v_{ij}^2 + w \sum V_{ij}^2 = \min \tag{9.17}$$

If the relative weight, w, is small the cross-over discrepancies are primarily minimized; if w is large the individual tracks are primarily fitted to the geoid model.

The effect of a crossover adjustment on the Geosat and ERS-1 geodetic mission observations, shown in Fig. 9.8, in a test example in the North Sea. The statistics of the effect of a-priori differences after a crossover adjustment, is shown in Fig. 9.9. The original Geosat and ERS-1 GM data prior to crossover adjustment is shown in Fig. 9.1. Notice that the colour scale for Figs. 9.1 and 9.8 is the same. The North Sea is known for very large dynamic sea surface topography signal. The crossover adjustment was carried out on the Geosat and ERS-1 independently and the resulting crossover adjusted picture shows a high degree of agreement between the two datasets. The comparison confirms that the residual signal is a consistent signal in both datasets and that the crossover adjustment has efficiently removed the dynamic ocean surface topography which leaves only the following signal in the residual sea surface height

$$h_{res} \approx \Delta N + e \tag{9.18}$$

A careful inspection of Fig. 9.8 reveals that the two datasets are not completely identical and some small differences still remain. The differences are typically outliers that can be picked up by the editing procedure described below, but also some residual track related signal can be seen.

The various steps described in this and the preceding sections have efficiently removed a reference geoid signal, the mean dynamic topography, and the time variable sea surface height signal in the geodetic mission data. This way, only the residual geoid signal remains in the altimetric sea surface height observations. This

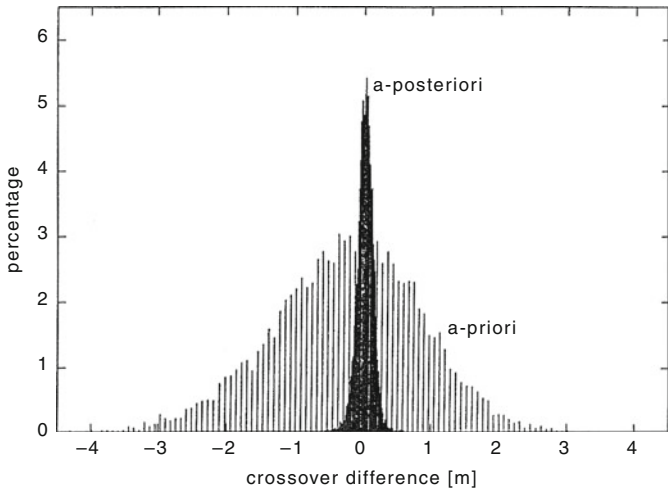


Fig. 9.9 Apriori and a-posteori crossover differences after a crossover adjustment of ERS-1 geodetic mission observations (Figure courtesy of Rummel et al. 1993)

data will be used to derive global high resolution marine gravity field in the next sections. However, prior to that it is important to describe the data editing and the error-budget of the sea surface height data as well as the huge improvement in data quality achieved through retracking the last 5–10 years.

9.6 Data Editing, Data Quality and Error-Budget

The quality of the derived high resolution gravity field is fundamentally dependent on the accuracy of the sea surface height observations and it is important to be aware of the accuracy of the input data as well as to carry out careful editing of the data.

All altimetric data have been edited for gross errors by the space agencies and the data distribution centres like AVISO, PODAAC or RADS. However, errors still remain due to wrong processing (e.g., retracking and wrong corrections), as well environmental errors like the presence of sea ice or coast. These errors frequently require more sophisticated methods to detect and remove.

Most outliers can be edited out by using standard editing criteria on the following information associated with the satellite altimetry data. All range and environment corrections should be present and within certain thresholds. The sea surface height and slope should be below a certain threshold.

As threshold either global numbers or local numbers based on the local conditions can be used (i.e., Hwang and Hsu 2003). One example of local conditional error removal technique is the technique used for the derivation of the DNSC08 global marine gravity field. This editing was applied on the residual geoid heights

after the removal of dynamic topography (crossover adjustment). This editing technique uses an efficient iterative de-spiking routine in which each altimetric observation is compared with the interpolated value from the nearest 64 points. For the interpolation a correlations length of 20 km is applied to ensure a smooth interpolation. If the point departs from the interpolated value by more than 2.5 times the standard deviation of the 64 local points the point is removed. This process was repeated iteratively using the reduced dataset until no further data points were removed. This was normally achieved in 3–5 iterations and generally removed between 3% and 6% of the altimetric sea surface height observations.

The main contributors to the errors e on the individual altimetric observations are the following:

$$e = e_{orbit} + e_{tides} + e_{range} + e_{retrack} + e_{environment} + e_{noise} \tag{9.19}$$

where

e_{orbit} is the radial orbit error
e_{tides} is the error due to residual tidal signal
e_{range} is the error on the range correction.
e_{retrak} is the error due to retracking
$e_{environment}$ is the error due to the presence of sea ice or coast
e_{noise} is the measurement noise.

For the use of DOV, the error on sea surface slopes must be derived. These are found from

$$e_{21} = \frac{\sqrt{e_1^2 + e_2^2}}{d} \tag{9.20}$$

Where e_1 and e_2 are the standard deviation on the consecutive sea surface height observations h_1 and h_2, and d is the distance.

For the un-retracked ERS-1 GM satellite altimetry, the error budget sums up to around 5–8 cm RMS (Scharroo, personal communication). This is roughly the same for the Geosat (Chelton and Schlax 1994). The error due to remaining tidal signal will increase in shallow water regions where the applied models are known to degrade (Andersen and Scharroo 2011). The various errors will be addressed more carefully in Sect. 9.12 which focuses on accuracy improvement. The error budget is naturally smaller than the sum of the errors in the applied models through the crossover adjustment which will removes long wavelength "errors" as well as long wavelength signal.

Figure 9.10 shows the huge improvement in the accuracy of sea surface slopes from retracking. The figure is courtesy of David Sandwell and shows the sea surface slope along six repeated ERS-1 profiles crossing the south Pacific with both high and low significant wave-heights. Slope errors were calculated using the 18 Hz measurements and slightly low pass filtering. A slope error of 1 μ-rad generally translate into a gravity error of 1 mGal (Sandwell and Smith 2005), so the retracking algorithm reduces the RMS error by 62% compared with the RMS error for the standard un-retracked data which corresponds to a 38% improvement in range precision (Sandwell and Smith 2005; Deng et al. 2003)

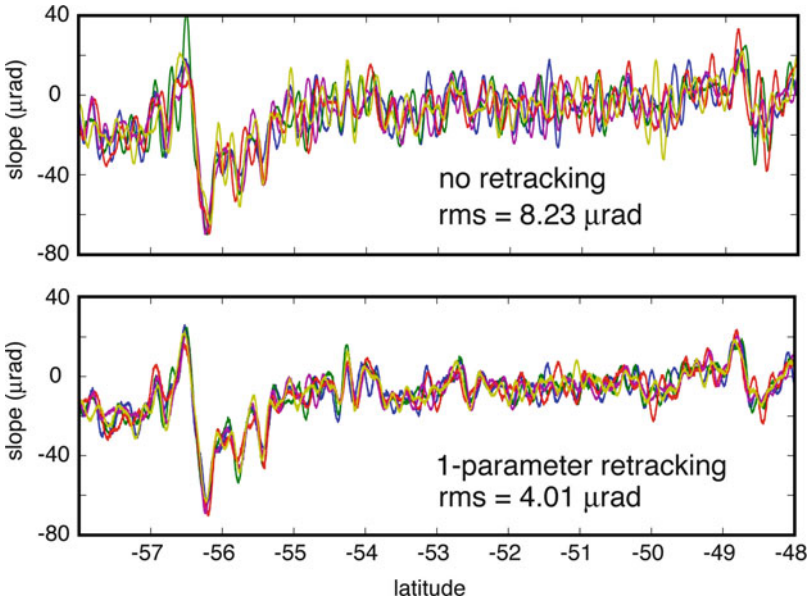


Fig. 9.10 Six repeated along-track sea surface slope profiles in the South Pacific Ocean. *Upper profile* is derived from the onboard tracker available in the waveform data record (RMS=8.23 μrad). *Lower profiles* have been derived from a one-parameter retracking algorithm constrained by smoothing the rise-time and amplitude parameters as in the text (RMS= 4.01 μrad) (Figure courtesy of David Sandwell)

Improvement in the height or slope accuracy through retracking directly translates into an improvement in both the accuracy but also of the resolution of the obtained gravity field. The higher accuracy of the sea surface height data means that the derived gravity field can be smoothed less, which again means that higher frequencies are retained in the derived gravity field. This was also demonstrated by Andersen et al. (2010b)

In many ways the ability to squeeze out more accurate gravity field information from retracking and reprocessing existing ERS-1 and Geosat GM datasets are close to being exhausted.

Fortunately, there are several new datasets coming in the near future which will bring a huge improvement in data accuracy, coverage and quality. The Cryosat-2 will firstly improve the coverage of the Arctic Ocean as it has an inclination of 88° bringing it 200 km from the North Pole. Secondly the repeat period is 369 days which gives higher ground track density than even the ERS-1 geodetic mission. Finally, the accuracy of the Delay-Doppler altimeter onboard Cryosat-2 will be a factor 2–3 better than the ERS-1 and Geosat altimeters theoretically bringing it down to 1 cm for 1-Hz data (Jensen and Raney 2005).

As for the error budget of the future ESA Sentinel-3 SRAL satellite to be launched around 2014 (R. Francis, personal communication, 2009) quotes a 0.8 cm

height accuracy for the Ku-band SAR altimeter for a SWH of 2 m. The height accuracy values for Cryosat-2 and Sentinel-3 should be compared with the 6 cm height accuracy for the ERS-1 and Geosat GM data so these satellites are expected to bring a quantum leap forward in accuracy of future gravity fields.

Finally, as both Jason-1 and Envisat are getting close to end of mission, there are a possibility that one of these satellites will be placed into a non-repeating geodetic mission for a limited time.

9.7 Gravity Recovery from Altimetry

For the use with satellite altimetry it is adequate to use a spherical approximation as described Part 2.5. The short wavelength residual geoid height N signal, isolated from satellite altimetry in the previous sections, can be expressed in terms of a linear functional applied on the anomalous potential T known as Brun's formula (2.36)

$$N = L_N(T) = \frac{T}{\gamma} \tag{9.21}$$

Where γ is normal gravity and T can be expanded into fully normalized spherical harmonic functions on the surface of a sphere with a radius R like in (Part III, 14.14). The anomalous potential T is a harmonic function satisfying Laplace's equation outside the masses

$$\Delta T = \frac{\partial^2 T}{\partial^2 \varphi} + \frac{\partial^2 T}{\partial^2 \lambda} + \frac{\partial^2 T}{\partial^2 r} = 0 \tag{9.22}$$

and Poissons equation ($\Delta \mathrm{T} = -4\pi\gamma\rho$) inside the masses ($\rho$ is density)

For the gravity anomaly Δg we use the spherical approximation which is related to the anomalous potential through the following functional similar to (2.100) like

$$\Delta g = L_{\Delta g}(T) = -\frac{\partial T}{\partial r} - 2\frac{T}{r} \tag{9.23}$$

This equation is frequently called the fundamental equation of physical geodesy.

Combining (9.23) and (9.21) shows that the gravity anomaly is related to the negative of the geoid slope (∂N) which is the quantity that can be computed from the altimetric sea surface heights.

For deriving gravity from altimetric sea surface slopes the deflections of the vertical (DOV) in the north and east direction (ξ, η) along the spherical unit vectors ($\boldsymbol{e}_\phi, \boldsymbol{e}_\lambda$) can be expressed similar to (1.183) like

$$\vec{\varepsilon} = L_\varepsilon(T) = \eta\vec{e}_\lambda + \xi\vec{e}_\varphi \tag{9.24}$$

where

$$\xi = -\frac{1}{\gamma\, r}\frac{\partial T}{\partial \varphi}$$
$$\eta = -\frac{1}{\gamma\, r\cos(\varphi)}\frac{\partial T}{\partial \lambda} \tag{9.25}$$

In the derivation of marine gravity from satellite altimetry two approaches are generally used.

One is the stochastic approach which predicts gravity directly from the observations using least-squares collocation (LSC). The major advantage of LSC for marine gravity field prediction is the fact that randomly spaced hybrid type data can be incorporated using statistical information about the errors in the data, and at the same time provide corresponding statistical information about the quality of the output gravity values. The drawback of LSC is the fact that it is very computational intensive, even with present day's computers. This approach is described in Sect. 9.8.

The other approach explores deterministic methods for the solution to Laplace's equation. This method requires global integration for the prediction of gravity in every single prediction point, which calls for huge computations and very fast computational methods. One particularly efficient method is a spectral approximation which requires that data have been interpolated onto a regular grid. This method has been widely used in the determination of global marine gravity during the last decade and is the scope of Sect. 9.9.

A hybrid approach in which LSC is used to interpolate the altimetric data points and fast spectral methods are used to evaluate (9.23) has also been widely used for local and global gravity field recovery and will be described in Sect. 9.10.

9.8 Least Squares Collocation for Altimetry

Least Squares Collocation (LSC) can be used to simultaneously determine both the signal and the error components (Wunsch and Zlotnicki 1984; Mazzega and Houry 1986; Knudsen and Brovelli 1991). The generalised form, which is presented here, has been documented by authors such as Tscherning and Rapp (1974), Rapp (1993) and applied to satellite altimetry by Knudsen (1993).

In its general form the relationship between the observations y_i and the anomalous potential can be written in the form

$$y_i = L_i(T) + e_i \tag{9.26}$$

where L_i is one of the functionals specified in Sect. 9.7, and e_i is an additive noise.

The gravity anomalies Δg are predicted from residual altimetric geoid anomalies h using the form

$$\Delta g = C_{\Delta gh}(C_{hh} + D_{hh})^{-1}h \tag{9.27}$$

Alternatively the gravity anomalies are predicted from residual geoid slope ε using

$$\Delta g = C_{\Delta g\varepsilon}(C_{\varepsilon\varepsilon} + D_{\varepsilon\varepsilon})^{-1}\varepsilon \tag{9.28}$$

An estimate of the a-posteriori error covariance of the gravity estimate is

$$\sigma_{\Delta g\Delta g} = C_{\Delta g\Delta g} - C_{\Delta gh}(C_{hh} + D_{hh})^{-1}C_{\Delta gh}^{T} \tag{9.29}$$

or for residual geoid slopes

$$\sigma_{\Delta g\Delta g} = C_{\Delta g\Delta g} - C_{\Delta g\varepsilon}(C_{\varepsilon\varepsilon} + D_{\varepsilon\varepsilon})^{-1}C_{\Delta g\varepsilon}^{T} \tag{9.30}$$

where the covariance matrices $C_{hh}, C_{\Delta gh}, C_{\Delta g\Delta g}, C_{\varepsilon\varepsilon}, C_{\Delta g\varepsilon}$ are the covariance matrices between height-height, gravity height, gravity-gravity, slope-slope and slope-gravity. The covariance matrices D_{hh} and $D_{\varepsilon\varepsilon}$ contain the noise variance of the geoid height and slopes, respectively. The elements of the covariance matrices in (9.27) and (9.28) can e.g. be calculated according to a mathematical model fitted to the observations using a program like "covfit" in the GRAVSOFT library (Forsberg and Tscherning 2008). If the different signal and error components are uncorrelated then the covariance values, C_{ij} and D_{ij}, are obtained by modifying the covariance to account for each of the signal and error components. For satellite observed sea surface height and the associated error the situation consists of several (assumed) uncorrelated terms and the covariances can be computed from (9.19) like

$$\begin{aligned} C_{hh} &= C_{NN} + C_{\xi\xi} \\ D_{hh} &= D_{e_{orbit}e_{orbit}} + D_{e_{tides}e_{tides}} + D_{e_{range}e_{range}} + D_{e_{rtrk}e_{rtrk}} + D_{e_{noise}e_{noise}} + D_{e_{env}e_{env}} \end{aligned} \tag{9.31}$$

The covariance values can be obtained using the kernel functions. The kernel associated with the gravity field can be derived using the spherical harmonic approximation for T (3.14) and the a-priori variances. The covariance between the anomalous potential T in the points P(φ,λ) and Q(φ',λ') is expressed as

$$E(P,Q) = \sum_{i=2}^{\infty}\sum_{j=0}^{i}\sigma_i^{TT}P_i(\cos\psi) \tag{9.32}$$

where σ_i^{TT} are degree variances and ψ is the spherical distance between the two points P and Q. Hence, (9.32) only depends on the distance between P and Q and neither on their locations nor on their azimuth (e.g. a homogeneous and isotropic kernel).

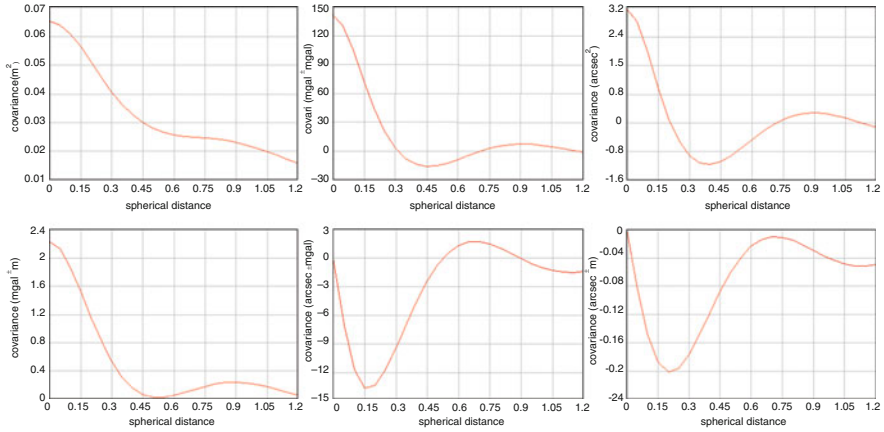


Fig. 9.11 Covariance functions associated with height anomalies, gravity anomalies and DOV in the upper panels. Cross-covariance between gravity and height, DOV and gravity and DOV and height in the lower three panels

Expressions associated with geoid heights and gravity anomalies and DOV can be obtained by applying their respective functionals on $E(\mathrm{P,Q})$, using covariance propagation like e.g. $C_{NN} = \mathrm{L}_N(\mathrm{L}_N(E(\mathrm{P}, Q)))$ for the geoid height following (5.48). Then

$$
\begin{aligned}
C_{NN} &= \sum_{i=2}^{\infty} \left(\frac{1}{\gamma}\right)^2 \sigma_i^{TT} P_i(\cos\psi) \\
C_{\Delta g \Delta g} &= \sum_{i=2}^{\infty} \left(\frac{i-1}{r}\right)^2 \sigma_i^{TT} P_i(\cos\psi) \\
C_{N\Delta g} &= \sum_{i=2}^{\infty} \left(\frac{i-1}{\gamma r}\right) \sigma_i^{TT} P_i(\cos\psi)
\end{aligned}
\tag{9.33}
$$

Accurate determination of the covariance function is important and is the subject of many studies in geodesy. An often used approach is to compute empirical covariances (program "empcov" of the GRAVSOFT package (Forsberg and Tscherning 2008)). Subsequently, these values might be fitted to preselected model-covariance functions like the Tscherning/Rapp model.

Modelling of the covariance function associated with the gravity field is described in Knudsen (1987a,b). As degree variance model, a Tscherning/Rapp model described in (9.9) can be used. The modeled covariance function associated with height, gravity anomalies and DOV is shown in Fig. 9.11 using degree variances for OSU91A to degree and order 360 (Rapp et al. 1991) with a scale factor of 0.207, gives the following typical covariance functions.

The ability to handle irregularly sampled data of various origins without the degradation due to interpolation makes LSC very well suited for computation of vertical gravity anomalies from along track satellite altimetry. The use of LSC should be considered upon creating local gravity fields where the computational cost is much smaller. A study by Hwang and Parsons (1995) demonstrated the use of LSC for computing gravity field in a limited area around Iceland.

The ability of LSC to handle irregularly sampled data of different origin is shown in Sect. 9.12, where the gravity field is predicted in coastal regions from altimetry and airborne gravity. In coastal regions LSC has the further advantage over methods FFT as the latter might suffer from high frequency noise due to the Gibbs phenomenon (Bracewell 1986a).

On global scales the computation of high-resolution gravity fields using LSC is simply not computationally feasible because of the huge amount of data ($>10^8$ altimetric observations globally). Even a computation of a local gravity field within a small cell of 1° by 1° can be problematic, as this cell might easily contains more than 2,000 altimetric data points which needs to be analysed in order to compute accurate covariance functions. However, computational power is steadily increasing, and the use of LSC for future evaluation of global high resolution altimetric gravity field will become feasible in the near future. Therefore methods and approximations are currently being investigated in order to enable global computation of marine gravity using LSC.

9.8.1 Interpolation Using Least Squares Collocation

For the production of existing global altimetric gravity fields LSC can conveniently be used in combination with spectral methods like Fast Fourier Techniques (described in the next section). This section will show how LSC can efficiently used to perform the interpolation of the altimetric observations onto a regular grid which is required for the evaluation of gravity using fast spectral techniques.

Interpolation by LSC can be handled by the "geogrid" program in the GRAVSOFT software package (Forsberg and Tscherning 2008) and will not be described here. This chapter will more focus on the adaption of the covariance function to local condition in the special case of interpolation of altimetric observations.

The local LSC prediction method in this program assumes a two dimensional isotropic covariance function described using a second order Markov function (Moritz 1980) as

$$C(r) = C_0 \left(1 + \frac{r}{\alpha}\right) e^{(-r/\alpha)} \tag{9.34}$$

r is the two-dimensional distance between the prediction point and computation point, and C_0 is the signal variance, and α is the correlation length (where a 50% correlation is obtained).

A special modification to the second order Markov function in (9.34) is sometimes applied for the interpolation of satellite altimetry due to the fact that the

satellite observations are provided along individual tracks and an error might be associated with all observations along a specific track. This is particularly so in coastal regions where the spatial scale of the sea surface variability can become so short and complicated that the assumption used in the cross-over adjustment (modelled using linear bias and tilt) becomes problematic.

A closer inspection of Figs. 9.8 and 9.18 illustrate this problem. To the north in the picture and in the German Bight in the lower right corner of the figure, some residual track related signal can be seen which also demonstrate that the crossover adjustment is not "perfect".

In order to limit the effect of this unwanted signal this error is modelled as an along track signal and in the interpolation this is accounted for by adding a covariance function for this error in the interpolation procedure. The error covariance function for this track related signal is applied to observations on the same track only (hereby assuming the error to be temporally uncorrelated)

Hence, for observations on the same track, the covariance function is modified to become

$$C(r) = C_0 \left(1 + \frac{r}{\alpha}\right) e^{(-r/\alpha)} + D_0 \left(1 + \frac{r}{\beta}\right) e^{(-r/\beta)} \tag{9.35}$$

where D_0 is the variance of the residual sea surface height and the β is the correlation length of this signal. For observations on different tracks D_0 is fixed at zero yielding an expression similar to (9.34).

Interpolation will unavoidably filter the observations; so much case must be taken in creating the optimal interpolation to limit this effect to create the most accurate gravity field anomalies. This along track modification to the second order Markov covariance function was originally derived for the KMS02 gravity field determination and subsequently refined for the DNSC08 gravity field prediction (Andersen and Knudsen 1998; Andersen et al. 2009). A practical use of this interpolation technique and the selection of interpolation parameters for the development of KMS02 will be shown in Sect. 9.11.

9.9 Deterministic Methods

The Stokes's integral formula and the solution to the Stokes's boundary value problem, have been described in Chap. 3 and the Stokes's formula (3.100) have been widely used to compute geoid undulations from gravity anomaly observations primarily on land. On the ocean, the problem is reversed as the satellites observes residual geoid signal. With satellite altimetry, the inverse Stokes formula, also known as the Molodensky's formula can be used to compute marine gravity anomalies from satellite altimeter derived sea surface heights (or geoid anomalies).

The inverse Stokes formula is a surface integral like

$$\Delta g_p = \gamma \frac{N_p}{r} - \frac{\gamma}{16\pi r} \int\int_{\sigma} \frac{N - N_p}{\sin^3(\psi/2)} d\sigma \tag{9.36}$$

where ψ is the spherical distance between the integration point (φ,λ) and the computation point (φ_p, λ_p).

Due to the properties of this integral kernel, the influence of more remote zones decreases rapidly and when using a remove/restore technique the integration radius can be limited to a few degrees (Wang 2001). There is a strong singularity at the innermost cell where $\sin^3(\psi/2)$ goes to zero. The approximation of this was treated by Lemoine et al. (1998).

The inverse Hotine's formula is related to the inverse Stokes formula and describes the relationship between the geoid undulations and the gravity disturbance and can be found in Zhang and Sideris (1996) and is similar to (3.20).

Gravity and geoid anomalies can also be derived from observations of north and east components of the DOV (ξ,η) using the inverse Vening Meinesz formula and the deflection-geoid formula (Hwang 1998).

$$\begin{Bmatrix} N \\ \Delta g \end{Bmatrix} = \frac{1}{4\pi} \begin{Bmatrix} R \\ \gamma \end{Bmatrix} \int \int_{\sigma} (\xi \cos\alpha + \eta \sin\alpha) \begin{Bmatrix} C \\ H \end{Bmatrix} d\sigma \tag{9.37}$$

where the kernel function H for the inverse Vening Meinesz formula related to deflection-geoid is given by

$$H(\psi) = \frac{\cos(\psi/2)}{2\sin(\psi/2)} \left(-\frac{1}{\sin(\psi/2)} + \frac{3 + 2\sin(\psi/2)}{1 + \sin(\psi/2)} \right) \tag{9.38}$$

where ψ is the spherical distance. The corresponding kernel function C for the deflection-geoid formula is given by

$$C(\psi) = -\cot\frac{\psi}{2} + \frac{3}{2}\sin\psi \tag{9.39}$$

Examples of the two kernel functions are shown in Fig. 9.12. Formulas for handling the innermost zone effect around zero spherical distance can be found in Hwang (1998) who showed that the asymptotic behaviour of the $H(\psi)$ and $C(\psi)$ for small ψ reduces to

$$H(\psi) \approx -\frac{2}{\psi^2} \qquad C(\psi) = -\frac{2}{\psi} \tag{9.40}$$

The global evaluation of both the inverse Stokes integral and the inverse Vening Meinis integral are allied to the surface spherical harmonic analysis and synthesis processes, and all the above formulas requires globally distributed observations for the accurate computation of a single gravity value. However, some modifications are required to make high frequency global gravity field modelling using this approach feasible. This is the subject of the following section.

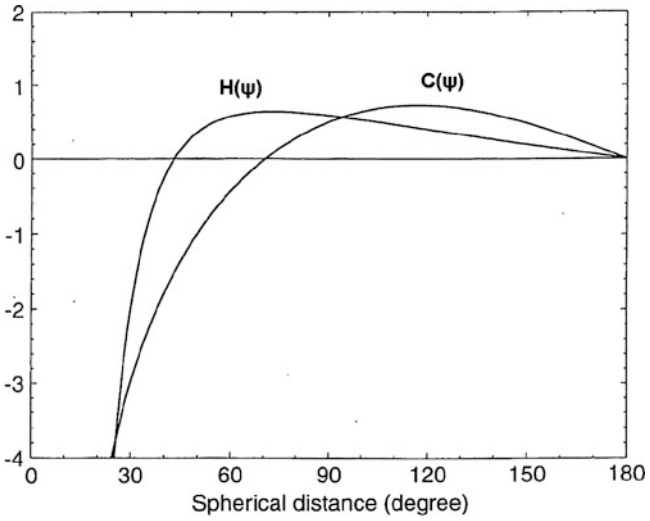


Fig. 9.12 The functions H(ψ) and C(ψ) as a function of spherical distance ψ

9.10 Fast Spectral Methods for Altimetric Gravity Prediction

Due to the enormous amount of altimetric data fast spectral methods have been used in all present global high-resolution gravity fields in one way or the other.

The most widely used spectral methods is the discrete Fast Fourier Techniques which has several advantages for fast computation, but which required data to be available as regular interpolated discrete values.

The fast spectral methods can be applied to evaluate the inverse Stokes integral (9.36) relating altimetric heights to gravity anomalies or for the evaluation of the surface integrals related to DOV in (9.38) or for evaluating the fundamental equation of physical geodesy relating geoid height to gravity (9.23).

It is still assumed that the long wavelength of the gravity field is adequately provided by set of spherical harmonic constituents (EGM96 or EGM2008) and that the long wavelength part of the signal is completely removed using the remove-restore technique. This way, accurate but approximate evaluations can be made over a limited spherical cap centred at the evaluation point. This way only a limited part of the global dataset must be investigated for the computation. The advantage is that this opens up for parallel computing as different areas can be computed independent of each other on different computers.

The second assumption is the data are regularly distributed in a grid. Such approximation requires that another step is introduced, namely, an interpolation or a gridding. One possible interpolation process was described in Sect. 9.8.1 using least squares collocation, but other interpolation processes like spline interpolation can also be used. Once data are available on a regular grid, the evaluation of the integral equations in (9.36) and (9.37) can very efficient be handled using spectral computational methods like the Fast Fourier Transform.

9.10.1 Fast Fourier Techniques for Altimetric Gravity

One of the fundamental advantages in terms of high resolution marine gravity field prediction is that FFT directly gives the result on the same grid as the input grid. This means that a single FFT run immediately gives the result in all data points. Furthermore the increased computational power is more or less linearly dependent on the number of grid points which makes evaluation on very dense grids like global 1 or 2 min grids possible. This means that the user should already in the interpolation step use the resolution of the wanted gravity grid.

The drawback of using FFT is the fact that data has to be provided on a homogenous interpolated grid which requires interpolation in the case of satellite altimetry. Furthermore FFT assumes data to be given at the same altitude but this is generally the case for satellite altimetry except for the few cases where data in e.g., lakes are used.

Gravity anomalies can be evaluated using spherical 1-D FFT methods. The spherical 1-D Fourier transformation was devised by Haagmans et al. (1993). In this method FFT is only applied in the longitude direction along each fixed parallel (φ_l). If a two dimensional grid is wanted, this can be achieved by combing sequences of 1-D FFT summarizing over all latitude bands. One dimensional spherical method has successfully been applied by e.g., Hwang et al. (2002) for recovering gravity anomalies from satellite altimetry.

2D FFT methods are available as spherical 2D FFT techniques (Strang van Hees 1990) or multiband 2D spherical FFT technique (Forsberg and Sideris 1993) as planar 2D FFT techniques (Schwarz et al. 1990). The detailed evaluation of the pros and cons of the various methods can be found in Part 7 of this book. The planar 2D requires the use of a flat Earth approximation and the introduction of a local Cartesian coordinate system.

For the sake of simplicity the derivation below is shown for a flat Earth approximation. Therefore, the computation of gravity anomalies is valid if the area only extends a few hundred kilometres in each direction (Part 7, Sect. 7.2). The flat Earth approximation is applicable as the remove-restore technique using either EGM96 or EGM2008, typically removes wavelengths longer than 100 km ensuring that only data within a limited cap is needed for the computation.

In the flat Earth approximation a local Cartesian coordinate system (x, y, z) is introduced and the formulas (9.24) and (9.25) reduces to

$$\begin{aligned} \Delta g &= -\frac{\partial T}{\partial z} - 2\frac{T}{R_e} \approx -\frac{\partial T}{\partial z} \\ \xi_y &= -\frac{1}{\gamma}\frac{\partial T}{\partial y} \\ \eta_x &= -\frac{1}{\gamma}\frac{\partial T}{\partial x} \end{aligned} \tag{9.41}$$

which in the frequency domain becomes

$$F(\Delta g) \approx -|k| F(T) = -\frac{|k|}{\gamma} F(N)$$
$$F(\xi_y) = -\frac{k_y}{\gamma} F(T)$$
$$F(\eta_x) = -\frac{k_x}{\gamma} F(T) \tag{9.42}$$

Where $F(T)$ represents the two dimensional discrete FFT of the grid of T values; $|k| = \sqrt{k_x^2 + k_y^2}$ and k_x, k_y are the wave-numbers equal to one over half the wavelength in the x and y direction, respectively.

Equation 9.42 shows, that the vertical derivative used to obtain gravity from residual geoid height in (9.41) is conveniently substituted by a Fourier transform and multiplication with the wave number followed by an inverse Fourier Transform.

The multiplication with the wave number amplifies short wavelength corresponding to high wave numbers, and filtering is required. This filtering process is treated in the section below.

Gravity anomalies can also be computed from DOV using the approximate relation (9.41) into the Laplace equation relating vertical gravity gradient with east and north DOV (Rummel and Haagmans 1990).

$$\frac{\partial^2 T}{\partial^2 z} + \frac{\partial^2 T}{\partial^2 x} + \frac{\partial^2 T}{\partial^2 y} = 0 \Rightarrow \frac{\partial \Delta g}{\partial z} = -\gamma \left(\frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y} \right) \tag{9.43}$$

In this way the vertical gravity gradient can be computed using a local grid of east and north DOV.

Applying the 2D Fourier transformation to (9.43) becomes

$$\frac{\partial (F(\Delta g))}{\partial z} = -i\,\gamma\, 2\,\pi\, (k_x F(\eta) + k_y F(\xi)) \tag{9.44}$$

In Cartesian approximation Δg is harmonic too and so the formulas for upward continuation holds, which gives:

$$F(\Delta g(k, z)) = F(\Delta g(k, 0) \exp^{(-2\pi |k| z)} \tag{9.45}$$

Using (9.45) in (9.44) the relationship in the Fourier domain between the DOV and gravity anomalies is given in the form

$$F(\Delta g) = -i \frac{\gamma}{|k|} (k_x F(\eta) + k_y F(\xi)) \tag{9.46}$$

So to compute gravity using altimetric DOV, initially grids of the north and east DOV components must be constructed (Sandwell and Smith 1997; Hwang and Parsons 1995). Then these grids are Fourier transformed and then the grids are multiplied and added as given by (9.46) and the resultant grid is inverse Fourier transformed.

Using the planar approximation of the inverse Vening Meinesz formula (9.37) for the prediction of gravity using DOV and using the asymptotic representation in (9.40) for small spherical distances, Hwang (1998) demonstrated that the deterministic approach using the inverse Vening Meinesz formula also leads to (9.46), and that in the frequency domain it was equivalent to the stochastic approach of least squares collocation.

Finally a word on edge effects should be given. Before the FFT transform is applied to the residual geoid grid it is important to extend the computation region outside the data region and to apply a cosine taper to the outer parts of the grid. This is done to avoid spectral leakage caused by wavelengths that are not periodic within the area. Detailed description of this can be found in Sect. 7.3.3

All available global altimetric gravity fields have taken advantage of the FFT in their derivation in one way of the other for the computation of gravity on 1 or 2 min global grids. The global marine gravity field by Sandwell and Smith (1997, 2009) and also the NCTU gravity field by Hwang et al. (2003) applied the formulas (9.46) using DOV derived from sea surface slopes, whereas the KMS and DNSC fields (Andersen and Knudsen 1998, 2000) applied the upper formula in (9.42) to the gridded residual geoid signal derived from the altimetric sea surface heights.

9.10.2 Filtering

For satellite, altimetry, noise will always be present due to un-modelled tides, orbit errors or other contributions to residual sea surface height variability as described in Sect. 9.6. This noise can be assumed to be of white noise nature, and will be amplified in the high-pass filtering operation of predicting gravity from geoid heights (9.42).

In order to limit this effect an optimal filter was designed that both handles the assumed white noise, but also handles the power spectral density of the gravity field signal. The power spectral density of the geoid spectrum is assumed to follow a Kaula rule power law (Kaula 1966) who demonstrated, that the geoid height power spectrum decays like k^{-4} where k is the radial wavenumber.

The filtering is obtained by frequency domain least squares collocation with a Wiener filter (Nash and Jordan 1987; Forsberg and Tscherning 1997)

$$F(\Delta G) = \frac{\Phi_{N\Delta g}}{\Phi_{NN} + \Phi_{ee}} F(N) \tag{9.47}$$

where Φ is the power spectral density and e is the noise on the interpolated altimetric geoid undulations. Forsberg and Solheim (1988) confirmed that, assuming white noise signal and a Kaula rule for the spectral decay assumption, the Φ_{NN} will decay like k^{-4} and devised the following modification to (9.42)

$$F(\Delta G) = \frac{k}{1 + ck^4} F(N) = k\,\beta(k)\,F(N) \tag{9.48}$$

The parameter c is an empirical parameter which can be interpreted as a proxy of the "resolution" that can be obtained given data. The parameter is normally fine-tuned from the local variability of the gravity field and noise on the residual geoid heights (see Fig. 9.16 below).

The "resolution" is here taken as the wavelength, corresponding to the wave-number k where $\beta(k) = 0.5$ corresponding to where $\lambda = 2\pi c^{1/4}$.

9.11 Practical Computation of Global High Resolution Marine Gravity

For most practical purposes the global marine gravity fields are computed or evaluated on 1 or 2 min global grids corresponding to 3.75 km or roughly 2 km at the Equator. Altimetry does not support 2 km spatial resolution with the densest cross-track and along track spacing between observations being around 6 km. Furthermore the interpolation and the filtering applied in (9.48) suppresses wavelength shorter than roughly 10–15 km (Yale et al., 1995). The 1 min grid is generally chosen to limit the loss of information in the interpolation process. For the DNSC08 gravity field, the 1 min resolution is also chosen to ease the joint use of the suite of global DNSC08 fields (gravity, bathymetry, mean sea surface, mean dynamic topography, and prediction error) by giving all on a common global grid.

Below, the way that the KMS02 and DNSC08 global marine gravity fields were computed are presented to illustrate the various parameters choice in order for the reader to be able to understand the physical meaning of the choices as well as to assist the reader to derive their own altimetric gravity fields making their own experiments and choices.

The way the gravity field is practically computed is by patching up the Earth in a number of tiles or regions and to compute each tile or region separately. The 2° latitude by 10° longitude used for KMS02 can be seen in Fig. 9.15 below. For the derivation of the DNSC08 gravity field smaller tiles of the size of 2° latitude by 5° longitude were used. For both fields a 0.5° additional margin outside the data region was added to taper the geoid signal to zero at the boundaries in order to avoid Gibbs effects in the FFT computation.

The interpolation of scattered along-track anomalies onto a regular grid is the first step in the process. This step is crucial to the accuracy of the gravity field so much care must be taken in choosing the optimum parameters for the covariance function

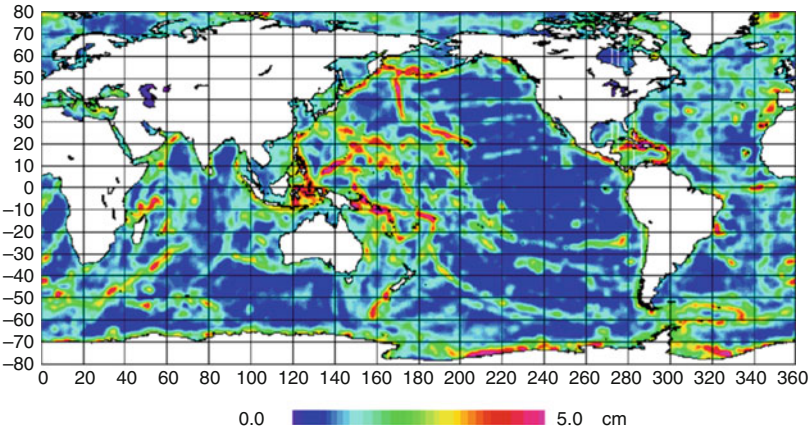


Fig. 9.13 Magnitude of residual geoid signal (Unit is cm)

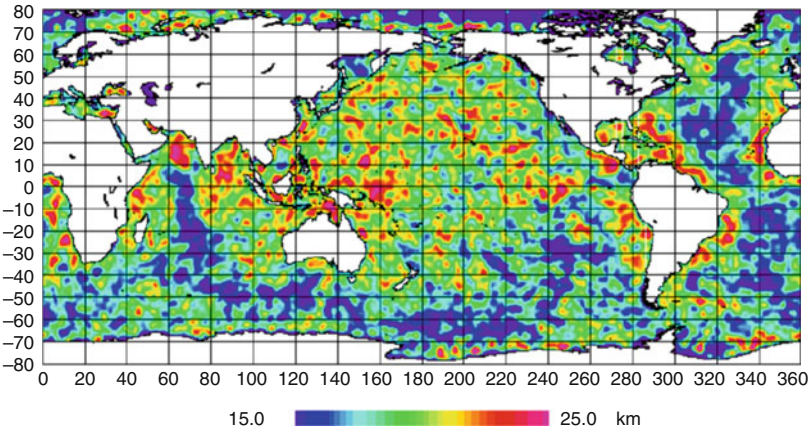


Fig. 9.14 Correlation length of the signal (α) computed as the half-width of the empirical covariance functions in 1° by 1° blocks

(9.35) used in this step. For the KMS02 gravity field, the following parameter choices were made for the signal variance (C_o) and the correlation length (α).

Figure 9.13 shows the magnitude of the residual geoid signal which was used for the computation of the signal variance (C_o). The signal and hence it variance, is seen to be largest in the tectonic active regions like the spreading and subduction zones.

The next parameter in (9.35) is the correlation length of the residual geoid signal (α). The correlation length is shown in Fig. 9.14 from a computation in 1° by 1° blocks. The correlation length largely reflects the depth of the ocean with relative small correlation length found for regions of smaller depths which are especially found along the spreading zones.

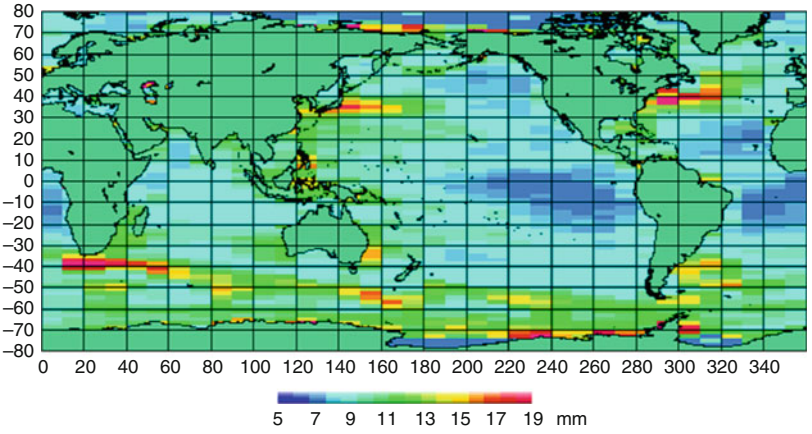


Fig. 9.15 The RMS of residual sea surface height used to determine the along track residual sea surface variance D_o parameter in (9.35) averaged over each 2° latitude by 10° longitude tiles. This clearly indicates the location of major current systems

The additional two parameter introduced into the second order Markov covariance function to model residual along-track errors are the variance (D_o) and correlation length (β) of this signal. The correlation length (β) was empirically determined to be 100 km assuming the error to be of long wavelength compared with the correlation length of the residual gravity signal (α).

In order to avoid problems with possible correlation between the quantities in (9.35), the D_o was kept fixed for the interpolation in each 2° by 10° tile. The value should reflect regions of high oceanographic noise. Hence it was approximated by a scaled version of the RMS of the sea surface height computed from 6 years of ERS-2 repeat observations and the magnitude range between $(0.5\,\text{cm})^2$ and $(4\,\text{cm})^2$. The average RMS of the sea surface variance within each tile is shown in Fig. 9.15.

The interpolated residual geoid height grids in each tile were then used to compute gravity anomalies using the multiband spherical 2D FFT technique. The conversion of geoid heights to gravity anomalies enhances shorter wavelength, and the Wiener filter described in (9.48) was applied using the filter parameter shown in Fig. 9.16.

Like several parameters for the interpolation, this parameter is strongly linked with the standard deviation of the sea surface height (see Fig. 9.5). The resolution parameters reflect the sea surface variability with high values in the major global current systems like the Gulf Stream, the Kuroshio and the Antarctic Circumpolar Currents. It should be interpreted in the way that increased filtering, thus resulting in "lower resolution" (higher c) is required in the most energetic regions to account for the increased "noise". Furthermore the presence of sea ice at latitude north and south of 70° requires increased filtering in these regions.

The DNSC08 and KMS02 were both derived in a global set of tiles (Andersen et al. 2005, 2009) but with different tile sizes and different processing parameters.

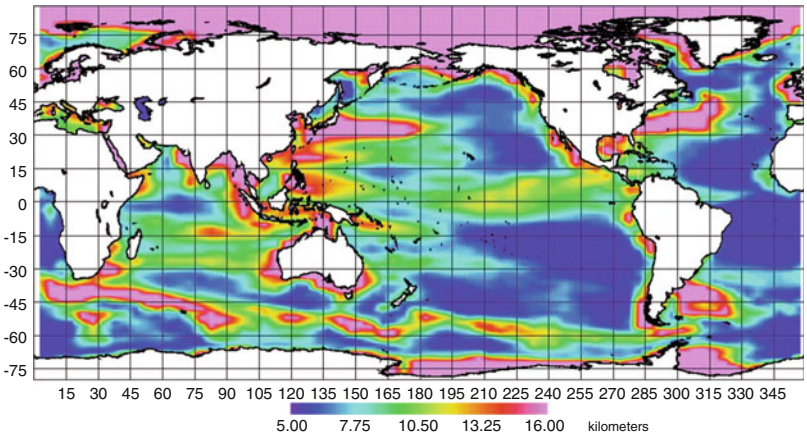


Fig. 9.16 The resolution parameter c in (9.47) used for the filtering of the gravity field (DNSC08GRA)

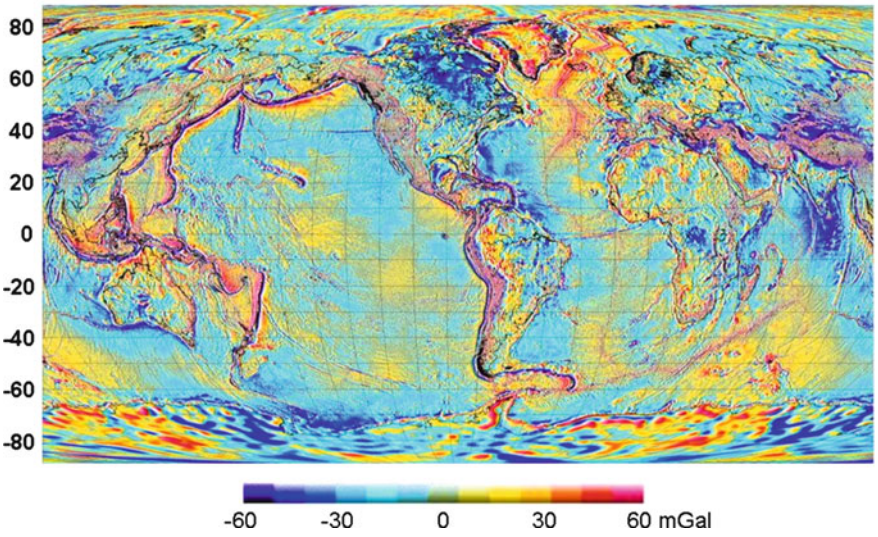


Fig. 9.17 The DNSC08 global gravity field. The altimetric gravity field over the oceans has been augmented with interpolated values from EGM2008 over land

For KMS02 the mosaic of 90 times 72 tiles were subsequently patched together, but for DNSC08 the smaller were tiled together with tapered overlay to avoid gradients along the tile-edges that could occasionally be seen in KMS02.

Finally the long wavelength gravity effect was restored using EGM96 in the case of KMS02 and EGM08 in the case of DNSC08 to give the total gravity field signal. This process also adds gravity on land. The final DNSC08 Global marine gravity field is shown in Fig. 9.17

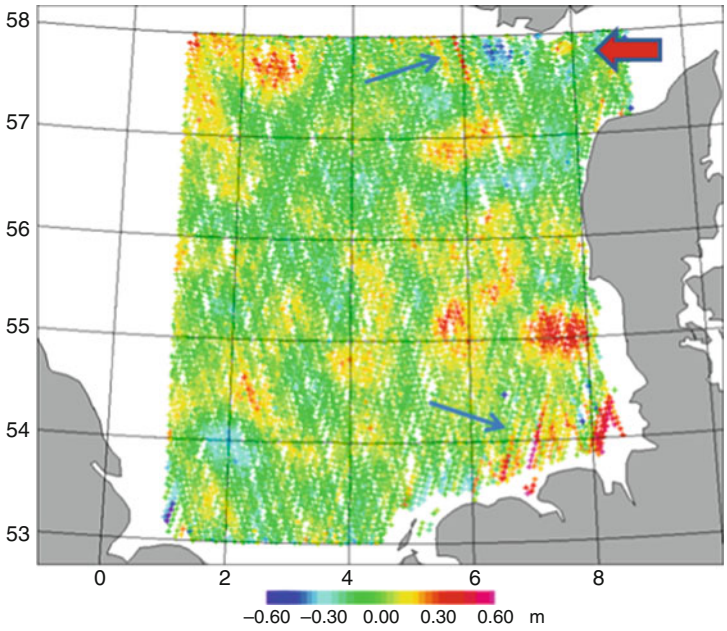


Fig. 9.18 The crossover adjusted ERS-1 geodetic mission sea surface height observations in the North Sea relative to the EGM96 geoid. The *blue arrows* indicate regions of imperfect crossover adjustment and the *red arrow* the location of a buried volcano

In this section the basic choice of parameter and their physical interpretation and fine-tuning was described for the conversion between gridded altimetric observations for the derivation of global marine gravity field. The subsequent section shows an example of the computation in one extended 5 by 9 degree tile in the North Sea with a geological interpretation.

9.11.1 North Sea Example

This section illustrate the practical steps in computing marine gravity from satellite altimetry starting using the same dataset as presented in Figs. 9.1 and 9.8. Here the process starts with residual geoid heights after the EGM96 have been removed and the data have been crossover adjusted.

The sea surface height observations representing the residual geoid height are shown in Fig. 9.18. Only ERS-1 GM data are considered in this example and only one 5° latitude by 9° longitude tile in the North Sea is considered. The standard deviation of the altimetric residual geoid heights are 4.8 cm with maximum value of 59 cm.

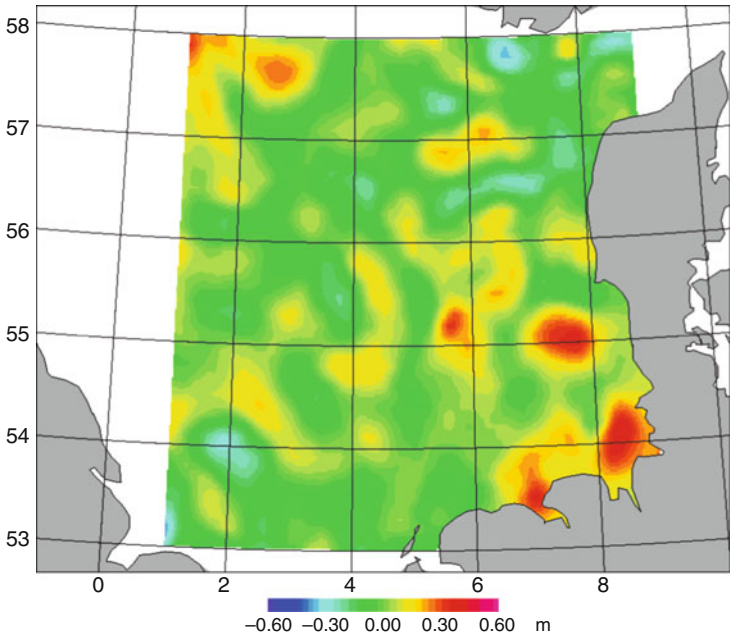


Fig. 9.19 The interpolated residual sea surface height observations

Some residual track-related errors are still visible after the crossover adjustment by the two blue arrows in Fig. 9.18 to the north towards the southern tip of Norway and to the south eastern part of the North Sea in the German Bight. Notice that the errors to the north are associated with tracks that also appear in the German Bight. These small errors can be handled using the extension to the Gauss Markov covariance function (9.34) as shown in (9.35).

In Fig. 9.18 the thick red arrow indicate a small positive signal which will be shown below to be associated with a strong gravity signal related to a buried volcano.

These values are subsequently interpolated by LSC using the modified second order Gauss Markov covariance functions formulas (9.35) with the fine-tuned parameters for signals and correlations length shown in Figs. 9.14 and 9.15. The result of the interpolated residual geoid height grid on 1 min resolution is shown in Fig. 9.19. The standard deviation of the interpolated grid is 4.1 cm with maximum value of 42 cm.

Notice that the residual along track geoid signal in the northern part of the region has been removed in the interpolated field. Also notice how the interpolation unavoidable extrapolate signal towards and onto the coast.

Subsequently the interpolated residual geoid height values in Fig. 9.19 were used to compute residual gravity anomalies using FFT applying a Wiener filter (9.48) with a choice of "resolution parameter" of 15 km taken from an inspection of

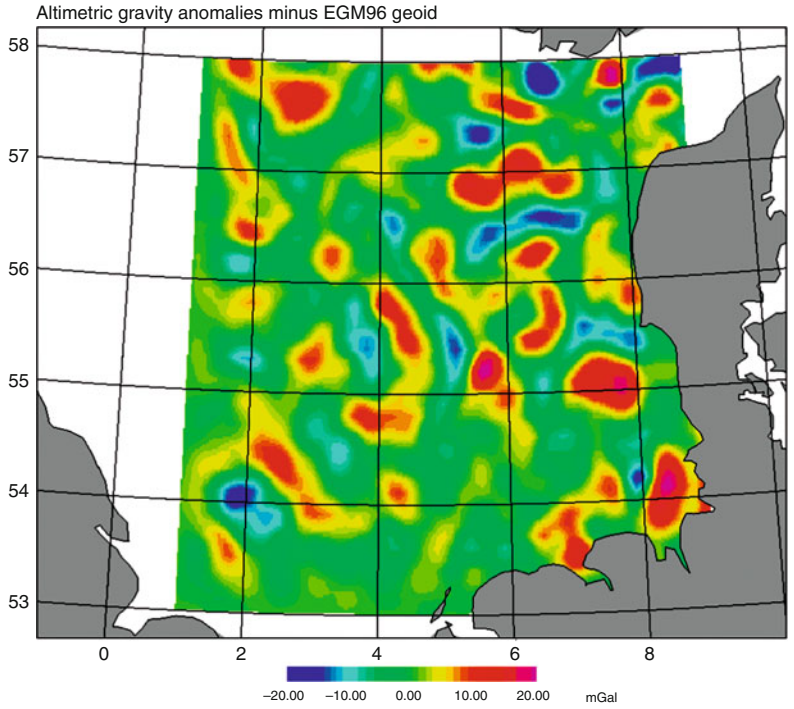


Fig. 9.20 The residual gridded gravity anomalies (relatively to EGM96)

Fig. 9.16. The residual gravity signal relative to EGM96 had a standard deviation of 5.2 mGal with a maximum value of 38 mGal on top of the buried volcano close to the southern tip of Norway marked in Fig. 9.20.

The final step in the gravity field prediction is to restore the EGM96 gravity contribution to obtain the full marine altimetric gravity field which is shown in Fig. 9.21. Now the standard deviation has been increased to 15 mGal and the maximum value is 42 mGal and the minimum value is −41 mGal. Comparison with local marine gravity observations in the region reduces from more than 8 mGal to better than 4 mGal.

The most distinct feature is a buried volcano south of Norway which is not resolved by EGM96, but clearly resolved using satellite altimetry. This peak anomaly of 42 mGal is found right at this buried volcano and the peak negative value of −41 mGal is found just to the east of this.

This free-air gravity field map also shows other distinct geological features related to the tectonics of the North Sea. One is the north-south going "Horn Graben" close to Denmark which is not resolved from EGM96. The other is the "Viking Graben" which is not very well resolved by EGM96 either.

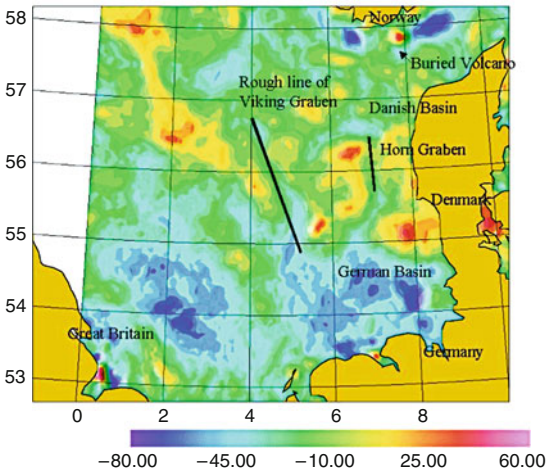


Fig. 9.21 The altimetric marine gravity field with EGM96 restored. All values are in mGal. Some major geological features of the region have been added to the map

9.12 Accuracy of Present-Day Altimetric marine Gravity Fields

Since the late 1990th several global marine gravity fields have become available on 1 or 2′ resolution for free download on the Internet: The NCTU fields (Hwang et al. 2003); the Sandwell and Smith fields – versions from 9.1 to version 18.1 (Sandwell and Smith 1997); the KMS02/DNSC08 fields (Andersen et al. 2005, 2009), and the GSFC fields (Wang 2001). During the last decade waveform retracking in one form or the other has been applied by Laxon and McAdoo (1998) who retracked altimetry in the Arctic Ocean using a robust retracker, Hwang et al. (2003) who retracked altimetry in the China Sea; Fairhead et al. (2004) who retracked/repicked data in several coastal regions, and finally the DNSC08 and Sandwell and Smith who applied retracking to the later versions of their marine gravity field (Andersen et al. 2009; Sandwell and Smith 2005, 2009).

Numerous local and global marine gravity anomalies have been created using a variety of successful techniques (e.g., Haxby 1983; Balmino et al. 1987; Sandwell 1992; Knudsen 1991; Knudsen et al. 1992; Tscherning et al. 1993; Hernandez and Schaeffer 2000; Kim 1996).

In order to illustrate the history of improvement in altimetric marine gravity field mapping over the last 10–15 years 321.400 unclassified marine gravity observations with accuracy of 2–4 mGal were provided by the National Geospatial-intelligence Agency (NGA) for the validation of altimetric gravity fields. This dataset covers the region between 25°N and 45°N and 275°E and 325°E corresponding to the region from the US east coast and out to the Mid-Atlantic spreading zone. The Gulf Stream flows northeast across the region and introduces an error of the order of 2–3 mGal because of increased sea surface height variability. Therefore, the comparison should NOT be viewed as representative for the general accuracy of global altimetric

Table 9.2 Comparison with 321.400 marine gravity field observations in the Gulf Stream region. For each of the global marine grids the standard deviation and the maximum difference are given. SS fields by Sandwell and Smith (1997, 2009); KMS02/DNSC08 by Andersen et al. (2005, 2009); EGM2008 by Pavlis et al. 2008); GSFC field by Wang (2001) and NCTU01 is by Hwang et al. (2003)

321.400 obs	Standard deviation (mGal)	Maximum difference (mGal)
KMS99	5.69	73.74
KMS02	5.05	49.38
DNSC08	3.92	36.91
EGM2008	3.94	36.90
SS V12.1	5.79	82.20
SS V16.1	4.88	45.29
SS V18.1	3.98	36.99
GSFC 00.1	6.14	89.91
NCTU01	6.10	92.10

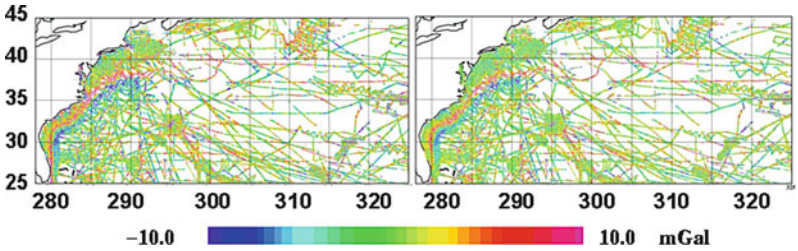


Fig. 9.22 Color coded difference between interpolated satellite altimetry gravity and 321.400 marine data in the Northwest Atlantic Ocean. The difference between marine data and the KMS02 global marine gravity field is shown in the *left panel*, and the corresponding comparison for DNSC08 is shown in the *right panel*. A closer inspection reveals that the DNSC08 is significant better in coastal regions

gravity fields, but more as an illustration of the general improvement in gravity field modeling during the last decade. Actually, the Gulf Stream region is one of the regions where altimetry performs the worst compared with marine gravity observations and where most smoothing has to be applied as shown in Figure 9.16.

A detailed comparison with this dataset is presented in Table 9.2 and the point by point difference between measured and interpolated gravity field values in the region is shown in Fig. 9.22. A total of nine global gravity fields released during the last decade have been tested. The oldest fields are the KMS99 field (1999), followed by the GSCF 0.1 (2000), the NCTU 01 (2001) and SS V12.1 (2001) and KMS02 (2002). All of these have standard deviation with the 321,400 gravity observations higher than 5 mGal.

A steady improvement in the accuracy of altimetric marine gravity field has been observed during the last decade. With the release of EGM2008 and the global gravity field (DNSC08 and SS V18.1) a consistent comparison below the 4 mGal level has been achieved. In terms of improvement this corresponds to more than 20% improvement in standard deviation compared with global marine gravity fields

7–10 years ago. One should notice that part of the 321.400 marine gravity field observations have been used for the EGM2008 geopotential model as 5 min mean anomalies.

The detailed comparison in Fig. 9.22 between individual marine gravity observations and interpolated gravity from KMS02 (left panel) and the DNSC08GRA (right panel) initially looks identical. However a close inspection of particularly the coastal regions indicates that DNSC08 is substantially better than KMS02 which is the effect of retracking and improved ocean tide modelling. Both maps show a red/blue anomaly pattern which closely follows the Gulf Stream. This could indicate that the correction for mean dynamic topography (ζ_{MDT}) using the PGM2007A mean dynamic topography model complete to degree and order 50 (Andersen and Knudsen 2009; Pavlis et al. 2007b) and used for the derivation of EGM2008 does not have adequate resolution, and that future corrections for mean dynamic topography should remove even higher degree and order of the signal.

9.13 Integrating Marine, Airborne and Satellite Derived Gravity

Marine gravity field are available from various different sources, like gravimeters onboard marine vessels (e.g., ships and submarines), onboard aircrafts, manually operated in the field, and finally from satellite altimetric measurements. These different data sources should not be considered as competitors of gravity information but rather a great opportunity to have complimentary sources of gravity information and the only way to create a truly global gravity field including the Polar Regions.

Airborne gravimetry is a fast and economic method for local to regional scale gravity mapping. Some of the biggest advantages are the uniform and seamless coverage of land and sea, and the ability to cover remote and otherwise inaccessible areas. The bias free property of airborne gravity data obtained by spring type gravimeters is also an important point for geodetic applications; see Childers et al. (2001) and Olesen et al. (2002). Ship borne gravity measurements are still one of the most accurate sources of gravity at sea, but the cost is large and furthermore the ship needs a minimum water depth in order to be feasible.

Airborne and marine gravimeters observe the gravity directly, and can be used to determine (any) offset, which might be present in the altimetric gravity field. The three set of data are shown in Fig. 9.23 for the test area on the west coast of Greenland around the Disko Bay.

The gravity field derived from altimetric residual geoid height observations h can be merged with airborne and/or marine gravity observations $\Delta g'$ using Least Squares Collocation. The expression for gravity and a-posteriori variance $\sigma^2_{\Delta g}$ on the predicted gravity anomalies Δg are

$$\Delta g = \left(C_{\Delta g h} C_{\Delta g \Delta g}\right) \begin{pmatrix} C_{hh} + D_h & C_{h\Delta g'} \\ C_{\Delta g' h} & C_{\Delta g' \Delta g'} + D_{\Delta g'} \end{pmatrix}^{-1} \begin{pmatrix} h \\ \Delta g' \end{pmatrix} \tag{9.49}$$

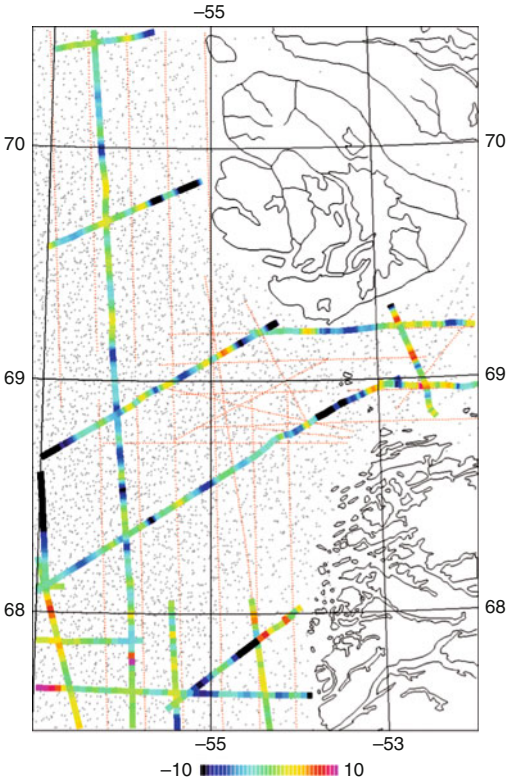


Fig. 9.23 Test area on the west coast of Greenland: The distribution of satellite altimetry (*gray dots*) and airborne gravity (*aligned black dots*) together with the difference between marine gravity and collocation results based on sea surface heights observations and airborne gravity

and

$$\sigma^2_{\Delta g} = C_{\Delta g \Delta g} - \left(C_{\Delta g h} C_{\Delta g \Delta g}\right) \begin{pmatrix} C_{hh} + D_h & C_{h\Delta g'} \\ C_{\Delta g' h} & C_{\Delta g' \Delta g'} + D_{\Delta g'} \end{pmatrix}^{-1} \begin{pmatrix} C^T_{\Delta g h} \\ C^T_{\Delta g \Delta g} \end{pmatrix} \tag{9.50}$$

$C_{hh}, C_{\Delta g h}, C_{\Delta g \Delta g}$ are the covariance matrices between height-height, gravity-height, gravity-gravity. The covariance matrices D_h and $D_{\Delta g'}$ contain the noise variance of the geoid height and gravity observations, respectively. Gravity anomalies with hyphen like $\Delta g'$ are the observed gravity from altimetry and/or ship. Δg are predicted gravity anomalies.

9.13.1 East Greenland Airborne and Altimetric Gravity Example

The Disko Bay (Illulisat fjord) coastal region located on Greenland's west coast around latitude of 69°N and longitude 55°W is used as test region, as it has good coverage of altimetric, marine and airborne observations as seen in Fig. 9.23. This

Table 9.3 Comparisons with marine gravity data (in mGal). Both the altimetric and airborne gravity field has been interpolated onto the location of the marine observations. Direct comparisons between co-located airborne observations and marine observations compares better than 2 mGal (Olesen et al. 2002)

Input data	Mean	Std. dev.	Abs. max.
Airborne gravity	−0.5	6.9	26
Satellite altimetry	−9.5	5.4	24
Satellite altimetry+airborne gravity	−0.7	3.6	18

area has seasonal ice cover and ice drift. A covariance function based on airborne gravity residuals has been estimated and an analytic expression has been determined (Knudsen 1987a).

For airborne gravity, an error model which takes into account the correlated noise is used in this study. However, the effect of incorporating this feature was found to be insignificant. This implies that even though the airborne data are filtered along track, they may be considered as point values for our use. Predicted gravity anomalies, as well as their associated error estimates, are finally derived from the normal equation solution. More information about the study can be found in Olesen (2003).

The result in Table 9.3 shows a very big improvement with the marine observations with the agreement being improved from 6.9 to 3.6 mGal. Similarly the bias of –9.5 mGal between marine and altimetric gravity is reduced to –0.7 mGal by the combined use of altimetry and airborne gravity. This demonstrates the potential for improving coastal marine gravity field by merging different types of observations.

9.14 Altimetric Gravity Research Frontiers

The previous sections have shown that the global altimetric gravity fields are generally very accurate in the open ocean, but in coastal and Polar Regions the error increases and this is naturally a focus area for future research. Gravity recovery is particularly difficult in these regions, but on here the largest improvement can still be gained from a dedicated effort in improving the accuracy of the sea surface height observations.

The problems in shallow water and Polar Regions are due to several factors: The waveform shape of the returned radar pulse will only infrequently follow a Brown model and hence data are frequently rejected by the automatic retracking by the space agencies. The presence of a coast will also distort the part of the illuminated region by the altimeter or the radiometer used to determine the range or range corrections. Sea state is also changing close to the coast and particularly the spatial extent of the tidal signal is scaled down creating very complex tidal patterns which furthermore include resonance and overtones. For the coastal regions, the use of spectral methods like the Fourier Transform will also be problematic even though the removal of the highly accurate EGM2008 model ensures that the recovered signal is not so much distorted by the presence of land.

For the next generation of global altimetric gravity fields dedicated effort into research in the following areas will be needed it for further gravity field improvement:

- Inclusion of new data types (ICESat, Cryosat-2, Sentinel-3)
- Improving the altimeter range corrections
- Improving the ocean tide correction
- Altimeter waveform re-tracking.

In the following an introduction into the problems and importance of these effects on gravity field determination will be presented with examples from ongoing research. Large part of the investigation relates directly to improving the accuracy of the sea surface height observations and hence lowering the error e on the altimetric observations (9.19)

9.14.1 ICESat and Cryosat-2

ICESat laser altimetry is a relative new and complementary data source to conventional radar altimetry (Zwally et al. 2002). The important aspect of ICESat is the fact that it has an inclination of 86° which brings it 400 km closer to the Pole than the ERS and Envisat satellites. In principle laser data can be processed and used very much like radar altimetry. For the DNSC08GRA these data were used in the partly ice-covered parts of the Arctic Ocean (between 70°N–86°N, 100°E–270°E) and at latitudes above 80°N in all of the Arctic Ocean in order to extend the MSS and gravity field towards the North Pole. Only a few months of the ICESat data were available for DNSC08 and since the termination of the mission in 2010 a total of around 19 month was recorded. One further advantage of ICESAT is its much smaller footprint compared with radar altimetry which means that it can in principle resolve shorter wavelength of the gravity spectrum. The footprint of the laser has a radius of roughly 70 m observing at each 120 m along track where the radius of conventional radar altimeter (ERS and Jason type) has a radius of 5–10 km depending on the sea state.

Cryosat-2 was successfully launched in 2010. To meet the challenges of measuring ice-sheet changes, Cryosat-2 carry a sophisticated radar altimeter called SIRAL (Synthetic Aperture Interferometric Radar Altimeter). It is capable of carrying out Delay-Doppler Processing in one direction during flight which means, that the resolution compared with conventional altimetry is increased by a factor of 20 to around 300 m. However over most of the oceans Cryosat-2 will operate as a conventional altimeter. The accuracy of Cryosat-2 will be well below the 1-cm level (Raney 2009) making it significantly better than conventional satellite altimeters as seen in Table 9.1. This means that besides being useful for the determination of the thickness of the ice, the Cryosat-2 can be used to recover gravity anomalies over the ocean with unprecedented accuracy compared with conventional satellite altimetry.

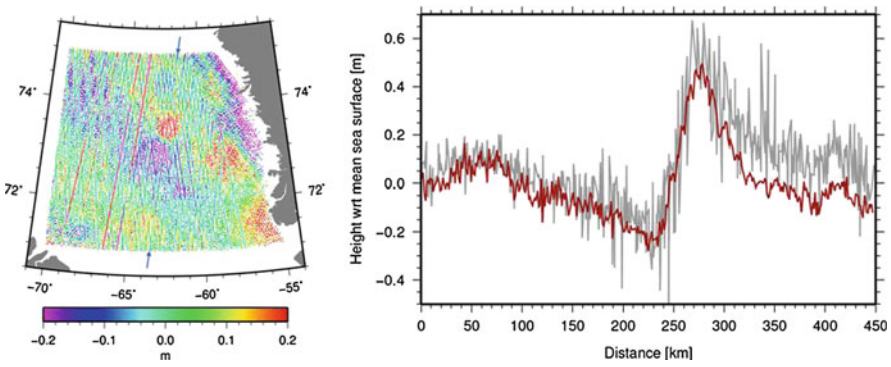


Fig. 9.24 Residual sea surface height observations in the Baffin Bay from ERS-1 and the first 3 month of SAR processed Cryosat-2 data. The profile marked with an *arrow* in the *left figure* is shown in the *right figure* with distance from the northern point

The first 3 month of Cryosat-2 SAR retracked residual sea surface height data in the Baffin Bay is shown in Fig. 9.24 overlayed the residual sea surface height from ERS-1 used for the prediction of DNSC08. Similar to the processing of ERS-1 (see Sect. 9.4–9.6) wavelength longer than 200 km have been removed from Cryosat-2 but no crossover adjustment were preformed.

One SAR 5 Hz profile is marked in the left Figure with a blue arrow. The figure to the right shows the Cryosat-2 residual sea surface height (relative to the DNSC08 Mean sea surface and not to the geoid) in red and ERS-1 observations (grey) within 5 km across-track from the SAR profile. Dramatic improvement in accuracy of the Cryosat-2 data is clearly visible compared with the older ERS-1 satellite data.

Cryoat-2 will furthermore improve the mapping of the Arctic Ocean as it has an inclination of 92° bringing it 200 km from the North Pole and for coastal regions the footprint of some 300 m of the Delay-Doppler signal will enable gravity field mapping much closer to the coast.

9.14.2 Altimeter Range Corrections

The determination of sea surface height close to the coast degrades due to the fact that several range corrections degrade as the altimeter approaches the coast. The radiometer used to correct the altimeter for the wet troposphere, has a much larger footprint than the altimeter and particularly the wet troposphere correction is affected. Although much smaller than the dry tropospheric range correction in magnitude, the wet troposphere correction is far more complex showing rapid variations in both time and space and therefore also needs careful attention in the coastal region. The correction can vary from just a few millimeters in dry, cold air to more than 30 cm in hot, wet air.

The footprint of the radiometer is dependent on the height of the spacecraft and the scanning frequency of the radiometer, but typical values of the footprint

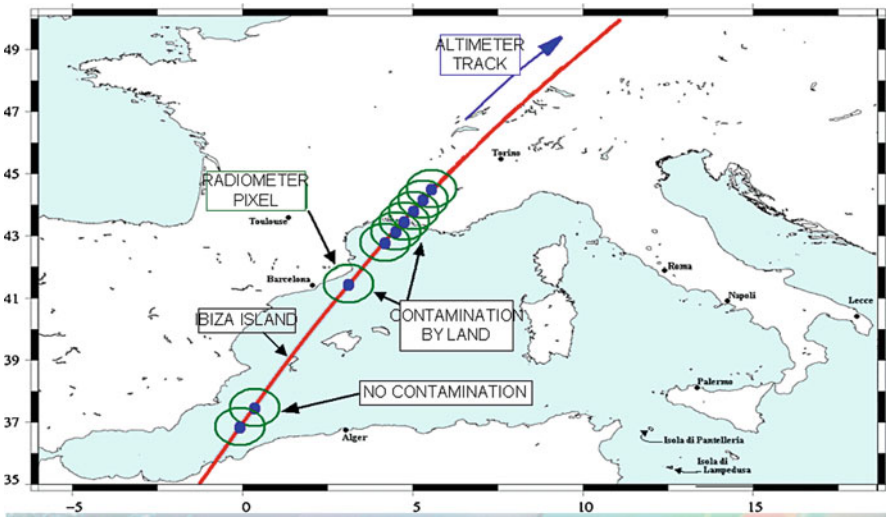


Fig. 9.25 An example of a Jason-1 track crossing the western Mediterranean Sea. *Blue dots* indicate the footprint of the altimeter and the *green circles* shows the size of the main radiometer beam (Figure from Eymard and Obligis 2006)

of the main beam ranges between 20 and 30 km. This is considerably larger than the 4–10 km footprint of the altimeter as illustrated in Fig. 9.25 for a pass across the Western Mediterranean Sea. Consequently, the radiometer is contaminated by the presence of land much earlier than the altimeter, as the spacecraft approaches and coast and generally the main beam is affected up to 30 km from the coast. The wet troposphere correction derived from the on-board radiometer is similarly affected, and currently intensive research is performed to improve the wet troposphere correction in coastal regions (e.g., Eymard and Obligis 2006)

The analysis by Andersen and Scharroo (2011) showed that the accuracy of the wet troposphere correction degrades from around 1.1 cm in the open ocean to roughly 2 cm around 30 km from the coast.

9.14.3 Ocean Tides

The largest contributor to sea surface height error in shallow water is unquestionably due to errors in present day ocean tide models (Andersen and Scharroo 2011). Even though the determination of the ocean tides have dramatically improved since the launch of TOPEX/Poseidon and most recent investigations indicate that global models are now accurate to around 1–2 cm in the global ocean (Andersen 1995; Shum et al. 1997), there are still problems close to the coast due to the fact that the tidal signal is scaled down and becoming increasingly complex

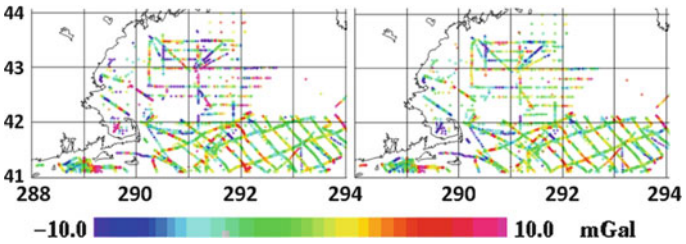


Fig. 9.26 The difference between marine observations and altimetric gravity in parts of the Gulf of Maine are colored. *Left* shows the differences for KMS99 and *right* the differences for KMS02. The color scale ranges +/-10 mGal, and the major difference between the two fields is explained by the use of GOT 00.2 compared with FES94 for the KMS99 gravity field to the left

with the presence of overtides in shallow water regions (Andersen 1999; Andersen et al. 2006)

Figure 9.26 shows the difference in gravity field mapping for the Gulf of Maine using two different ocean tide models. The plot to the left is a comparison between marine gravity and interpolated gravity using KMS99 which used FES94 in its derivation. The figure to the right represents the differences between marine gravity and interpolated gravity from KMS02, which used the GOT00.2 ocean tide model (Ray 2001). The largest improvement are clearly seen north of 42°N, which is the location of the shelf break, which indicate the significant improvements from the use of GOT 00.2 ocean tide model. Since this investigation was performed, ocean tide modelling has improved even further with the release of new ocean tide models called GOT 4.7.

In the deep ocean, recent investigations showed that ocean tide has a height accuracy of around 1.4 cm (Bosch 2008). However, global ocean tide models still have errors exceeding 10–20 cm close to the coast as also demonstrated by Ray (2006). Such signal can easily generate 5–10 mGal gravity error very close to the coast. So improved coastal ocean tide modeling is still one of the keys to improved altimetric gravity field recovery in shallow water regions in the future.

9.14.4 Retracking in Coastal and Polar Regions

As the satellite approaches the coast the characteristics of the sea surface changes, and it is important to retrack the existing GM data using more tolerant methods in order to increase the amount of data available to derive altimetric gravity. Similarly, it is important to retrack satellites to increase the accuracy of the sea surface height observations. This process involves two runs of retracking – a so called double retracking – where the first retracking run is performed to increase the number of observations, whereas the second run is performed to increase the accuracy of the sea surface height retrieval also demonstrated in Fig. 9.10.

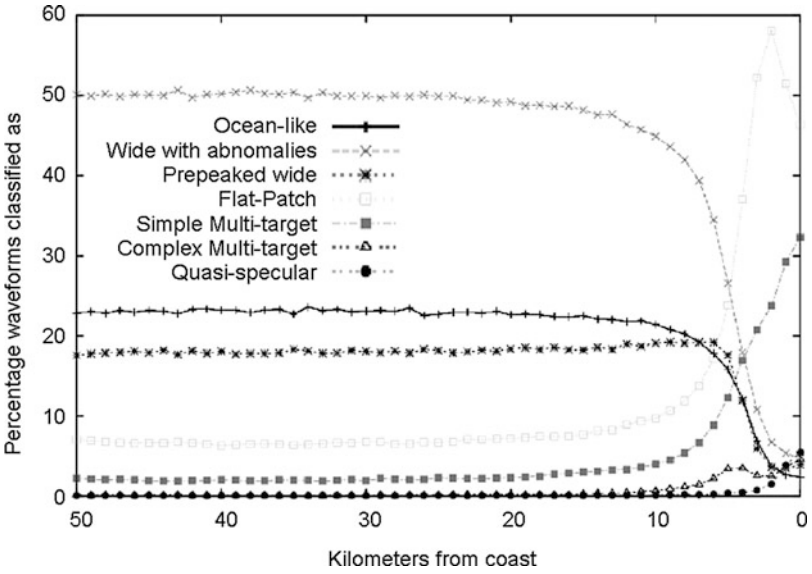


Fig. 9.27 Waveform shape distribution from 18 Hz un-averaged observations in global coastal zone (excluding sea-ice) from the ERS1 GM as a function of the distance to the coast. Detailed description of waveform characteristics can be found in Dowson and Berry (2006)

The Geosat GM does not benefit much from retracking as it was very carefully investigated and retracked originally by the US Navy. The data was recently recompiled from various archives, reprocessed and retracked at NOAA, who kindly provided the dataset to the scientific community (Lillibridge et al. 2004).

Due to special properties and the high inclination of the ERS-1 GM mission, the data from this satellite clearly gains most from retracking. With the Arctic Ocean being mostly permanently ice-covered, and the ERS-1 satellite covering up to the 82 parallel, retracking is the only way of obtaining altimetric gravity data at high latitudes where very few of these data resemble open ocean Brown waveforms.

Another benefit of tolerant retracking of the ERS-1 data is the fact that the waveform changes rapidly in complexity as the altimeter approached the coast. Numerous different echo shapes appear in the coastal zones caused by a variety of surface effects including land contamination of the echo, off-ranging to inland water, and the presence of unusually calm water in sheltered areas. For a detailed description of different waveforms see Dowson and Berry (2006).

Even though coastal zone echoes are complex and rapidly changing, the waveforms can be successfully retracked. Figure 9.27 illustrate that within 10 km of the coast, a rapid increase in non-Brown model waveforms is seen; within 5 km of the coast the majority of the echoes are non Brown model shaped. For the derivation of the DNSC08 gravity field, the Earth and Planetary Science Lab (EAPRS) expert system (Berry et al. 2005) was adapted to retrack ten complex waveform shapes of the ERS-1 GM waveforms corresponding to ice, inland water and land. In order to

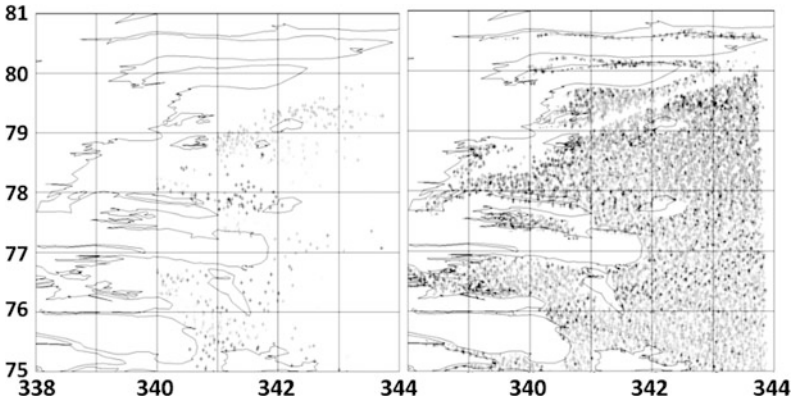


Fig. 9.28 Altimetric height observations in the ice-covered regions east of Greenland. The *upper figure* shows the number of data points that can be retrieved using standard ESA retracked 1 Hz data. The *lower figure* shows the amount of 1 Hz data that can be retrieved using more tolerant retrackers

include ocean waveform retracking the Southampton Ocean Center ocean retracker (Challenor and Srokosz 1989) was added to the system.

Due to the presence of sea-ice in Polar Regions, these will be the regions where retracking using multiple tolerant retrackers will provide the most new data and the most significant improvements to gravity field determination. The region east of Greenland (75°N << 80°N, 320°E << 350°E) is well known for the presence of sea ice. Here the number of ERS-1 data points that can be retrieved from retracking is increased from 750 data points (un-retracked) to 22,200 data points (retracked) using the more tolerant retrackers as seen in Fig. 9.28. Even data in the narrow fjords are recovered.

This vast improvement in data carries forward into an improvement of the derived gravity field. This can be seen from a comparison with 900 airborne gravity data from the Greenland/Svalbard KMS9803 survey bounded by 77°N–80°N, 30°W–5°E. The accuracy of these airborne measurements is better than 2 mGal (Olesen 2003).

The results of the comparison with 900 airborne gravity observations are shown in Table 9.4 for six different gravity fields; the KMS02, Laxon and McAdoo (version 97), ArcGP (version (01–06), SS v. 16.1 and v. 18.1 and DNSC08GRA. The Laxon and McAdoo polar gravity field (version 97) was developed using an early attempt with tolerant retracking of the ERS data (Laxon and McAdoo 1998). The ArcGP gravity field is derived from a combination of data from different sources such as marine, airborne, altimetry etc. (Kenyon and Forsberg 2008). For KMS02 the lack of retracked altimetry over the ice means that this field is not good at all. The Laxon and McAdoo gravity field from retracked ERS data is significantly better, and the ArcGP compilation of data performs even better. The DNSC08GRA is partly based on the ArcGP data, as ArcGP is part of data dataset used to derive

Table 9.4 Comparison with 900 airborne gravity observations from the KMS9803 airborne survey. The standard deviation and maximum difference between the airborne observation and various gravity fields are given

900 points	Std (mGal)	Max (mGal)
KMS02	9.4	51.2
Laxon and McAdoo(97)	7.2	46.2
ArcGP (01–06)	5.8	34.4
SS 16.1/18.1	8.2/5.9	44.9/37.4
DSNC08	4.1	24.0

the EGM2008 geoid and gravity field. The huge amount of new data that can be retrieved using suite of tolerant retrackers and particularly the sea-ice designed retracker (Berry et al. 2005) brings the standard deviation of the comparison for DNSC08GRA all the way down to 4.1 mGal. In terms of variance reduction this is nearly a 6-times improvement over KMS02.

Appendix A Data Resources

A.1 Altimetry Data

Some of the major distributors of satellite altimetry are the following:
Radar Altimetry Database system (RADS)
http://rads.tudelft.nl
Archiving Validation, interpretation of satellite data (AVISO)
www.aviso.oceanobs.com/en/altimetry/index.html
National Ocean and Atmosphere Administration (NOAA)
http://ibis.grdl.noaa.gov/SAT/ocean_links.html
Jet Propulsion Lab (JPL-PODAAC)
http://podaac.jpl.nasa.gov/DATA_CATALOG/index.html
International Altimeter Service (IAS):
http://ias.dgfi.badw.de/IAS

A.2 Altimetric Gravity Field Resources

DTU Space (DNSC, DTU gravity field models)
http://space.dtu.dk (data and models)
University of California, San Diego (Sandwell and Smith gravity field models)
http://topex.ucsd.edu/marine_grav/mar_grav.html

NCTU National Chaotung University (Taiwan)
: The NCTU1 global marine gravity field model is available on request from Cheinway Hwang at hwang@geodesy.cv.nctu.edu.tw

Arctic Gravity Field Project (ArcGP)
: Arctic gravity field grid
http://earth-info.nga.mil/GandG/wgs84/agp/readme_new.html

Chapter 10
Geoid Determination by FFT Techniques

Michael G. Sideris

10.1 Outline of the Chapter

This chapter introduces Fourier-based methods, and in particular the fast Fourier transform (FFT), as a tool for the efficient evaluation of the convolution integrals involved in geoid determination. An attempt was made to make this document as self-contained as possible for the benefit of readers inexperienced in spectral methods. Therefore, the Fourier transform and its properties are presented in the appendix following the chapter (Appendix A), and reference is made to the particular formulas and properties employed in geoid determination. Readers familiar with the Fourier transform theory can skip Appendix A and concentrate on Chap. 10, which discusses its application for efficient determination of the geoid. Although an extensive, but definitely not exhaustive, list of references containing more details, applications and numerical results is provided, it is hoped that the reader will be able to find herein (in Appendix A) the fundamental Fourier transform theory necessary for understanding the developments presented in the following pages.

The chapter begins with a quick review of the Stokes boundary value problem and its solution by the remove-restore technique, using Hermert's second condensation method for the terrain reduction. It then shows how error propagation can be accomplished with FFT methods, and discusses the input–output system theory that uses gridded heterogeneous noisy data in the frequency domain. The similarities and differences with the space-domain least-squares collocation method are pointed out. The chapter also shows the FFT evaluation of other convolution integrals useful in gravity field approximation.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_10,

10.2 Review of Stokes's Integral and Its Evaluation

10.2.1 Stokes's Boundary Value Problem

As discussed in Chap. 3, Stokes's boundary value problem (BVP) is the gravimetric determination of the geoid S. Stokes's problem deals with the determination of a potential, harmonic outside the masses, from gravity anomalies Δg given everywhere on the geoidal surface; see (3.94). Consequently, since no masses are allowed outside S, the topography of the Earth must be elliminated mathematically. We will come back to this point later; for now, we assume that S encloses all masses.

The classical BVP is to determine the disturbing potential T, which satisfies Laplace's equation

$$\Delta T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0, \tag{10.1}$$

under the a boundary condition on S, Which, in spherical approximation, i.e., neglecting relative errors of the order of flattening of the reference ellipsoid (Moritz 1980), is

$$\frac{\partial T}{\partial r} + \frac{2}{r} T + \Delta g = 0. \tag{10.2}$$

The solution of (10.1) under the condition of (10.2) provides T as a function of the gravity anomalies Δg on the geoid, and is given by Stokes's integral of (3.98) or (3.101), which is written here in a form that lends itself to FFT determination as follows:

$$T = \frac{R}{4\pi} \iint_S \Delta g S(\psi) d\sigma = \mathbf{S}(\Delta g), \tag{10.3}$$

where R is the mean radius of the Earth and $\mathbf{S}$ denotes the Stokes integral operator. S(ψ) is Stokes's function of (3.100), which is rewritten here as follows:

$$S(\psi) = \frac{1}{\sin(\psi/2)} - 6\sin\frac{\psi}{2} + 1 - 5\cos\psi - 3\cos\psi \ln\left(\sin\frac{\psi}{2} + \sin^2\frac{\psi}{2}\right), \tag{10.4}$$

$$sin^2\frac{\psi}{2} = sin^2\frac{\varphi_P - \varphi}{2} + sin^2\frac{\lambda_P - \lambda}{2} cos\ \varphi_P\ cos\,\varphi. \tag{10.5}$$

Recall that ψ is the spherical distance between the data pont (φ, λ) and the computation point (φ_P, λ_P). Note that, in contrast to Chap. 3, in this chapter we will use subscript P to denote the computational point and no subscript for the running point; this will simplify the notation in the FFT-based formulas that will be developed later on.

The geoid undulation at point P is then obtained by applying Bruns's equation (2.36):

$$N = \frac{T}{\gamma} = \frac{R}{4\pi\gamma} \iint_S \Delta g S(\psi) d\sigma = \frac{1}{\gamma}\mathbf{S}(\Delta g). \tag{10.6}$$

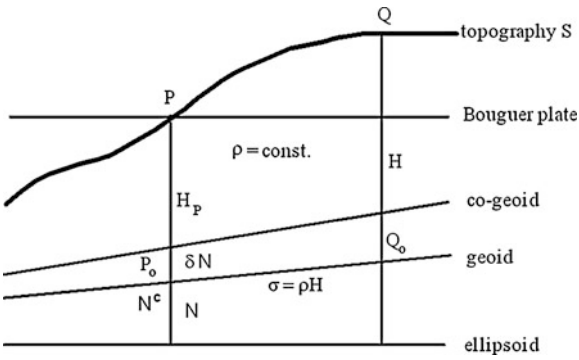


Fig. 10.1 Actual and condensed topography, in planar approximation (After Sideris 1990)

10.2.2 Geoid Undulations and Terrain Reductions

Equation (10.6) gives the undulation of the geoid N provided that there are no masses outside the geoidal surface. One way to take care of the topographic masses of density ρ – usually assumed constant – is Helmert's condensation reduction (see also Chap. 3, Sect. 3.5), which is used here as a representative from a number of possible terrain reductions, applied as follows:

(a) Remove all masses above the geoid;
(b) Lower station from P to P_o (see Fig. 10.1) using the free-air reduction F; and
(c) Restore masses condensed on a layer on the geoid with density $\sigma = \rho$H.

This procedure gives Δg on the geoid computed from the expression

$$\Delta g = \Delta g_P - A_P + F + A^c_{P_o} = \Delta g_P + F + \delta A. \tag{10.7}$$

$(\Delta g_P + F)$ is the free-air gravity anomaly at P, A_P is the attraction of the topography above the geoid at P, and $A^c_{P_o}$ is the attraction of the condensed topography at P_o.

It must be mentioned here that eq. (10.7) holds only if gravity anomalies are linearly dependent on heights. In the general case of the "Helmertized" Stokes BVP solution, A^c must be also computed at P, and the gravity anomalies must first be reduced by adding $\delta A_P = -(A_P - A^c_P) \neq c_P$ and then be downward-continued to the geoid (thus not resulting in simple Faye anomalies); details can be found in Martinec et al. (1993). Nevertheless, for the sake of simplicity, we are continuing here by accepting the approximation involved in eq. (10.7).

Obviously, the attraction change δA is not the only change associated with this reduction. Due to the shifting of masses, the potential changes as well by an amount called the *indirect effect on the potential*, given by the following equation:

$$\delta T = T_{P_o} - T^c_{P_o}, \tag{10.8}$$

where T_{P_o} is the potential of the topographic masses at P_o and $T^c_{P_o}$ is the potential of the condensed masses at P_o. Due to this potential change, the use of (10.6) with Δg

from (10.7) produces not the geoid but a surface called the *co-geoid*. Thus, before applying Stokes's equation, the gravity anomalies must be transformed from the geoid to the co-geoid by applying a small correction $\delta\Delta$g called the *indirect effect on gravity*:

$$\delta\Delta g = -\frac{1}{\gamma}\frac{\partial\gamma}{\partial h}\delta T. \tag{10.9}$$

The final expression giving N can now be written as

$$N = \frac{1}{\gamma}\mathbf{S}(\Delta g + \delta A + \delta\Delta g) + \frac{1}{\gamma}\delta T = N^c + \delta N, \tag{10.10}$$

where N^c is the co-geoidal height and δN is the indirect effect on the geoid; see Fig. 10.1.

In planar approximation, δT and δA can be expressed using the vertical derivative operator $\mathbf{L} = \partial/\partial$z (Sideris 1990), which can also be expressed by the surface integral (Moritz 1980; Sideris 1987a)

$$\mathbf{L}f = \frac{\partial f}{\partial z} = \frac{1}{2\pi}\iint \frac{f - f_P}{l^3}dxdy, \qquad \mathbf{L}^n = \frac{\partial^n f}{\partial z^n} \tag{10.11}$$

which is the planar approximation of (3.20). By definition, $\mathbf{L}$ annihilates any function f that is constant on the plane. The potential change is

$$\delta T = -\pi G\rho H_P^2 - 2\pi G\rho\sum_{r=1}^{\infty}\frac{1}{(2r+1)!}L^{2r-1}H^{2r+1}, \tag{10.12}$$

and the attraction change is equal to the classical terrain correction c:

$$\delta A = c = 2\pi G\rho\sum_{r=1}^{\infty}\frac{1}{(2r)!}L^{2r-1}(H - H_P)^{2r}, \tag{10.13}$$

where G denotes Newton's gravitational constant of (1.2). These series expansions are valid for low slopes of the terrain, namely for $(H - H_P)/l \leq 1$. It is important to remember that the attraction of the condensed topography in (10.7) must be computed on the geoidal surface in order for the reduced gravity to refer to the geoid (actually, the co-geoid) and be used as input to Stokes's formula. Also, if ρ is not constant, then it has to be included under the integral in the $\mathbf{L}$ operator. For more discussion, see Wichiencharoen (1982), Wang and Rapp (1990), and Sideris (1990).

Expressions like (10.11)–(10.13) have generallly only a formal meaning, despite their extensive use in geodesy. This is because $\mathbf{L}$ is an unbounded singular integral operator and, therefore, the application of its powers of any order is feasible only under the hypothesis of extreme smoothness of the function f. Nevertheless, in practice we apply $\mathbf{L}$ only in a discretized form, assuming implicitly that the spectrum of f is zero above a certain frequency (or, more precisely, that it is aliased into lower frequencies). Under these assumptions, f is an analytic function for which (10.11)–(10.13) become meaningful and thus their implementation is justified.

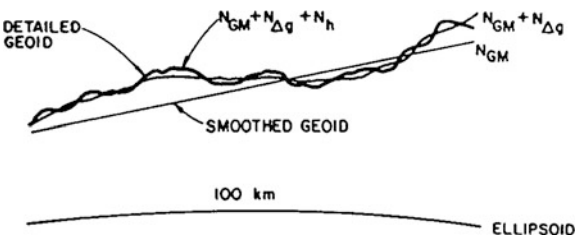


Fig. 10.2 Contributions of different data to regional geoid determination (After Schwarz et al. 1987)

10.2.3 Practical Evaluation of Stokes's Integral

The remove-restore technique. The use of (10.6) requires gravity anomalies all over the geoid for the computation of a single geoid undulation. Obviously, this is impractical and thus, in practice, some modifications of the technique are necessary. Firstly, we can only apply (10.6) in a limited region. Then, the long wavelenth contributions of the gravity field will not be present in the results and must be computed in another way. They are provided by a set of spherical harmonic coefficients (geopotential model). Secondly, the integral is discretized and is computed as a summation using discrete data. Due to the limited density of the gravity data, the short wavelengths will not be present (aliased). They are computed by using topographic heights, which are usually given in the form of a Digital Terrain Model (DTM). These frequency contributions are shown in Fig. 10.2.

Utilizing the remove-restore concept (see Chap. 2, Sect. 2.5), the computation of geoid undulations N by combining a geopotential model (GM), mean free-air gravity anomalies Δg_{FA}, and heights H in a DTM is based on the following formula:

$$N = N^{GM} + N^{\Delta g} + N^{H}, \qquad \Delta g = \Delta g^{FA} - \Delta g^{GM} - \Delta g^{H}. \tag{10.14}$$

Although geoid undulations are more sensitive to the low to medium frequencies of the field, in rough topography all three data sets are necessary for estimating N. Note that the gravity anomalies used in Stokes's equation have the contributions of the topography and the GM removed. Thus, the remove (pre-processing) stage involves the computation and removal of the GM and direct terrain contributions from the free-air gravity anomalies, and the restore (post-processing) step involves the restoration of the GM contribution and the terrain contribution to N via the indirect effect term N^H.

Formulas for the GM-contributions. In spherical approximation, the geopotential model part of Δg and N is given by the following formulas (see, e.g., Kearsley et al. 1985):

$$\Delta g^{GM} = \overline{g} \sum_{n=2}^{n_{\max}} (n-1) \sum_{m=0}^{n} [C_{nm} \cos\, m\lambda_P + S_{nm} \sin\, m\lambda_P] P_{nm}(\sin \varphi_P), \tag{10.15}$$

$$N^{GM} = R\sum_{n=2}^{n_{\max}}\sum_{m=0}^{n}[C_{nm}\cos m\lambda_P + S_{nm}\sin m\lambda_P]P_{nm}(\sin\varphi_P), \tag{10.16}$$

where C_{nm}, S_{nm} are the fully normalized geopotential coefficients of the anomalous potential (see Chap. 3, Sect. 3.4), P_{nm} are the fully normalized Legendre functions (see Chap. 3, Sect. 3.3), n_{max} is the maximum degree of the geopotential model, $\overline{g}$ is the mean gravity ($= GM/R^2$) and R is the mean radius of the Earth.

Formulas for the Δg -contribution. The contribution of gravity anomalies can be computed in a variety of ways. Here, as an example suitable for FFT-evaluation, the planar approximation of Stokes's integral is briefly discussed. For small distances inside an area E, we can use the planar approximation, where the first term of $S(\psi)$ is the dominant one. Thus we have

$$\frac{1}{\sin(\psi/2)} \approx \frac{2}{\psi} \approx \frac{2R}{l}, \tag{10.17}$$

$$R^2 d\sigma = dxdy, \tag{10.18}$$

and (10.6) reduces to

$$N_P^{\Delta g} = \frac{1}{2\pi\gamma}\iint_E \frac{\Delta g}{l}dxdy, \tag{10.19}$$

$$l = [(\mathrm{x}-\mathrm{x_P})^2 + (\mathrm{y}-\mathrm{y_P})^2]^{1/2}, \tag{10.20}$$

where x, y are the coordinates of the data points and x_P, y_P are the coordinates of the computation point.

Note that (10.19) can also be interpreted as an equivalent of Green's identity (see 1.61 in Chap. 1), with S being just the plane $z = 0$, $u = T$, and $-\partial u/\partial z = \Delta g$.

Formulas for the direct and inverse Terrain contribution. Keeping only the terms for $r = 1$ in (10.12) and (10.13), the terrain effect on Δg and the indirect effect on N take the following form:

$$\begin{aligned}\delta A_P = c_P &= -\Delta g_P^H = \pi G\rho \mathbf{L}(H-H_P)^2 = \pi G\rho[\mathbf{L}H^2 - 2H_P\mathbf{L}H] \\ &= \frac{1}{2}G\rho\iint_E \frac{(H-H_P)^2}{l^3}dxdy = \frac{1}{2}G\rho\iint_E \frac{H^2-H_P^2}{l^3}dxdy \\ &\quad -H_P G\rho\iint_E \frac{H-H_P}{l^3}dxdy,\end{aligned} \tag{10.21}$$

$$\delta N_P = -\frac{\pi G\rho}{\gamma}H_P^2 - \frac{\pi G\rho}{3\gamma}\mathbf{L}H^3 = -\frac{\pi G\rho}{\gamma}H_P^2 - \frac{G\rho}{6\gamma}\iint_E \frac{H^3-H_P^3}{l^3}dxdy. \tag{10.22}$$

In the first line of (10.21), there is no $\mathbf{L}H_P^2$ term since $\mathbf{L}H_P^2 = H_P^2\mathbf{L}(1) = 0$. Also note that the above two formulas, and their imlementation, have already been discussed in Chap. 4, Sects. 4.4 and 4.5.

10.2.4 The Need for Spectral Techniques

Due to the fact that it is very time-consuming to evaluate Stokes's integral, it is often attempted to reduce the size of the area E by modifying Stokes's kernel function. The principle idea, due to Molodensky et al. (1962), is that the truncation error committed by limiting the area of the integration of the terrestrial gravity anomalies to a spherical cap can be reduced by a suitable modification of Stokes's kernel (Jekeli 1982; Hsu 1984). In a different approach, an increased area of integration has been shown to improve the results (Schwarz 1984; Sjöberg 1986). These kinds of methods increase the computational requirements and have not always provided superior results to those from the simple remove-restore technique with the unmodified Stokes kernel.

Integrals of the form of (10.19) are called *convolution integrals* and lend themselves to efficient evaluation by FFT techniques, provided that the data are given on regular grids. The terrain correction integrals of (10.21) and (10.22) can also be formulated as convolution integrals. Using the properties of the Fourier transform, there is no need for time-consuming point-wise numerical summations, and the evaluation of convolution integrals is replaced by very efficient multiplications in the frequency domain. In addition, FFT gives results on the same grid as the grid the data were given on. In other words, in a single run of the FFT software one obtains geoid undulations on all points of the Δg-grid. Thus, spectral techniques based on the FFT overcome very successfully the problem of slow computation speed and provide a homogeneous coverage of results, which is very suitable for interpolation and plotting purposes. Consequently, it may not always be necessary to modify Stokes's kernel function, which becomes even more obvious when the remove-restore technique is employed (Schwarz et al. 1987; Sideris and Forsberg 1990). Instead, the use of spectral techniques is recommended for the computation of large regional and continental geoids, especially since gravity and terrain data are now readily available on regular grids. With some clever techniques for efficient data handling and improved computational speed (see Appendix A, Sect. A.5), the geoid of very large areas can now be computed on any ordinary desktop personal computer.

Section 10.3 discusses in detail the FFT evaluation of Stokes's integral and its advantages and drawbacks in relation to the other available methods. An introduction to the necessary Fourier transform theory is given in Appendix A.

10.3 Geoid Undulations by FFT

10.3.1 Planar Approximation of Stokes's Integral

As stated in Sect. 10.2.3, (10.19), the geoidal height $N^{\Delta g}$ computed from gravity anomalies given by (10.7) in an area E can be expressed in planar approximation by the following two-dimensional convolution integral (Kearsley et al. 1985):

$$N(x_P, y_P) = \frac{1}{2\pi\gamma} \iint_E \frac{\Delta g(x, y)}{\sqrt{(x_p - x)^2 + (y_p - y)^2}} dx dy$$
$$= \frac{1}{\gamma} \Delta g(x_P, y_P) * l_N(x_P, y_P), \tag{10.23}$$

where the superscript Δg has been omitted from N for the sake of simplicity, and l_N is the planar form of Stokes's kernel function:

$$l_N(x, y) = (2\pi)^{-1}(x^2 + y^2)^{-1/2}. \tag{10.24}$$

Using (10.160) and (10.23) is evaluated by two direct and one inverse Fourier transforms as follows:

$$N(x, y) = \frac{1}{\gamma} \mathbf{F^{-1}}\{\mathbf{F}\{\Delta g(x, y)\}\mathbf{F}\{l_N(x, y)\}\} = \frac{1}{\gamma} \mathbf{F^{-1}}\{\Delta G(u, v) L_N(u, v)\}. \tag{10.25}$$

Point gravity anomalies as input. Using $M \times N$ gridded point gravity anomalies with spacing Δx and Δy, the geoid undulation at a point (x_k, y_l) can be evaluated by the following convolution, which is just the discrete form of (10.23):

$$N(x_k, y_l) = \frac{1}{2\pi\gamma} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \Delta g(x_i, y_j) 1_N(x_k - x_i, y_l - y_j) \Delta x \Delta y, \tag{10.26}$$

$$l_N(x_k - x_i, y_l - y_j) = \begin{cases} [(x_k - x_i)^2 + (y_l - y_j)^2]^{-1/2}, & x_k \neq x_i \text{ or } y_l \neq y_j \\ 0, & x_k = x_i \text{ and } y_l = y_j \end{cases}. \tag{10.27}$$

To account for the singularity of l_N, the kernel in (10.27) has been set to zero at the origin and the contribution to N of the gravity anomaly at the computation point (grid element) must be evaluated separately. Approximately, this contribution is (Heiskanen and Moritz 1967; Schwarz et al. 1990)

$$\delta N(x_k, y_l) \approx \frac{\sqrt{\Delta x \Delta y}}{\gamma \sqrt{\pi}} \Delta g(x_k, y_l) \tag{10.28}$$

and expresses the effect on N of a circular region around the computation point with area equal to $\Delta x \Delta y$, having constant gravity anomaly value of $\Delta g(x_k, y_l)$.

A slightly better approximation for δN can be found in Haagmans et al. (1993). Geoid undulations can then be evaluated by FFT as follows:

$$N(x_k, y_l) = \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\Delta G(u_m, v_n) L_N(u_m, v_n)\}. \tag{10.29}$$

ΔG has to be computed by the discrete Fourier transform (DFT) of (10.210)

$$\Delta G(u_m, v_n) = \mathbf{F}\{\Delta g(x_k, y_l)\} = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} \Delta g(x_k, y_l) e^{-i2\pi(mk/M + nl/N)} \Delta x \Delta y. \tag{10.30}$$

L_N can be evaluated either by the DFT

$$L_N(u_m, v_n) = \mathbf{F}\{l_N(x_k, y_l)\} = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} l_N(x_k, y_l) e^{-i2\pi(mk/M + nl/N)} \Delta x \Delta y \tag{10.31}$$

or by the continuous Fourier transform (CFT) of (10.208)

$$L_N(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} l_N(x, y) e^{-i2\pi(ux+vy)} dx dy = \frac{1}{(u^2 + v^2)^{1/2}} = \frac{1}{q}, \tag{10.32}$$

where q is the radial frequency, and then be discretized for use in (10.29). L_N given by (10.32) is called the analytically-defined spectrum of Stokes's kernel. As it will be shown later on, the use of the analytical Fourier transform is not recommended if one wants to obtain results identical to those obtained by numerical integration.

Equations 10.7 and 10.10 show clearly the filtering effect of convolution. The Δg-spectrum is divided by q resulting in attenuation of the high frequencies present in the gravity anomalies. In other words, Stokes's kernel can be considered as a type of low-pass filter, which indicates that the geoid undulations are primarily affected by the low and medium frequencies of the gravity field.

Mean gravity anomalies as input. If the input data are $M \times N$ gridded mean gravity anomalies $\overline{\Delta g}$, the planar Stokes formula can be formulated as

$$N(x_k, y_l) = \frac{1}{2\pi\gamma}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} \overline{\Delta g}(x_i, y_j)\overline{l_N}(x_k - x_i, y_l - y_j)\Delta x \Delta y, \tag{10.33}$$

$$\begin{aligned}\overline{l_N}(x_k, y_l) &= \int_{x_k-\Delta x/2}^{x_k+\Delta x/2}\int_{y_l-\Delta y/2}^{y_l+\Delta y/2} \frac{1}{\sqrt{x^2 + y^2}} dx dy \\ &= x\ln(y + \sqrt{x^2 + y^2}) + y\ln(x + \sqrt{x^2 + y^2}) \Big|_{x_k-\Delta x/2}^{x_k+\Delta x/2} \Big|_{y_l-\Delta y/2}^{y_l+\Delta y/2}.\end{aligned} \tag{10.34}$$

Equation 10.33 can also be efficiently evaluated via FFT, i.e.,

$$N(x_k, y_l) = \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\mathbf{F}\{\overline{\Delta g}(x_k, y_l)\}\mathbf{F}\{\overline{l_N}(x_k, y_l)\}\}$$
$$= \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\overline{\Delta G}(u_m, v_n)\overline{L_N}(u_m, v_n)\}. \quad (10.35)$$

To distinguish between the spectra defined by (10.5) and (10.12), we call $\overline{L_N}$ the mean Stokes kernel spectrum.

It is worth mentioning here that a simple 2D sinc function can also be used to relate the spectrum of data considered as either representing point values at the nodes of a grid or mean values in the area of a grid element (Sideris and Tziavos 1988). By using this technique, if the input gravity anomalies are mean values, the geoid undulations can be expressed as

$$N(x_k, y_l) = \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\text{sinc}\left(\frac{m}{M}\right)\text{sinc}\left(\frac{n}{N}\right)\overline{\Delta G}(u_m, v_n)L_N(u_m, v_n)\}, \quad (10.36)$$

where $\overline{\Delta G}$ is the spectrum of mean gravity anomalies as in (10.35), and L_N is the spectrum of the kernel function as expressed in (10.5). By comparing (10.35) with (10.36), we see that

$$\overline{L_N}(u_m, v_n) = \text{sinc}\left(\frac{m}{M}\right)\text{sinc}\left(\frac{n}{N}\right)L_N(u_m, v_n). \quad (10.37)$$

Equation 10.37 indicates that the Fourier transform of the mean kernel function can theoretically be obtained by multiplying the Fourier transform of the point kernel function, obtained either analytically or by the discrete transform, by a 2D sinc function. For more explanations and a complete discussion, Sideris and Tziavos (1988) should be consulted.

Analytical versus discrete kernel spectrum. Although the analytically-defined spectrum has some advantages compared with the discrete one, such as no DFT required for its evaluation and no effect of leakage and aliasing, it is not suitable for the computation of discrete convolution if we want the results to be the identical to those from numerical integration. Equation 10.3 is, considering the symmetry of the kernel function, equivalent to

$$N(x, y) = \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\Delta G(u, v)L_N^1(u, v)\}$$
$$+ \frac{1}{2\pi\gamma}\mathbf{F}^{-1}\{\Delta G(u, v)L_N^2(u, v)\}, \quad (10.38)$$

$$L_N^1(u, v) = \int_{-T_x/2}^{T_x/2}\int_{-T_y/2}^{T_y/2} l_N(x, y)e^{-i2\pi(ux+vy)}dxdy, \quad (10.39)$$

$$L_N^2(u,v) = 2\int_{T_x/2}^{\infty}\int_{T_y/2}^{\infty} l_N(x,y)e^{-i2\pi(ux+vy)}dxdy, \tag{10.40}$$

where T_x, T_y are the dimensions of the area E. If the grid interval is small enough and the effect of aliasing is negligible when discretizing L_N and Δg, the first term of the right-hand side in (10.38) would be equal to the discrete convolution, i.e., (10.7)–(10.9), and the second term of the right-hand side would be the error due to the analytically defined spectrum used. This error can reach a few decimetres (Sideris and Li 1993) and thus the use of the analytical spectrum should be avoided. For more details and numerical results, Li (1993) and Sideris and Li (1993) should be consulted.

Effects of planar approximation: spherical corrections. The flat-Earth formulas for N developed in the previous sections are valid in the vicinity of the computation point. To avoid long-wavelength errors, the area of local data should not extend to more than several hundreds of kilometers in each direction. This approximation can be improved to any desired accuracy, at least in theory, by using matched asymptotic expansions. Jordan (1978) combined inner and outer expansions into a composite expansion which is valid for small as well as for large distances ψ and describes accurately the gravity field over all wavelengths.

The composite expansion T_c for the disturbing potential depends only on the spherical distance ψ from the coordinate origin located at the centre of the local area and is given by the expression

$$T_c(\psi,\alpha) = \varepsilon_T T(x,y), \quad \varepsilon_T = \left(\frac{\psi/2}{\sin(\psi/2)}\right)^k, \quad k=1 \quad \text{or} \quad k=3, \tag{10.41}$$

where α is the spherical azimuth and ε_T is the correcting factor. ε_T can also be used to correct geoid undulations N and height anomalies ζ. This factor is plotted in Fig. 10.3 along with the corresponding factor for correcting deflections of the vertical ξ and η, which is given by the expression

$$\begin{Bmatrix}\xi(\psi,\alpha)\\ \eta(\psi,\alpha)\end{Bmatrix} = \varepsilon_{\xi,\eta}\begin{Bmatrix}\xi(x,y)\\ \eta(x,y)\end{Bmatrix}, \quad \varepsilon_{\xi,\eta}=\frac{\sin\psi}{\psi}\left(\frac{\psi/2}{\sin(\psi/2)}\right)^{k+2}, \quad k=1 \text{ or } k=3 \tag{10.42}$$

According to Jordan (1978), $k=1$ should be used in (10.41) and (10.42) when $\mathbf{F}\{T(x,y)\}$ at $u=v=0$ is bounded and non-zero, and $k=3$ otherwise. The theoretically correct choice would be $k=3$, since the DC-value of the T-spectrum is zero because the gravity anomalies have a zero mean over the globe. In practice, though, $k=1$ can be used since in limited areas the DC-value of the T-spectrum is neither infinite nor zero. In any case, the whole debate is of minute importance because $\varepsilon_T < 1.01$ and $0.99 < \varepsilon_{\xi\eta} < 1.01$ for $\psi \le 15°$ as Fig. 10.3 indicates. Thus, the corrections are less than 1% for results at distances up to 15° from the centre of the local area and can, in most cases, be safely neglected. Moreover, when the gravity anomalies have been referred to a geopotential model, the outer expansion contribution vanishes and the corrections become truly insignificant.

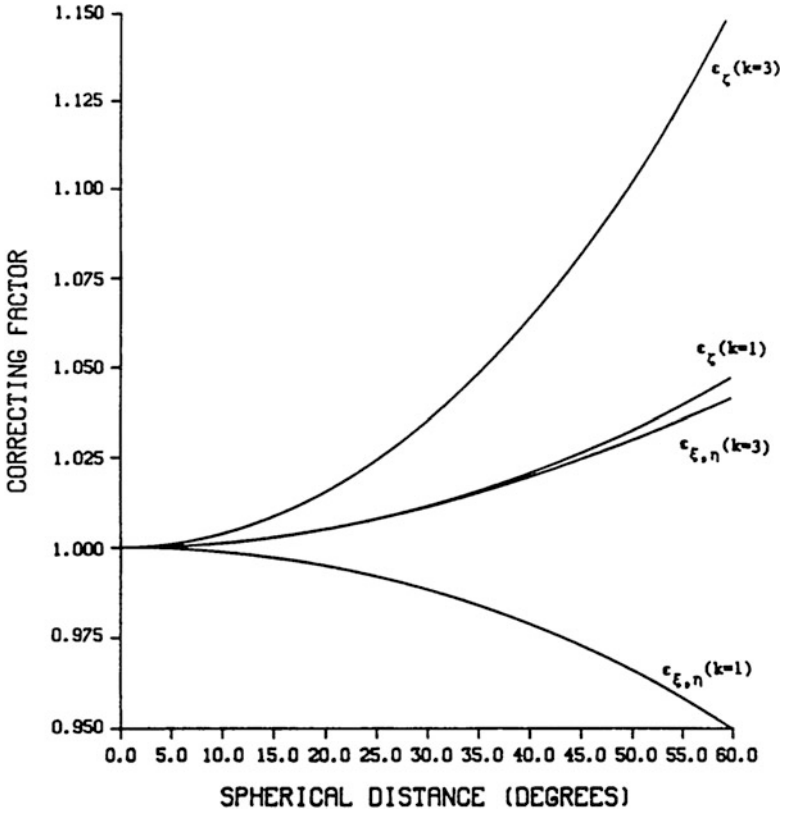


Fig. 10.3 Factors for correcting planar ξ, η, N (or T or ζ) for the earth's curvature (After Sideris 1987)

10.3.2 Spherical Form of Stokes's Integral

The approximations introduced by the planar form of Stokes's integral can be minimized or avoided by using the spherical Stokes integral. Taking into account (10.5), the spherical form of Stokes's integral, i.e., (10.7), can be written explicitly as

$$N(\varphi_p, \lambda_p) = \frac{R}{4\pi\gamma} \iint_E \Delta g(\varphi, \lambda) S(\varphi_p, \lambda_p, \varphi, \lambda) \cos\varphi d\varphi d\lambda. \tag{10.43}$$

With gridded gravity anomalies, (10.43) can be written as

$$N(\varphi_l, \lambda_k) = \frac{R}{4\pi\gamma} \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} \Delta g(\varphi_j, \lambda_i) \cos\varphi_j S(\varphi_l, \lambda_k, \varphi_j, \lambda_i) \Delta\varphi\Delta\lambda. \tag{10.44}$$

With different approximations of Stokes's kernel function on the sphere, geoid undulations can be evaluated at all gridded points simultaneously by means of either the one-dimensional or the two-dimensional fast Fourier transform. These developments are presented in the next few sections.

Approximated spherical kernel. In order to transform (10.44) into a convolution integral, Strang van Hees (1990) suggested to approximate $cos\varphi_P cos\varphi$ in (10.5) by $cos^2\overline{\varphi}$, or by the slightly more accurate $\cos^2\overline{\varphi} - \sin^2(\phi_P - \phi)/2$, where $\overline{\varphi}$ is the mean latitude of the computation area. In this case, (10.5) becomes

$$\begin{aligned}\sin^2\frac{\psi}{2} &\approx \sin^2\frac{\varphi_p-\varphi}{2} + \sin^2\frac{\lambda_p-\lambda}{2}\cos^2\overline{\varphi}\\ &\approx \sin^2\frac{\varphi_p-\varphi}{2} + \sin^2\frac{\lambda_p-\lambda}{2}\left(\cos^2\overline{\varphi} - \sin^2\frac{\varphi_p-\varphi}{2}\right)\end{aligned} \tag{10.45}$$

and (10.44) takes the convolution form

$$\begin{aligned}N(\varphi_l,\lambda_k) &= \frac{R}{4\pi\gamma}\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}\Delta g(\varphi_j,\lambda_i)\cos\varphi_j S(\varphi_l-\varphi_j,\lambda_k-\lambda_i,\overline{\varphi})\Delta\varphi\Delta\lambda\\ &= \frac{R}{4\pi\gamma}[\Delta g(\varphi_l,\lambda_k)\cos\varphi_l] * S(\varphi_l,\lambda_k,\overline{\varphi}).\end{aligned} \tag{10.46}$$

With this approximation, (10.46) can be evaluated efficiently by means of the two-dimensional DFT:

$$N(\varphi_l,\lambda_k) = \frac{R}{4\pi\gamma}\mathbf{F}^{-1}\{\mathbf{F}\{\Delta g(\varphi_l,\lambda_k)\cos\varphi_l\}\mathbf{F}\{S(\varphi_l,\lambda_k,\overline{\varphi})\}\}. \tag{10.47}$$

The approximation of (10.44) by (10.46) makes it possible to compute geoid undulations over large areas on the sphere on all grid points simultaneously by using the two-dimensional Fourier transform. Its disadvantages are that it requires considerable amounts of computer memory because 100% zeros are padded in the latitude and longitude direction, and that additional errors are introduced due to the approximation made on the kernel function. This error can be minimized by the use of the multi-band spherical FFT method proposed by Forsberg and Sideris (1993), which is briefly described below.

Approximated spherical kernel with many bands. Since the errors of the above approximation increase from the centre of the area to the north and south edges, Forsberg and Sideris (1993) proposed to subdivide the area in narrow bands along the longitude direction; see Fig. 10.4. To improve the approximation in (10.45), $\cos\varphi_P \cos\varphi$ can be written as $\cos\varphi_P \cos[\varphi_P - (\varphi_P - \varphi)]$ In each sub-area, φ_P can be considered as constant and again taken as equal to the mean latitude $\overline{\varphi}_i$. In this case, (10.5) is approximated by

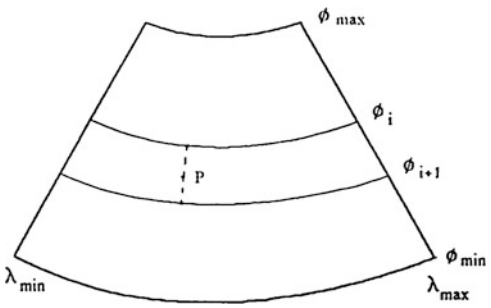


Fig. 10.4 Latitude bands used in the multi-band spherical FFT approach (After Forsberg and Sideris 1993)

$$\begin{aligned}\sin^2\frac{\psi}{2} &\approx \sin^2\frac{\varphi_p-\varphi}{2}+\sin^2\frac{\lambda_p-\lambda}{2}\cos\overline{\varphi}_i\cos[\overline{\varphi}_i-(\overline{\varphi}_i-\varphi)]\\ &\approx \sin^2\frac{\varphi_p-\varphi}{2}+\sin^2\frac{\lambda_p-\lambda}{2}[\cos^2\overline{\varphi}_i\cos(\overline{\varphi}_i-\varphi)\\ &\quad+\cos\overline{\varphi}_i\sin\overline{\varphi}_i\sin(\overline{\varphi}_i-\varphi)] \end{aligned} \tag{10.48}$$

and again the computations are done using (10.47) for each band (with $\overline{\varphi}_i$ in place of $\overline{\varphi}$). Note that for all points along the parallel of mean latitude, an exact solution to the spherical Stokes integral is obtained. By subdividing the area into ν even overlapping latitude zones with mean latitude $\overline{\varphi}_i$, a composite solution for N may be obtained by linear interpolation between the solutions obtained in two consecutive bands with mean latitudes $\overline{\varphi}_i$ and $\overline{\varphi}_{i+1}$:

$$N(\varphi)=\frac{\varphi-\overline{\varphi}_{i+1}}{\overline{\varphi}_i-\overline{\varphi}_{i+1}}N_i+\frac{\overline{\varphi}_i-\varphi}{\overline{\varphi}_i-\overline{\varphi}_{i+1}}N_{i+1}. \tag{10.49}$$

The number of zones ν may be selected according to the required accuracy level and the computer's speed, memory and storage specifications. When $\nu=1$, the solution is identical to the one obtained by the approximation of the previous section.

Rigorous spherical kernel. To overcome the limitations of the previous 2D FFT method, Haagmans et al. (1993) made use of the fact that it provides the exact undulations for all the points along the parallel of mean latitude. Using this property and the addition theorem of the Fourier transform, they came up with an approach which allows for the evaluation of the true discrete spherical Stokes integral without approximation, parallel by parallel, by means of the 1D FFT. In fact, for results on a certain parallel of latitude φ_l using data along a parallel of latitude φ_j, ψ changes only with $\lambda_k-\lambda_i$ and Δg changes only with λ_j and thus the 2D discrete Stokes integral of (10.46) takes the form

$$\begin{aligned}N(\varphi_l,\lambda_k)&=\frac{R}{4\pi\gamma}\sum_{j=0}^{N-1}\left[\sum_{i=0}^{M-1}\Delta g(\varphi_j,\lambda_i)\cos\varphi_j S(\varphi_l,\varphi_j,\lambda_k-\lambda_i)\Delta\lambda\right]\Delta\varphi,\\ \varphi_l&=\varphi_1,\varphi_2,\ldots,\varphi_N.\end{aligned} \tag{10.50}$$

The brackets in (10.50) contain a one-dimensional discrete convolution with respect to λ, i.e., along a parallel, and can be evaluated by the 1D FFT. By employing the addition theorem of DFT, the discrete Stokes integral for the fixed parallel can be evaluated by (Haagmans et al. 1993)

$$N(\varphi_l, \lambda_k) = \frac{R}{4\pi\gamma}\mathbf{F_1^{-1}}\left\{\sum_{j=0}^{N-1}\mathbf{F_1}\{\Delta g(\varphi_j, \lambda_k)\cos\varphi_j\}\mathbf{F_1}\{S(\varphi_l, \varphi_j, \lambda_k\}\right\},$$
$$\varphi_l = \varphi_1, \varphi_2, \ldots, \varphi_N, \tag{10.51}$$

where $\mathbf{F_1}$ and $\mathbf{F_1^{-1}}$ denote the 1D Fourier transform operator and its inverse. Equation 10.51 yields the geoidal heights for all the points on one parallel, which are identical to those obtained by direct summation using (10.46) point by point.

The major advantage of the 1D spherical FFT approach is that it gives exactly the same results as those obtained by direct numerical integration. In addition, it only needs to deal with one one-dimensional complex array each time, resulting in a considerable saving in computer memory as compared to the 2D FFT technique discussed before. Moreover, the adoption of FFT makes it far more computationally efficient than the classical direct numerical integration. Detailed comparisons of various techniques can be found in Haagmans et al. (1993) and Forsberg and Sideris (1993).

10.3.3 Elimination of Edge Effects and Circular Convolution

It must be noted that the Stokes formula is expressed in its various forms by linear convolutions while most fast Fourier transform algorithms are designed for the computation of circular convolutions. Distortion of the results will occur due to the edge effect introduced by using the circular convolution instead of the linear convolution. Figure 3.3 illustrates the effect of circular convolution and that of different zero-padding methods when the computation is either at the centre or at a corner of the computation area. The small circle represents the computation point for the geoid undulation and, at the same time, the maximum kernel function value.

Figure 10.5a shows the correct kernel functions corresponding to numerical integration. Figure 10.5b gives the mirrored kernel functions of the circular convolutions without zero-padding. When the computation point is not at the centre, it can be seen that the periodically mirrored kernel function values are not correct. The conventional method to eliminate the edge effect is to append 100% zeros at each row and column of both convolved functions (Brigham 1988; Bracewell 1986a). However, this method still does not provide correct results at non-central points, as shown in Fig. 10.5c since only part of the data is used in the evaluation. The correct method is to append 100% zeros around the gravity anomalies only, and to compute the kernel function in both the area covered by gravity anomalies and the zero-padded area. As shown in Fig. 10.5d, the kernel function values are identical to those

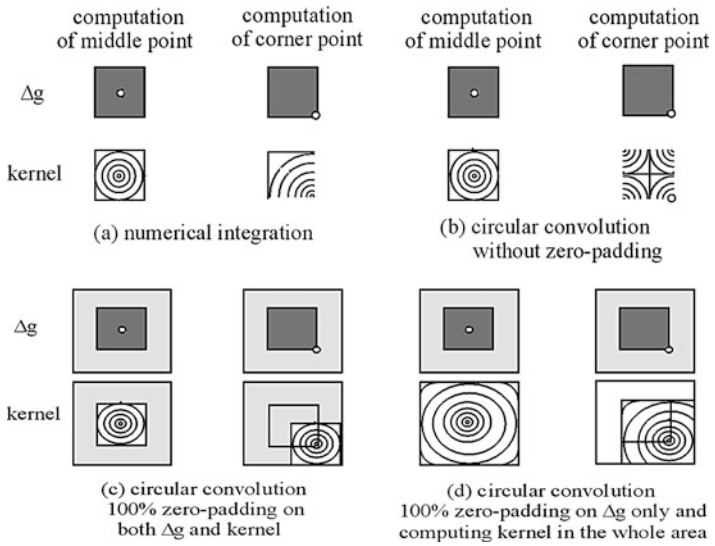


Fig. 10.5 Edge effects and circular convolution in FFT evaluations of Stokes's integral (After Li 1993)

given in Fig. 10.5a. Consequently, with this method, the results computed by the fast Fourier transform techniques are identical to those from rigorous discrete numerical integration. For more details, see Sideris and Li (1992, 1993) and Li (1993).

The above comments are valid for both the spherical and the planar approximations of Stokes's integral. They also hold for the terrain correction integrals that will be discussed below and, in general, for any other gravity field convolution integrals evaluated by FFT.

10.4 FFT-Evaluation of Terrain Effects

10.4.1 2D Formulas for Terrain Effects

Defining the kernel function

$$l_c(x, y) = (x^2 + y^2)^{-3/2}, \tag{10.52}$$

the terrain correction given in Sect. 10.2.3 by (10.46) can be written in any of the following two equivalent convolution forms, where (a) corresponds to the case of constant density and (b) to the case of horizontally varying density:

$$c(x_P, y_P) = \frac{1}{2} G\rho \iint\limits_E \frac{[h(x,y) - h(x_P, y_P)]^2}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G\rho \iint\limits_E \frac{h^2(x,y) - 2h(x_P, y_P)h(x,y) + h^2(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G\rho \{h^2(x_P, y_P) * l_c(x_P, y_P) - 2h(x_P, y_P)[h(x_P, y_P) * l_c(x_P, y_P)]$$

$$+ h^2(x_P, y_P)[o(x_P, y_P) * l_c(x_P, y_P)]\}, \tag{10.53a}$$

$$c(x_P, y_P) = \frac{1}{2} G \iint\limits_E \rho(x,y) \frac{[h(x,y) - h(x_P, y_P)]^2}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G \iint\limits_E \rho(x,y) \frac{h^2(x,y) - 2h(x_P, y_P)h(x,y) + h^2(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G\{[\rho(x_P, y_P)h^2(x_P, y_P)] * l_c(x_P, y_P)$$

$$- 2h(x_P, y_P)[[\rho(x_P, y_P)h(x_P, y_P)] * l_c(x_P, y_P)]$$

$$+ h^2(x_P, y_P)[\rho(x_P, y_P) * l_c(x_P, y_P)]\}, \tag{10.53b}$$

$$c(x_P, y_P) = \frac{1}{2} G\rho \iint\limits_E \frac{h^2(x,y) - h^2(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$- h(x_P, y_P) G\rho \iint\limits_E \frac{h(x,y) - h(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G\rho \{h^2(x_P, y_P) * l_c(x_P, y_P) - h^2(x_P, y_P)[o(x_P, y_P) * l_c(x_P, y_P)]$$

$$- 2h(x_P, y_P)[h(x_P, y_P) * l_c(x_P, y_P)$$

$$- h(x_P, y_P)[o(x_P, y_P) * l_c(x_P, y_P)]]\}, \tag{10.54a}$$

$$c(x_P, y_P) = \frac{1}{2} G \iint\limits_E \rho(x,y) \frac{h^2(x,y) - h^2(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$- h(x_P, y_P) G \iint\limits_E \rho(x,y) \frac{h(x,y) - h(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= \frac{1}{2} G\{[\rho(x_P, y_P)h^2(x_P, y_P)] * l_c(x_P, y_P) - h^2(x_P, y_P)[\rho(x_P, y_P) * l_c(x_P, y_P)] - 2h(x_P, y_P)[[\rho(x_P, y_P)h(x_P, y_P)] * l_c(x_P, y_P) - h(x_P, y_P)[\rho(x_P, y_P) * l_c(x_P, y_P)]]\}, \tag{10.54b}$$

where $o(x, y)$ has the value of one, i.e., $o(x, y) = 1$, for all grid points. Similarly, the indirect effect on the geoid, which was given in Sect. 10.2.3 by (10.23), can be written in the convolution form

$$\delta N(x_P, y_P) = -\frac{\pi G\rho}{\gamma} h^2(x_P, y_P) - \frac{G\rho}{6\gamma} \iint_E \frac{h^3(x, y) - h_P^3(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= -\frac{\pi G\rho}{\gamma} h^2(x_P, y_P) - \frac{G\rho}{6\gamma}\{h^3(x_P, y_P) * l_c(x_P, y_P) - h^3(x_P, y_P)[o((x_P, y_P) * l_c(x_P, y_P)]\}. \tag{10.55a}$$

$$\delta N(x_P, y_P) = -\frac{\pi G\rho}{\gamma} h^2(x_P, y_P) - \frac{G}{6\gamma} \iint_E \rho(x, y) \frac{h^3(x, y) - h_P^3(x_P, y_P)}{[(x_P - x)^2 + (x_P - x)^2]^{3/2}} dxdy$$

$$= -\frac{\pi G\rho}{\gamma} h^2(x_P, y_P) - \frac{G}{6\gamma}\{[\rho(x_P, y_P)h^3(x_P, y_P)] * l_c(x_P, y_P) - h^3(x_P, y_P)[\rho((x_P, y_P) * l_c(x_P, y_P)]\}. \tag{10.55b}$$

The singularity of the l_c kernel function is again bypassed by setting $l_c(0, 0) = 0$. This is of no practical consequence because all integrals above contain not the heights but the height differences which are zero when $x = x_P$ and $y = y_P$. Although (10.54) presents a weaker singularity than (10.53), the latter has been used in practice more often because it requires fewer Fourier transforms. For a detailed discussion on the singularity of the terrain correction formula, Klose and Ilk (1992) should be consulted.

We gave the above three formulas both for constant and for variable density to show how the convolutions are set up in each case. However, as we have done in the proceeding sections, we will continue here developing the formulas only for the simpler case of constant density and the reader can easily modify them in the case of variable density.

Using Fourier transforms, the above equations are evaluated as follows:

$$c(x,y) = \frac{1}{2}G\rho[\mathbf{F}^{-1}\{H_2(u,v)L_c(u,v)\} - 2h(x,y)\mathbf{F}^{-1}\{H(u,v)L_c(u,v)\} + h^2(x,y)\mathbf{F}^{-1}\{O(u,v)L_c(u,v)\}], \tag{10.56}$$

$$c(x,y) = \frac{1}{2}G\rho[\mathbf{F}^{-1}\{H_2(u,v)L_c(u,v)\} - h^2(x,y)\mathbf{F}^{-1}\{O(u,v)L_c(u,v)\} - 2h(x,y)(\mathbf{F}^{-1}\{H(u,v)L_c(u,v)\} - h(x,y)\mathbf{F}^{-1}\{O(u,v)L_c(u,v)\})], \tag{10.57}$$

$$\delta N(x,y) = -\frac{\pi G\rho}{\gamma}h^2(x,y) - \frac{G\rho}{6\gamma}[\mathbf{F}^{-1}\{H_3(u,v)L_c(u,v)\} - h^3(x,y)\mathbf{F}^{-1}\{O(x,y)L_c(u,v)\}], \tag{10.58}$$

where $H_i(u,v) = \mathbf{F}\{h^i(x,y)\}$ for $i = 2,3$, $O(u,v) = \mathbf{F}\{o(x,y)\}$, and $L_c(u,v) = \mathbf{F}\{l_c(x,y)\}$.

As an example, in the following we will give more details on the FFT-evaluation of (10.56) for c. Equations 10.54 and 10.55 can be treated in the same manner and will not be explicitly discussed here.

Point heights as input. Using $M \times N$ gridded point heights, (10.53) can be replaced by

$$c(x_k,y_l) = \frac{1}{2}G\rho\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\frac{[h(x_i,y_j)-h(x_k,y_l)]^2}{[(x_k-x_i)^2+(y_l-y_j)^2]^{3/2}}\Delta x\Delta y, \tag{10.59}$$

and can be efficiently evaluated via FFT (Sideris 1984)

$$c(x_k,y_l) = \frac{1}{2}G\rho[\mathbf{F}^{-1}\{H_2(u_m,v_n)L_c(u_m,v_n)\} - 2h(x_k,y_l)\mathbf{F}^{-1}\{H(u_m,v_n)L_c(u_m,v_n)\} + h^2(x_k,y_l)\mathbf{F}^{-1}\{O(u_m,v_n)L_c(u_m,v_n)\}]. \tag{10.60}$$

If zero-padding is not adopted for the heights, we can easily see that the last convolution in (10.53) or the last inverse Fourier transform in (10.60) reduces to the DC value of L_c, $L_c(0,0)$ (Sideris 1985). However, in order to avoid circular convolution, we should use the zero-padding technique. In this case, $o(x,y) = 1$ at data points and $o(x,y) = 0$ at the zero-padded points and, therefore, $\mathbf{F}^{-1}\{\mathbf{F}\{o(x,y)\}\mathbf{F}\{l_c(x,y)\}\}$ has to be computed explicitly.

Mean heights as input. If the input are $M \times N$ mean gridded heights $\overline{h}$, in place of (10.53) we can write

$$c(x_k, y_l) = \frac{1}{2} G\rho \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [\overline{h}(x_i, y_j) - \overline{h}(x_k, y_l)]^2 \overline{l}_c(x_k - x_i, y_l - y_j), \quad (10.61)$$

$$\begin{aligned} \overline{l}_c(x_k, y_l) &= \int_{x_k - \Delta x/2}^{x_k + \Delta x/2} \int_{y_1 - \Delta y/2}^{y_l + \Delta y/2} \frac{1}{(x^2 + y^2)^{3/2}} dx dy \\ &= \frac{(x^2 + y^2)^{1/2}}{xy} \Big|_{x_k - \Delta x/2}^{x_k + \Delta x/2} \Big|_{y_l - \Delta l/2}^{y_l + \Delta l/2}, \end{aligned} \quad (10.62)$$

Denoting the spectrum of the mean height kernel $\overline{l}_c$ by $\overline{L}_c$, (10.61) can be efficiently computed by using FFT as follows:

$$\begin{aligned} c(x_k, y_l) = \frac{1}{2} G\rho [& \mathbf{F}^{-1}\{\overline{H}_2(u_m, v_n)\overline{L}_c(u_m, v_n)\} \\ & - 2\overline{h}(x_k, y_l)\mathbf{F}^{-1}\{\overline{H}(u_m, v_n)\overline{L}_c(u_m, v_n)\} \\ & + \overline{h}^2(x_k, y_l)\mathbf{F}^{-1}\{O(u_m, v_n)\overline{L}_c(u_m, v_n)\}]. \end{aligned} \quad (10.63)$$

Analytical versus discrete kernel spectrum. It is interesting to mention here that the terrain correction can also be evaluated by using an analytical kernel spectrum. Starting from Laplace's (10.1) and the derivative property of the Fourier transform of (10.149), we can obtain the spectrum of Laplace's equation for a gravity field harmonic function $f(x, y, z)$ as

$$\mathbf{F}\{\Delta f(x, y, z)\} = [(i2\pi u)^2 + (i2\pi v)^2 + (i2\pi w)^2]\mathbf{F}\{f(x, y, z)\} = 0. \quad (10.64)$$

Since in general $\mathbf{F}\{f(x, y, z)\} \neq 0$, (10.64) yields

$$u^2 + v^2 + w^2 = 0 \quad \text{or} \quad -w^2 = u^2 + v^2 = q^2 \quad \text{or} \quad iw = \pm q, \quad (10.65)$$

where the plus sign corresponds to increasing values of f towards the positive z-axis. Thus, since the gravity field quantities decrease with height, the vertical derivative spectrum is

$$\mathbf{F}\left\{\frac{\partial^n f}{\partial z^n}\right\} = \mathbf{F}\{\mathbf{L}^n f\} = \mathbf{F}\{(d_z *)^n \mathbf{F}\{f\} = D_z^n \mathbf{F}\{f\} = (-2\pi q)^n \mathbf{F}\{f\}, \quad (10.66)$$

where we have made use of the vertical derivative operator $\mathbf{L}$ of (10.11) and we have used d_z to denote a function which, when convolved with another function, yields its derivative.

Since any function given on the $x - y$ plane can be considered as harmonic in three-dimensional space, we will apply the above findings when f is a power of height. In this case, and using (10.11), (10.21) and (10.22) can be written in the following convolution form:

$$c(x_P, y_P) = \pi G\rho\{h^2(x_P, y_P) * d_z(x_P, y_P) - 2h(x_P, y_P)[h(x_P, y_P) * d_z(x_P, y_P)]\}, \quad (10.67)$$

$$\delta N(x_P, y_P) = -\frac{\pi G\rho}{\gamma} h^2(x_P, y_P) - \frac{\pi G\rho}{3\gamma}[h^3(x_P, y_P) * d_z(x_P, y_P)], \quad (10.68)$$

and can be evaluated by FFT as follows:

$$\begin{aligned} c(x, y) &= \pi G\rho[\mathbf{F}^{-1}\{H_2(u, v) D_z(u, v)\} - 2h^2(x, y)\mathbf{F}^{-1}\{H(u, v) D_c(u, v)\} \\ &= \pi G\rho[\mathbf{F}^{-1}\{-2\pi q H_2(u, v)\} - 2h^2(x, y)\mathbf{F}^{-1}\{-2\pi q H(u, v)\}, \end{aligned} \quad (10.69)$$

$$\begin{aligned} \delta N(x, y) &= -\frac{\pi G\rho}{\gamma} h^2(x, y) - \frac{\pi G\rho}{3\gamma}\mathbf{F}^{-1}\{H_3(u, v) D_z(u, v)\} \\ &= -\frac{\pi G\rho}{\gamma} h^2(x, y) - \frac{\pi G\rho}{3\gamma}\mathbf{F}^{-1}\{-2\pi q H_3(u, v)\}. \end{aligned} \quad (10.70)$$

The above equations, although they require fewer Fourier transformations than the formulas of the previous sections, are not recommended for numerical evaluations. The reasons are the same as those given in Sect. 10.2.3 for Stokes's integral. Thus, to obtain by FFT identical results as those from numerical integration, the discrete kernel should be used and all convolutions should be evaluated using proper zero-padding (see Sect. 10.3.3). The above formulas, however, are illustrating clearly the dependence of terrain effects on higher derivatives of powers of heights, thus demonstrating the high-pass filtering nature of these operations and the need for dense topographic information for accurate results.

10.4.2 Terrain Corrections by 3D FFT

Gravity terrain corrections can also be computed by the three-dimensional fast Fourier transform (3D FFT) method. By using density values on a 3D grid, a 3D grid of terrain corrections is produced from which the terrain corrections of the points on the Earth's surface are evaluated by interpolation. The technique gives directly the results at the geoid level, i.e., the indirect effect of the topographic reduction, and at a flight level, which finds a very important application in airborne gravimetry and gradiometry measurements.

The topographic vertical attraction at a point $P(x_P, y_P, z_P)$ on the surface of the Earth is the negative first-order derivative of the potential of the topographic masses in the z-direction and can be expressed as

$$T_z(x_P, y_P, h_P) = G \iiint_{E\,0}^{h} \frac{(h_P - h)\rho(x, y, z)}{[(x_P - x)^2 + (y_P - y)^2 + (h_P - z)^2]^{3/2}} dz dx dy. \tag{10.71}$$

The topographic effect on gravity can be separated into two parts: the Bouguer plate effect B and the terrain correction c. Equation 10.71, therefore, can be rewritten as

$$T_z(x_p, y_p, h_p) = B(x_p, y_p, h_p) - c(x_p, y_p, h_p), \tag{10.72}$$

$$B(x_P, y_P, h_P) = G \iiint_{E\,0}^{h_P} \frac{(h_P - z)\rho(x, y, z)}{[(x_P - x)^2 + (y_P - y)^2 + (h_P - z)^2]^{3/2}} dz dx dy, \tag{10.73}$$

$$c(x_p, y_p, h_p) = G \iiint_{E\,h}^{h_P} \frac{(h_p - z)\rho(x, y, z)}{[(x_p - x)^2 + (y_p - y)^2 + (h_p - z)^2]^{3/2}} dz dx dy. \tag{10.74}$$

Assuming constant density and that the area E is bounded by x_{min} and x_{max} in the x-direction and y_{min} and y_{max} in the y-direction, the Bouguer effect B can be evaluated (Nagy 1966) by

$$\begin{aligned} B(x_p, y_p, h_p) = G\rho\{&(x_p - x)\ln[(y_p - y) + r] + (y_p - y)\ln[(x_p - x) + r] \\ &- (h_p - z)\tan^{-1}\left[\frac{(x_p - x)(y_p - y)}{(h_p - z)r}\right]\Bigg\}\Bigg|_{x_{\min}}^{x_{\max}}\Bigg|_{y_{\min}}^{y_{\max}}\Bigg|_{0}^{h_P}, \\ r = [&(x_p - x)^2 + (y_p - y)^2 + (h_p - z)^2]^{1/2}. \end{aligned} \tag{10.75}$$

If the radius of the area E is infinite, the Bouguer effect B can be expressed as

$$B\left(x_p, y_p, h_p\right) = 2\pi G\rho h_p. \tag{10.76}$$

The computation of (10.75) and (10.76) is straightforward. In order to get the terrain correction of (10.74) from (10.72), we need to discuss how to evaluate (10.71) accurately and effectively. Actually, (10.71) can be evaluated directly by two methods: one is the numerical integration method, which is rigorous but very time-consuming; the other one is the 3D FFT method, which will be discussed in the following.

Suppose that the masses between the geoid and the topography can be divided into many small prisms with the same $x - y$ cross-section. If the density within each prism can be taken as constant, (10.71) can be discretized as follows:

$$T_z(x_k, y_l, z_\mu) = G \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{\kappa=1}^{K} \rho(x_i, y_j, z_k) \bar{l}_3(x_k - x_i, y_l - y_j, z_\kappa - z_k), \quad (10.77)$$

$$\begin{aligned} \bar{l}_3(x_k, y_l, z_\mu) &= \iiint_{\Delta v_{kl\mu}} \frac{z}{(x^2 + y^2 + z^2)^{3/2}} dx dy dz \\ &= \Bigg[x \ln(y + r') + y \ln(x + r') \\ &\quad - z \arctan\left(\frac{xy}{zr'}\right) \Bigg] \Bigg|_{x_k - \Delta x/2}^{x_k + \Delta x/2} \Bigg|_{y_l - \Delta y/2}^{y_l + \Delta y/2} \Bigg|_{z_\mu - \Delta z/2}^{z_\mu + \Delta z/2}, \quad (10.78) \\ r' &= (x^2 + y^2 + z^2)^{1/2}. \end{aligned}$$

M, N, K are the actual dimensions of the 3D grid in the x-, y- and z-direction, respectively and Δv_{ijk} is the volume of each grid element, defined by grid spacing Δx, Δy and Δz. The singularity of the $\bar{l}_3$ kernel function is again treated by setting $\bar{l}_3(0,0,0) = 0$. Equation 10.77 can then be expressed in the convolution form

$$T_z(x_k, y_l, z_\mu) = G\rho(x_k, y_l, z_\mu) * \bar{l}_3(x_k, y_l, z_\mu), \quad (10.79)$$

and be efficiently evaluated by the 3D FFT as follows:

$$\begin{aligned} T_z(x_k, y_k, z_k) &= G\mathbf{F}^{-1}\{\mathbf{F}\{\rho(x_k, y_k, z_k)\}\mathbf{F}\{\bar{l}_3(x_k, y_k, z_k)\}\} \\ &= G\mathbf{F}^{-1}\{P(u_m, v_n, w_\lambda)\overline{L}_3(u_m, v_n, w_\lambda)\}. \quad (10.80) \end{aligned}$$

This method is not seen as a replacement of the 2D FFT method since the latter, when more terms are kept in the Taylor series expansion, is capable of producing results of the same accuracy. The 3D FFT method has, however, two important advantages over the 2D FFT method. First, it is unaffected by terrain inclination and thus avoids the numerical difficulties present in the 2D FFT method. And second, and most important, it can handle varying density in the z-direction, which is not possible with the 2D FFT method. This latter property makes the 3D FFT the only efficient alternative to numerical integration in situations where the three-dimensional density distribution is known from the geology of the area or from geophysical surveys. In the above two cases, even when the results are only needed on the Earth's surface or at a specific level, the extra effort and computer resources required are justified in order to avoid (1) the shortcomings of the 2D FFT method and (2) the long computation time required by numerical integration. A detailed description of the 3D FFT method along with numerical results can be found in Peng (1994) and Peng et al. (1995).

10.5 Optimal Spectral Geoid Determination

10.5.1 Error Propagation

The FFT method can use heterogeneous data, provided that they are given on a grid, and can produce error estimates, provided that the power spectral densities (PSDs), which are the Fourier transform of the covariance functions, of the data and their noise are known. In this case, the technique is equivalent to frequancy-domain collocation. To illustrate how error propagation can be used with Stokes's integral, we first rewrite (10.23) for noisless data in the following convolution form:

$$N = \Delta g * s \tag{10.81}$$

where $s = l_N/\gamma$. We now assume that the "observed" gravity anomalies have errors n, i.e., $\Delta g_o = \Delta g + n$, with known PSD P_{nn}. In this case, (10.81) becomes

$$\hat{N} = \Delta g_o * s_o = (\Delta g + n) * s_o \tag{10.82}$$

and then we can write

$$N = \hat{N} + e = \Delta g_o * s_o + e = (\Delta g + n) * s_o + e \tag{10.83}$$

where e is the error of the estimated undulations $\hat{N}$; see also Fig. 10.6. In the frequency domain, (10.82) and (10.83) have the form

$$\mathbf{F}\{\hat{N}\} = \mathbf{F}\{\Delta g_o\}\mathbf{F}\{s_o\} = (\mathbf{F}\{\Delta g\} + \mathbf{F}\{n\})\mathbf{F}\{s_o\} \tag{10.84}$$

$$\begin{aligned}\mathbf{F}\{N\} &= (\mathbf{F}\{\Delta g\} + \mathbf{F}\{n\})\mathbf{F}\{s_o\} + \mathbf{F}\{e\} \\ &= \mathbf{F}\{\Delta g_o\}\mathbf{F}\{s_o\} + \mathbf{F}\{e\} = \mathbf{F}\{\hat{N}\} + \mathbf{F}\{e\}\end{aligned} \tag{10.85}$$

Multiplying the above expression by its complex conjugate first and then by the complex conjugate of the Δg-spectrum, and assuming no correlation between signal and noise and between input and output noise, i.e., the corresponding cross-PSDs are zero, the PSD of N and the cross-PSD of Δg and N can be derived by taking the expectations of the corresponding expression (see 10.184). They are

$$\begin{aligned}P_{NN} &= P_{\hat{N}\hat{N}} + P_{ee} = S_o(P_{\Delta g\Delta g} + P_{nn})S_o^* + P_{ee} \\ &= S_o P_{\Delta g_o \Delta g_o} S_o^* + P_{ee} = |S_o|^2 P_{\Delta g_o \Delta g_o} + P_{ee},\end{aligned} \tag{10.86}$$

$$P_{N\Delta g} = S_o(P_{\Delta g\Delta g} + P_{nn}) = S_o P_{\Delta g_o \Delta g_o}, \tag{10.87}$$

where S_o is the spectrum of s_o and $P_{\Delta g\Delta g}$ and $P_{\Delta go\Delta go}$ are the PSDs of the noise-free and observed gravity anomalies, respectively. From the above equations, the spectrum of the "modified" (so as o filter out the input noise) Stokes kernel and

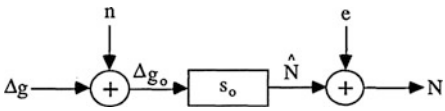


Fig. 10.6 Stokes's convolution (Δg-input, N-output system) with noisy data (After Sideris 1996)

the ouput noise PSD become dependent on the signal-to-noise ratio $P_{\Delta g\Delta g}/P_{nn}$, as follows:

$$S_o = \frac{P_{N\Delta g_o}}{P_{\Delta g_o \Delta g_o}} = \frac{P_{N\Delta g}}{P_{\Delta g\Delta g} + P_{nn}} = S\left(1 + \frac{P_{nn}}{P_{\Delta g\Delta g}}\right)^{-1}, \quad S = \frac{P_{N\Delta g}}{P_{\Delta g\Delta g}} \tag{10.88}$$

$$\begin{aligned} P_{ee} &= P_{NN} - P_{N\Delta g}(P_{\Delta g\Delta g} + P_{nn})^{-1} P_{\Delta gN} \\ &= |S|^2 P_{\Delta g\Delta g}[1 - (1 + P_{nn}/P_{\Delta g\Delta g})^{-1}], \end{aligned} \tag{10.89}$$

where S is the spectrum of Stokes's kernel for noiseless data ($n = 0$). In this case, $e = 0$, $P_{ee} = 0$, and $S_o = S$. The spectrum of the estimated undulations of (10.84) can then be written as

$$\mathbf{F}\{\hat{N}\} = \mathbf{F}\{\Delta g_o\}\mathbf{F}\{s_o\} = \frac{P_{N\Delta g}}{P_{\Delta g\Delta g} + P_{nn}} \mathbf{F}\{\Delta g_o\}. \tag{10.90}$$

It is now evident that, using the signal-to-noise ratio, Stokes's kernel can be modified to filter out the noise of the input data. Moreover, error estimates can be computed for the results by obtaining the inverse Fourier transform of P_{ee} which yields the error covariance matrix of the predicted geoid undulations.

Recalling that the PSD function is the spectrum of the covariance function, (10.90) and (10.89) are the frequency-domain representation of the least-squares collocation (LSC) equations (Moritz 1980). Actually, similarly to collocation, (10.88)–(10.90) can be obtained by minimizing P_{ee} with respect to S_o (or S_o^*). We will show this below, starting from the expression for the output noise spectrum, which, from Fig. 10.6 and (10.85), is

$$\mathbf{F}\{e\} = \mathbf{F}\{N\} - \mathbf{F}\{\hat{N}\} = \mathbf{F}\{N\} - (\mathbf{F}\{\Delta g\} + \mathbf{F}\{n\})\mathbf{F}\{s_o\} \tag{10.91}$$

Multiplying the above expression by its complex conjugate first and then taking the expectations of the resulting terms yields the PSD of e:

$$P_{ee} = P_{NN} - P_{\hat{N}\hat{N}} = P_{NN} - S_o P_{\Delta gN} - S_o^* P_{N\Delta g} - S_o(P_{\Delta g\Delta g} + P_{nn})S_o^* \tag{10.92}$$

The optimal S_o is the one that minimizes P_{ee}, and can be obtained by setting the derivative of P_{ee} with respect to S_o^* equal to zero (see also Bendat and Piersol 1986, Sect. 6.2.4):

$$\frac{\partial P_{ee}}{\partial S_o^*} = -2P_{N\Delta g} + 2S_o(P_{\Delta g\Delta g} + P_{nn}) = 0 \tag{10.93}$$

which yields the S_o of (10.87). Substituting it into (10.84) and (10.91) we obtain (10.90) and (10.89), respectively.

It is obvious from the above discussion that the FFT method can, like LSC, use heterogeneous noisy data, provided that they are given on a grid, by use of the multiple-input, multiple-output systems theory; for details, consult Bendat and Piersol (1980) and Sideris (1996). Note, however, that for the PSDs to be the Fourier ransform of the covariace functions (CVs) used in collocation, these CVs have to be stationary, which is not the case in practice with the noise CVs. Thus, the FFT method, although it is much more efficient that LSC collocation (because it does not require any matrix inversion), has to approximate non-stationary noise covariance functions (which are easily handled by LSC) by stationary ones. For a detailed dicussion on this, Sansò and Sideris (1997) and Kotsakis and Sideris (2001) should be consulted.

10.6 Other Examples of FFT Evaluation of Geodetic Operators

10.6.1 The Vening Meinesz Integral

Since deflections of the vertical are the horizontal derivatives of the geoid, the general property of the Fourier transform of the derivative of a function (see 10.153) can be used to derive the planar approximation form of the Vening Meinesz integral from the Stokes integral:

$$\begin{Bmatrix} \xi(x_p, y_p) \\ \eta(x_p, y_p) \end{Bmatrix} = \begin{Bmatrix} -\partial N(x_p, y_p)/\partial y_P \\ -\partial N(x_p, y_p)/\partial x_P \end{Bmatrix}$$

$$= -\frac{1}{2\pi\gamma} \begin{Bmatrix} \Delta g(x_p, y_p) * \partial l_N(x_p, y_p)/\partial y_P \\ \Delta g(x_p, y_p) * \partial l_N(x_p, y_p)/\partial y_P \end{Bmatrix} \tag{10.94}$$

or, equivalently,

$$\begin{Bmatrix} \xi(x_p, y_p) \\ \eta(x_p, y_p) \end{Bmatrix} = \frac{1}{2\pi\gamma} \iint_E \Delta g(x, y) \frac{1}{\left[(x_p - x)^2 + (y_p - y)^2\right]^{3/2}} \begin{Bmatrix} y_p - y \\ x_p - x \end{Bmatrix} dxdy$$

$$= -\frac{1}{2\pi\gamma} \Delta g(x_p, y_p) * \begin{Bmatrix} l_\xi(x_p, y_p) \\ l_\eta(x_p, y_p) \end{Bmatrix}, \tag{10.95}$$

where

$$\begin{Bmatrix} l_\xi(x, y) \\ l_\eta(x, y) \end{Bmatrix} = -\begin{Bmatrix} \partial l_N(x, y)/\partial y \\ \partial l_N(x, y)/\partial x \end{Bmatrix} = (x^2 + y^2)^{-3/2} \begin{Bmatrix} y \\ x \end{Bmatrix}. \tag{10.96}$$

Using (10.149), the spectra of the Vening Meinesz kernels can be obtained directly from the spectrum of the Stokes kernel (see 10.32) as

$$\mathbf{F}\left\{\begin{array}{c} l_\xi(x,y) \\ l_\eta(x,y) \end{array}\right\} = -\left\{\begin{array}{c} i2\pi v \\ i2\pi u \end{array}\right\} \mathbf{F}\{l_N(x,y)\} = -\left\{\begin{array}{c} i2\pi v \\ i2\pi u \end{array}\right\} \frac{1}{q}$$
$$= -\left\{\begin{array}{c} i2\pi v \\ i2\pi u \end{array}\right\} \frac{1}{(u^2+v^2)^{1/2}}. \tag{10.97}$$

The deflections of the vertical can thus be also evaluated by FFT as follows:

$$\left\{\begin{array}{c} \xi(x_p,y_p) \\ \eta(x_p,y_p) \end{array}\right\} = -\frac{1}{2\pi\gamma}\left\{\begin{array}{c} \mathbf{F}^{-1}\{\mathbf{F}\{\Delta g(x_p,y_p)\}\mathbf{F}\{l_\xi(x_p,y_p)\}\} \\ \mathbf{F}^{-1}\{\mathbf{F}\{\Delta g(x_p,y_p)\}\mathbf{F}\{l_\eta(x_p,y_p)\}\} \end{array}\right\}. \tag{10.98}$$

As in the case of the geoid undulations and for exactly the same reason, the use of the analytical spectrum of (10.97) is not recommended in practice. Instead, the spectra of the kernels of (10.96) should be computed numerically, and proper zero padding should be applied when (10.98) is evaluated.

It must also be noted that deflections of the vertical can also be evaluated by the 2D and 1D FFT on the sphere by use of the spherical Vening Meinesz kernels, analogously to the procedure followed in Sect. 10.3.2 for the geoid undulations. Formulas and numerical tests can be found in Liu et al. (1997).

10.6.2 The Analytical Continuation Integrals

Continuation integrals are often used in applications such as airborne gravimetry to relate gravity anomalies at flight altitude, $h = z_o$, to gravity anomalies at geoid level, $h = 0$. In planar approximation, this relationshp is given by the following integral:

$$\Delta g(x_P,y_P,z_0) = \frac{1}{2\pi}\iint_E \Delta g(x,y,0)\frac{z_0}{[(x_P-x)^2+(y_P-y)^2+z_0^2]^{3/2}}\,dxdy$$
$$= \Delta g(x_P,y_P,0) * l_u(x_P,y_P,z_0), \tag{10.99}$$

where the upward continuation kernel is

$$l_u(x,y,z_0) = \frac{z_0}{2\pi[(x^2+y^2+z_0^2]^{3/2}}. \tag{10.100}$$

Since (10.99) is a convolution, it can be evaluated as follows:

$$\Delta g(x_P,y_P,z_0) = \mathbf{F}^{-1}\{\mathbf{F}\{\Delta g(x_P,y_P,0)\}\mathbf{F}\{l_u(x_P,y_P,z_0)\}\} \tag{10.101}$$

The upward continuation kernel does have an analytically defined spectrum

$$\mathbf{F}\{l_u(x_P, y_P, z_o)\} = L_u(u, v, z_o) = e^{-2\pi z_o (u^2+v^2)^{1/2}} = e^{-2\pi z_o q}, \tag{10.102}$$

which illustrates clearly that upward continuation attenuates the high frequencies of the gravity field.

Equation 10.101 can be reversed to obtain the formula for downward continuation and the spectrum of the downward continuation kernel l_d:

$$\begin{aligned} \Delta g(x_P, y_P, 0) &= \mathbf{F}^{-1}\left\{\frac{\mathbf{F}\{\Delta g(x_P, y_P, z_o)\}}{\mathbf{F}\{l_u(x_P, y_P, z_o)\}}\right\} \\ &= \mathbf{F}^{-1}\{\mathbf{F}\{\Delta g(x_P, y_P, z_0)\}\mathbf{F}\{l_d(x_P, y_P, z_0)\}\} \qquad (10.103) \\ \mathbf{F}\{l_d(x_P, y_P, z_o)\} &= 1/L_u(u, v, z_o) = e^{2\pi z_o (u^2+v^2)^{1/2}} = e^{2\pi z_o q}. \qquad (10.104) \end{aligned}$$

As expected, downward continuation amplifies the high frequencies and therefore the data noise as well, and the solution obtained by (10.103) is usually stabilized by use of a winner filter, which makes use of the PSD of the data noise P_{nn} (similar to (10.88) and (10.90) for the undulations). Again, the use of the analytical spectra of l_u and l_d is not recommended for numerical evaluations.

10.6.3 The Inverse Stokes and Inverse Mening Meinesz Formulas

It is easy in planar approximation to invert the spectrum of Stokes's equation (10.23) to obtain

$$\mathbf{F}\{\Delta g\} = \frac{\gamma \mathbf{F}\{N\}}{\mathbf{F}\{l_N\}} = 2\pi q \gamma \mathbf{F}\{N\} \tag{10.105}$$

This equation is useful to obtain gravity anomalies from altimetry-derived undulations. Gravity anomalies can also be obtained from deflections of the vertical. As discussed in Sect. 9.9.1, (9.43), a straightforward result of Laplace's equation on the plane is the relationship

$$-\frac{\partial \Delta g}{\partial z} = \gamma\left(\frac{\partial \xi}{\partial y} + \frac{\partial \eta}{\partial x}\right) \tag{10.106}$$

By taking the Fourier transform of both sides and making use of (10.149) and (10.64)–(10.66) we obtain

$$-2\pi q \mathbf{F}\{\Delta g\} = \gamma 2\pi i (v\mathbf{F}\{\xi\} + u\mathbf{F}\{\eta\}) \tag{10.107}$$

Then the spectrum of Δg can be obtained from the spectra of ξ and η (see also 9.46):

$$\mathbf{F}\{\Delta g\} = -\gamma \frac{i}{q}(v\mathbf{F}\{\xi\} + u\mathbf{F}\{\eta\}) \tag{10.108}$$

By combining (10.105) and (10.108) we can also get the spectrum of N from the spectra of ξ and η (see also 9.47):

$$\mathbf{F}\{N\} = -\gamma \frac{i}{2\pi q^2}(v\mathbf{F}\{\xi\} + u\mathbf{F}\{\eta\}) \tag{10.109}$$

As discussed is Sect. 9.9.2, Winner filtering is employed in altimetry applications to evaluate the above equations with noisy data. This is more critical for (10.105) than (10.109), as the former becomes unstable due to noise amplification at high frequencies.

10.7 Concluding Remarks

The main advantage of spectral methods is that they can efficiently handle large amounts of gridded data and give results on all grid points simultaneously, which has made them a standard and indispensable tool for geoid computations. There are some problems that affect the accuracy of the results and are usually believed to be unique to FFT methods. Actually, many of these problems, such as aliasing, leakage, the singularity of kernel functions at the origin, and the proper handling of mean and point data, are common to all methods using the same data. The problems that are indeed unique to spectral methods only include phase shifting, edge effect or circular convolution, and, sometimes, planar approximation (Sideris 1987). In fact, phase shifting can be very easily corrected by using the time/space shifting property of the Fourier transform. The effect of planar approximation is not significant in most local applications. For regional applications, FFT can also be used on the sphere (Strang van Hees 1990; Vermeer and Forsberg 1992; Forsberg and Sideris 1993; and Haagmans et al. 1993) and it is recommended over the planar approximation.

The method used in digital signal processing to eliminate circular convolution is to append 100% zeros to each row and column of the two data arrays (Brigham 1988). This, however, still does not provide perfect results in our case. Sideris and Li (1992) suggested to append 100% zeros at each row and column of the signal array only, such as the gravity anomalies and the heights, and to compute the kernel function at both the signal-covered and the zero-expanded areas. With this method, FFT spectral techniques provide identical results to those from the rigorous numerical integration. In summary, no additional errors will be brought into the results when spectral techniques are used for the evaluation of gravity field convolutions.

The main drawbacks of FFT-based spectral techniques are that they only take gridded data as input and require much more computer memory. Gridded data, such as gravity anomalies, can be obtained from the irregularly distributed observations. As for any other technique, smoothing of the gravity anomalies is necessary to provide better interpolated results in areas without observations (Heiskanen and Moritz 1967). It is worth mentioning here that Sideris (1995) has developed a hybrid method by which a grid of undulations can be computed from a set of irregularly distributed gravity anomalies. The fact that all signals in physical geodesy are real and the FFT is a complex operation makes half of the computer core memory required by an FFT-based program useless, and at the same time, the complex operations take twice as much time as real operations do (Hartley 1942; Bracewell 1984; Li and Sideris 1992). Nevertheless, judicious use of the Fourier transform properties, or the use of the fast Hartley transform, can overcome these limitations; see Sects. A.5 and A.6 in Appendix A of Chap. 10.

The data types used are more important for the overall accuracy than the methods applied (Schwarz 1984; Mainville et al. 1992). This does not mean that modifications of the methodology are not important. It means, however, that we have to analyze the different data types in view of the resolution they provide for the gravity field spectrum, and to find out which improvements may be needed to most effectively combine data types with different spectral characteristics (Schwarz 1984). From the frequency point of view, most of the conventional methods are restricted to dealing with the long and the medium wavelength information of the gravity field, i.e., the gravity anomalies and geopotential models. To meet, for example, the oceanographic and geophysical requirements and provide a geoid with an absolute accuracy at the cm level and a relative accuracy of better than 1 ppm, more attention has to be paid to dealing with the short wavelength information of the gravity field, such as the detailed topography data. Naturally, to improve long-wavelength errors and biases, the data that the GOCE satellite mission will provide will be indispensible.

It can be thus concluded that spectral geoid determination techniques can be further improved. The objective of current research is to determine the gravimetric geoid with an accuracy of 1 cm. Spectral geoid determination techniques need to be optimized in terms of error propagation and accuracy, computational efficiency and computer memory required. Special attention should be paid to the investigation of the gravimetric geoid accuracy in mountainous areas. Ongoing research is focused on the following aspects: refinement of terrain correction formulas, improvement of the gravity reductions, optimization of geopotential model contributions, minimization of the effect of gravity anomaly errors, as well as increase of the computational efficiency and reduction of the required computer memory.

Appendix A: Definition, Properties and Application of the Fourier Transform

A.1 Basic Definitions

A.1.1 Sinusoids

A real sinusoid of amplitude A, cyclic frequency ω_o and phase angle θ_o is a function of the form:

$$s(t) = A\cos(\omega_o t + \theta_o), \tag{10.110}$$

where t is time or, usually in geodetic applications, distance. The cyclic frequency is related to the period T and the (linear) frequency f_o by the expression

$$\omega_o = 2\pi/T = 2\pi f_o. \tag{10.111}$$

Expanding the cosine term in (10.110) yields

$$s(t) = a\cos\omega_o t + b\sin\omega_o t, \quad a = A\cos\theta_o, \quad b = -A\sin\theta_o. \tag{10.112}$$

which allows for the computation of A and θ_o from the coefficients a and b:

$$A = (a^2 + b^2)^{1/2}, \tag{10.113}$$

$$\theta_o = \arctan(-b/a). \tag{10.114}$$

With i being the imaginary unit, a complex sinusoid has the form

$$s_c(t) = a\cos\omega_o t \pm i a\sin\omega_o t = ae^{\pm i\omega_o t}, \tag{10.115}$$

which can be used to express a real sinusoid as a function of complex sinusoids:

$$s(t) = A\cos(\omega_o t + \theta_o) = A\frac{e^{i(\omega_o t+\theta_o)} + e^{-i(\omega_o t+\theta_o)}}{2} = \frac{A}{2}e^{i\omega_o t}e^{i\theta_o} + \frac{A}{2}e^{-i\omega_o t}e^{-i\theta_o}. \tag{10.116}$$

A.1.2 Fourier Series

If a function $g(t)$ is periodic with period T, i.e., if

$$g(t) = g(t+T); \quad \int_0^T g(t)dt = \int_{t_o}^{t_o+T} g(t)dt, \tag{10.117}$$

then, making use of the orthogonality properties of sine and cosine, $g(t)$ can be expanded into the following series with coefficients a_n and b_n:

$$g(t) = \sum_{n=0}^{\infty} \left[a_n \cos\left(\frac{2\pi n}{T} t\right) + b_n \sin\left(\frac{2\pi n}{T} t\right) \right], \tag{10.118}$$

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \cos\left(\frac{2\pi n}{T} t\right) dt,$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \sin\left(\frac{2\pi n}{T} t\right) dt, n = 0, 1, 2, \ldots., \tag{10.119}$$

provided that $g(t)$ has a finite number of maxima and minima in a period, a finite number of finite discontinuities in a period, and is absolutely integrable over a period (Dirichlet's conditions; see also Sect. A.2.1, Chap. 10).

Making use of (10.115) and (10.116), the above Fourier series expansion can be written in the following complex form:

$$g(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} G_n e^{i\omega_n t}, \omega_n = \frac{2\pi n}{T} = \omega_o n, \tag{10.120}$$

$$G_n = \int_{-T/2}^{T/2} g(t) e^{-i\omega_n t} dt = \frac{1}{2}(a_n - i b_n), n = 0, \pm 1, \pm 2, \ldots., \tag{10.121}$$

which shows that a Fourier expansion decomposes a periodic function into a sum of sinusoids with cyclic frequencies $2\pi n/T$.

Denoting by $\Delta\omega$ the frequency 'spacing' or 'step' $2\pi/T$, we get

$$\Delta\omega = \frac{2\pi}{T}, \quad \omega_n = n\Delta\omega, \quad \frac{1}{T} = \frac{\Delta\omega}{2\pi}, \tag{10.122}$$

and thus (10.120) is finally written as a series, i.e., linear combination, of complex sinusoids:

$$g(t) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} G_n e^{i\omega_n t} \Delta\omega. \tag{10.123}$$

A.2 The Continuous Fourier Transform and Its Properties

A.2.1 Definition of the Continuous Fourier Transform

We give here a heuristic definition of the continuous Fourier transform (CFT), or continuous spectrum, based on the Fourier series. By letting $T \to \infty$, the periodic function $g(t)$ becomes non-periodic. Also, $n \to \infty, \omega_o$ becomes vanishingly small, say $\omega_o = \Delta\omega \to 0$ and $\omega_n = n\Delta\omega \to \omega$. Then at the limit, $T \to \infty, \Delta\omega \to d\omega$ the summation becomes integration, i.e, $G_n = \mathrm{G}(\omega), \sum \Delta\omega = \int d\omega$, and (10.121) and (10.123), respectively, become

$$G(\omega) = \int_{-\infty}^{\infty} g(t)e^{-i\omega t}dt, \tag{10.124}$$

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)e^{i\omega t}d\omega, \tag{10.125}$$

which define the direct and inverse CFT. Since $\omega = 2\pi\mathrm{f}$, the factor $1/2\pi$ can be avoided by expressing the spectrum as a function of f instead of ω as follows:

$$G(f) = \int_{-\infty}^{\infty} g(t)e^{-2\pi i f t}dt = \mathbf{F}\{g(t)\}, \tag{10.126}$$

$$g(t) = \int_{-\infty}^{\infty} G(f)e^{2\pi i f t}df = \mathbf{F^{-1}}\{G(f)\}, \tag{10.127}$$

where $\mathbf{F}$ and $\mathbf{F}^{-1}$ denote the direct and inverse Fourier transform, respectively. The direct and inverse CFT are called a Fourier transform pair and are usually abbreviated as

$$g(t) \leftrightarrow G(f). \tag{10.128}$$

$G(f)$ is, in general, a complex function with real part $G_R(f)$ and imaginary part $G_I(f)$. It thus contains information both about the amplitude $|G(f)|$ and the phase angle $\theta(f)$. Similarly to (10.112)–(10.116), these quantities are

$$G(f) = G_R(f) + iG_I(f) = |G(f)|\, e^{i\theta(f)}, \tag{10.129}$$

$$|G(f)| = [G_R^2(f) + G_I^2(f)]^{1/2}, \tag{10.130}$$

$$\theta(f) = Arg\{G(f)\} = \arctan\frac{G_I(f)}{G_R(f)}. \tag{10.131}$$

$G(f)$ exists when $g(t)$ is absolutely integrable, i.e., the integral of $|g(t)|$ from $-\infty$ to ∞ exists (is $< \infty$), and $g(t)$ has only finite discontinuities. If $g(t)$ is periodic or impulse, $G(f)$ does not exist unless the theory of distributions is introduced. This leads to the definition of the impulse function that is given below.

A.2.2 The Impulse Function

The unit impulse or Dirac delta function $\delta(t)$ is usually defined by the relationships

$$\delta(t - t_o) = 0, t \neq t_o; \int_{-\infty}^{\infty} \delta(t - t_o)dt = 1. \tag{10.132}$$

Other definitions are based on treating the impulse function as a distribution or a generalized limit of a sequence of functions. An alternative definition is

$$\delta(t) = \lim_{a \to 0} f(t, a), \tag{10.133}$$

where $f(t, a)$ is a function in a series of functions that progressively increase in amplitude, decrease in duration, and have a constant area of unit; see Fig. 10.7. Using $f(t, a) = \sin(at)/\pi t$, the following expression for $\delta(t)$ is obtained (Papoulis 1977, 1984), which is of importance in evaluating the otherwise non-existent CFT of periodic and other particular functions:

$$\int_{-\infty}^{\infty} \cos(2\pi f t)df = \int_{-\infty}^{\infty} e^{i2\pi f t}dt = \delta(t). \tag{10.134}$$

As an example, the CFT of the sinusoid function of (10.110) will be derived for $\theta_o = 0$; see Fig. 10.8. Equation 10.126, using (10.111), (10.117) and (10.134) gives

$$\begin{aligned} S(f) &= \mathbf{F}\{s(t)\} = \int_{-\infty}^{\infty} A\cos(2\pi f_o t)e^{-i2\pi f t}dt \\ &= \frac{A}{2}\int_{-\infty}^{\infty} (e^{i2\pi f_o t} + e^{-i2\pi f_o t})e^{-i2\pi f t}dt \\ &= \frac{A}{2}\int_{-\infty}^{\infty} (e^{-i2\pi(f - f_o)t} + e^{-i2\pi(f + f_o)t})dt \end{aligned}$$

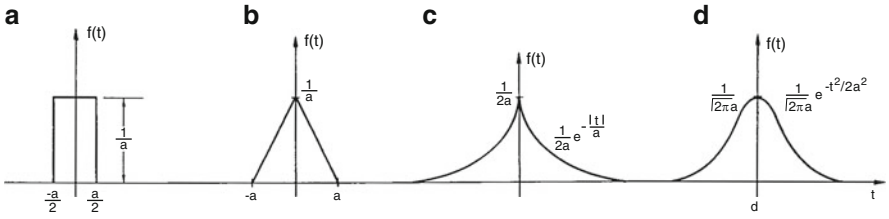


Fig. 10.7 The impulse function as the limit of a function sequence

$$= \frac{A}{2}\delta(f - f_o) + \frac{A}{2}\delta(f + f_o), \tag{10.135}$$

$$A\cos(2\pi f_o t) \leftrightarrow \frac{A}{2}\delta(f - f_o) + \frac{A}{2}\delta(f + f_o).$$

Similarly, for the sine function (see Fig. 10.8), it can be proven that

$$A\sin(2\pi f_o t) \leftrightarrow i\frac{A}{2}\delta(f + f_o) - i\frac{A}{2}\delta(f - f_o). \tag{10.136}$$

Important properties of the impulse function are listed below:

$$\delta(t_o)h(t) = h(t_o)\delta(t_o) \tag{10.137}$$

$$\int_{-\infty}^{\infty} \delta(t - t_o)h(t)dt = h(t_o), \tag{10.138}$$

$$\delta(at) = |a|^{-1}\,\delta(t), \tag{10.139}$$

$$\mathbf{F}\{a\delta(t)\} = a, \tag{10.140}$$

$$\Delta(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \leftrightarrow \Delta(f) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right). \tag{10.141}$$

The last expression describes a sequence of impulse functions, sometimes called 'comb' function, which repeat at intervals T in the time (space) domain and $1/T$ in the frequency domain. The multiplication of any continuous function with $\Delta(t)$ produces *digitization*. Thus, $\Delta(t)$ is very important for *sampling* and for deriving formulas for the discrete Fourier transform from those for the continuous Fourier transform.

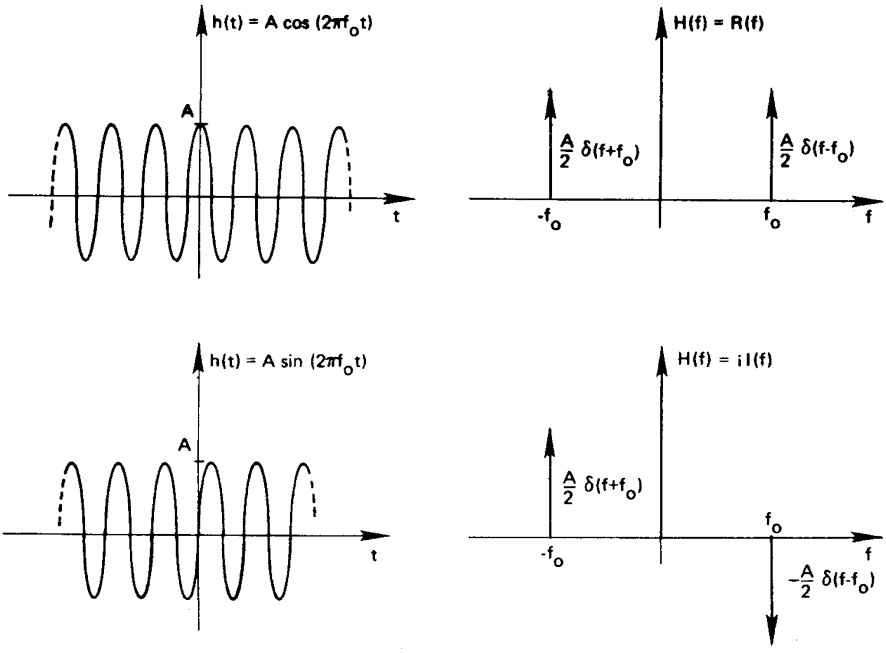


Fig. 10.8 The Fourier transform pairs of the cosine and the sine function

A.2.3 The Rectangle and the Sinc Functions

Also important for deriving formulas for the discrete Fourier transform from those for the continuous Fourier transform are the rectangle and the sinc functions, which actually form a Fourier transform pair. The rectangle function of base T_o and amplitude A is defined as follows:

$$\Pi(t) = \begin{cases} A, & |t| = T_0/2 \\ A/2, & t = \pm T_0/2 \\ 0, & |t| > T_0/2 \end{cases} . \tag{10.142}$$

The sinc function, which is very important in interpolation problems, is defined as

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t}, \tag{10.143}$$

and the Fourier transform pair (see Fig. 10.9) is

$$\Pi(t) \leftrightarrow 2AT_o\mathrm{sinc}(2T_o f). \tag{10.144}$$

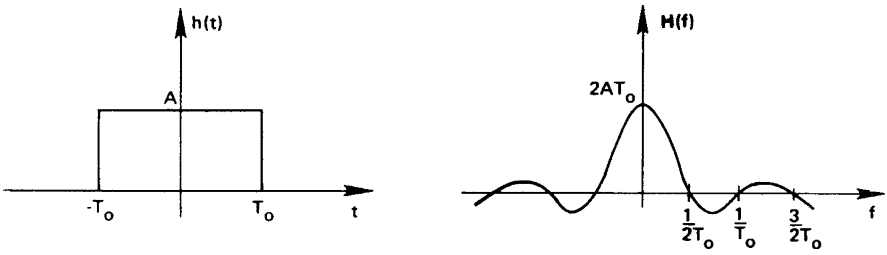


Fig. 10.9 The rectangle function and its Fourier transform, the sinc function

A.2.4 Interpretation of the Fourier Transform and the Fourier Series

Equation 10.120 indicates that a periodic function can be represented as a sum of harmonics of amplitudes G_n and cyclic frequencies ω_n, with fundamental frequency ω_o. Comparing (10.125) to (10.120), $G(\omega)d\omega/2\pi$ can be viewed as the infinitesimal magnitude of a 'harmonic' with cyclic frequency ω. These 'harmonics' have zero fundamental frequency ($\omega_o \rightarrow d\omega$) and are 'spaced' infinitesimally far apart. In other words, a non-periodic function can be represented as a sum of exponentials (harmonics) with fundamental frequency tending to zero!

From the above interpretation, and also from (10.129) to (10.131), it is clear that the Fourier transform contains information regarding the amplitude and the phase of the 'harmonics' that constitute the function. This becomes easily apparent in the examples of Fig. 10.8, where the spectra show both the amplitude and frequency of the sine and cosine functions and the fact that they have a phase difference of $\pi/2$ [recall that $\cos(-x) = \cos x$ while $\sin(-x) = -\sin(x)$]. Basically, we assume that any given function has two equivalent representations: one in the time (or space) domain and another one in the frequency domain. Equation 10.126 analyzes the time (space) function into a frequency spectrum (in terms of magnitude and phase or, equivalently, in terms of real and imaginary part; see 10.129), while (10.127) synthesizes the frequency spectrum to regain the time (space) function. Equation 10.130 gives the magnitude spectrum while (10.131) gives the phase spectrum of the function.

A.2.5 Properties of the CFT

The following properties are listed here without proof. The proofs, based directly on the definition equations of the CFT, can be found in Brigham (1988).

$$ah(t) + bg(t) \leftrightarrow aH(f) + bG(f) \quad \text{Linearity} \quad (10.145)$$

$$H(t) \leftrightarrow h(-f) \quad \text{Symmetry} \quad (10.146)$$

$$h(at) \leftrightarrow \frac{1}{|a|} H\left(\frac{f}{a}\right) \quad \text{Time scaling} \quad (10.147)$$

$$h(t - t_o) \leftrightarrow H(f) e^{-i2\pi f t_o} \quad \text{Time shifting} \quad (10.148)$$

$$\frac{\partial^n h(t)}{\partial t^n} \leftrightarrow (i2\pi f)^n H(f) \quad \text{Differentiation} \quad (10.149)$$

$$\int_{-\infty}^{t} h(x)dx \leftrightarrow \frac{1}{i2\pi f} H(f) + \frac{1}{2} H(0)\delta(f) \quad \text{Integration} \quad (10.150)$$

$$\int_{-\infty}^{\infty} h(t)dt = H(0) \quad \text{DC} - \text{value} \quad (10.151)$$

$$h_E(t) \leftrightarrow H_E(f) = R_E(f) \quad \text{Even function} \quad (10.152)$$

$$h_O(t) \leftrightarrow H_O(f) = iI_O(f) \quad \text{Odd function} \quad (10.153)$$

$$h(t) = h_R(t) \leftrightarrow H(f) = R_E(f) + iI_O(f) \quad \text{Real function} \quad (10.154)$$

$$h(t) = ih_I(t) \leftrightarrow H(f) = R_O(f) + iI_E(f) \quad \text{Imaginary function} \quad (10.155)$$

In the above formulas, R and I stand for the real and imaginary part of H, respectively, and the subscripts E, O, R, I stand for even, odd, real and imaginary function, respectively.

A.2.6 Convolution and Correlation

The convolution and correlation of two functions $g(t)$ and $h(t)$, denoted by $*$ and $\otimes$, respectively, are defined as follows:

$$\begin{aligned} x(t) &= \int_{-\infty}^{\infty} g(\tau)h(t-\tau)d\tau = g(t) * h(t) = h(t) * g(t) \\ &= \int_{-\infty}^{\infty} h(\tau)g(t-\tau)d\tau, \end{aligned} \quad (10.156)$$

$$y(t) = \int_{-\infty}^{\infty} g(\tau)h(t+\tau)d\tau = g(t) \otimes h(t) \neq h(t) \otimes g(t)$$
$$= \int_{-\infty}^{\infty} h(\tau)g(t+\tau)d\tau. \tag{10.157}$$

The most important property of the convolution is that its spectrum is the product of the spectra of the two functions. Similarly, correlation transforms to multiplication of the complex conjugate of the second spectrum, denoted by superscript *, with the spectrum of the first function. These constitute *the convolution theorem* and *the correlation theorem*, respectively, which in abbreviated form are

$$x(t) = g(t) * h(t) \leftrightarrow X(f) = G(f)H(f), \tag{10.158}$$
$$y(t) = g(t) \otimes h(t) \leftrightarrow Y(f) = G(f)H^*(f). \tag{10.159}$$

The process of convolution in the time (space) domain comprises four steps: (1) *folding*, i.e., taking the mirror image of $h(\tau)$ about the ordinate axis; (2) displacement, i.e., shifting $h(-\tau)$ by the amount t; (3) multiplication of $h(t-\tau)$ by $g(\tau)$; and (4) integration, i.e., computation of the area under the product of $h(t-\tau)$ and $g(\tau)$. In correlation, the procedure is the same without the folding step; see Fig. 10.10.

Although this four-step process shows what needs to be done to evaluate a convolution (or a correlation) integral numerically, it really gives no clear 'physical' interpretation of what a convolution is. This, however, becomes rather obvious from the frequency domain representation of convolution. The multiplication of the two spectra indicates that the whole process is nothing else but *filtering* of one of the functions by the other. In other words, regions of the spectrum of one of the functions are either attenuated, or amplified, or otherwise altered according to the shape of the spectrum of the other function. This interpretation is important in the frequency-domain evaluation of gravity field convolution integrals like, e.g., Stokes's integral.

The simple spectral representations of (10.158) and (10.159) are of great practical importance. It is now obvious that instead of computing the tedious convolution and correlation integrals by numerical integration one could evaluate them by multiplication of the spectra and use of the inverse Fourier transform. Two direct and one inverse Fourier transforms are needed in each case, and the process is made clear by the following equations:

$$x(t) = g(t) * h(t) = \mathbf{F}^{-1}\{X(f)\} = \mathbf{F}^{-1}\{G(f)H(f)\}$$
$$= \mathbf{F}^{-1}\{\mathbf{F}\{g(t)\}\mathbf{F}\{h(t)\}\}, \tag{10.160}$$
$$y(t) = g(t) * h(t) = \mathbf{F}^{-1}\{Y(f)\} = \mathbf{F}^{-1}\{G(f)H^*(f)\}$$
$$= \mathbf{F}^{-1}\{\mathbf{F}\{g(t)\}[\mathbf{F}\{h(t)\}]^*\}. \tag{10.161}$$

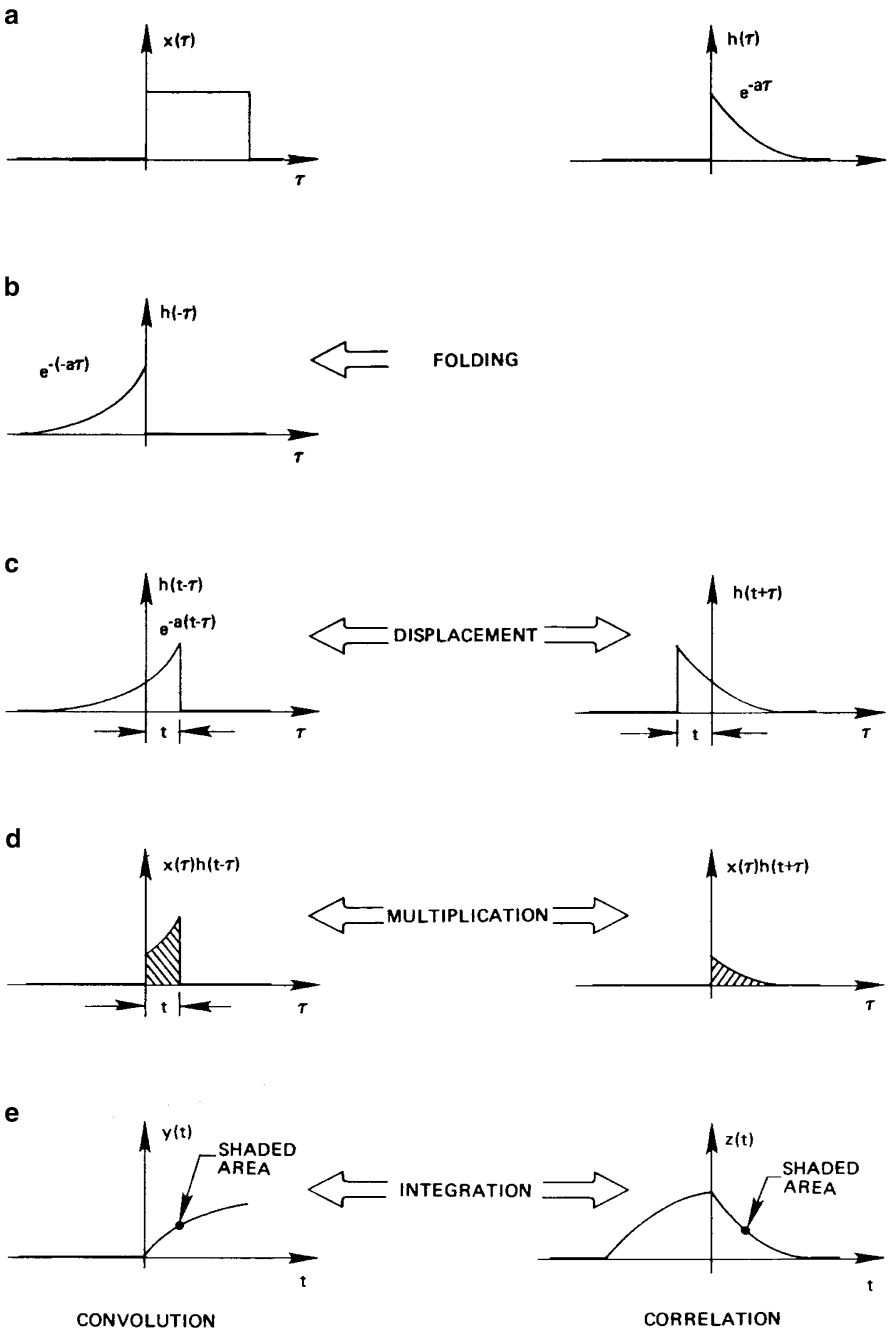


Fig. 10.10 Graphical illustration of time-domain convolution and correlation (After Brigham 1988)

Important properties of convolution are listed below:

$$g(t) * h(t) = g(t) \otimes h(t), \text{if either } g(t) \text{ or } h(t) \text{ is even;} \tag{10.162}$$

$$\delta(t + t_o) * h(t) = h(t + t_o),\ \delta(t) * h(t) = h(t); \tag{10.163}$$

$$\frac{\partial x(t)}{\partial t} = \frac{\partial [g(t) * h(t)]}{\partial t} = \frac{\partial g(t)}{\partial t} * h(t) = g(t) * \frac{\partial h(t)}{\partial t}; \tag{10.164}$$

$$g(t)h(t) \leftrightarrow G(f) * H(f). \tag{10.165}$$

A.3 The Discrete Fourier Transform

A.3.1 From the Continuous to the Discrete Fourier Transform: Aliasing and Leakage

In the practical implementation of the Fourier transform formulas, two approximations are employed: (a) the continuous integrations are replaced by discrete summations and (b) the infinite limits of summation are replaced by finite ones. Obviously, such approximations will introduced errors due to the digitization and the truncation of the series that may or may not be significant depending of the properties of the transformed function. Figure 10.11 illustrates graphically the process of going from the continuous to the discrete Fourier transform (DFT).

First, the function $h(t)$ is sampled or digitized with a sampling interval $\Delta t = T$ by multiplying it with a comb function $\Delta_o(t)$. According to (10.141) and (10.165), this leads to the convolution of the spectrum of $h(t)$ with the spectrum of $\Delta_o(t)$ which is another comb function consisting of impulses at intervals $1/T$. $H(f)*\Delta_o(f)$ is thus a repeating, i.e., periodic, version of the true spectrum. Depending on the value of T, this repetition can cause overlap, which alters the spectrum producing an error caused *aliasing*. The next step is to limit the extent of the function to a finite length, say T_o, containing N sampled points. This is accomplished by multiplying the discretized function by a rectangular function of base T_o and unit height, denoted $x(t)$ in Fig. 10.11, which leads to the multiplication of $H(f)*\Delta_o(f)$ by a sinc function (see 10.142, 10.143 and 10.144). Consequently, another distortion is introduce to the resulting spectrum $H(f) * \Delta_o(f) * X(f)$ called *leakage*. The last step is to discretize the resulting spectrum by multiplying it by a frequency-domain comb function $\Delta_1(f)$ with frequency spacing $\Delta f = 1/T_o$, which of course leads to the repetition of the discretized time (or space) domain function. The DFT is thus periodic in both domains:

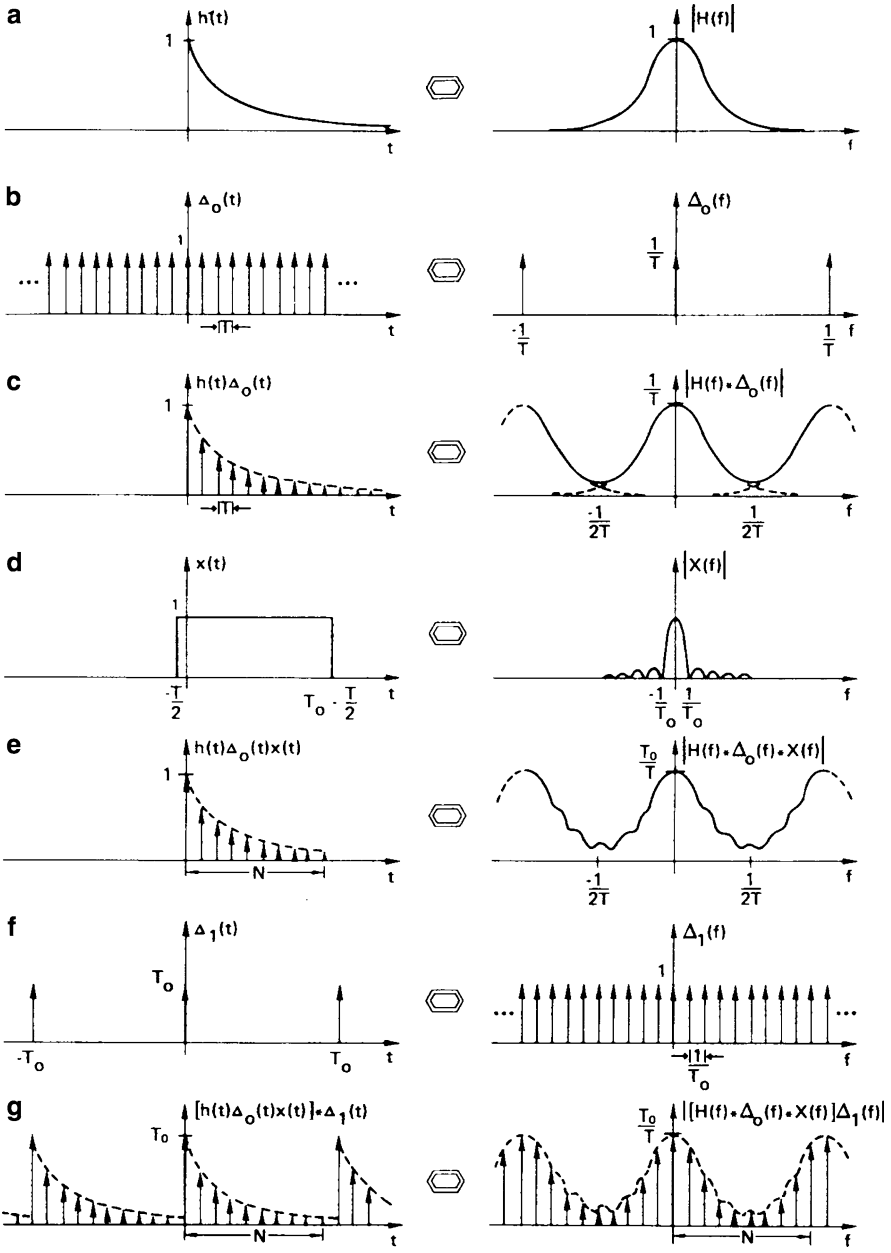


Fig. 10.11 From the continuous to the discrete Fourier transform (After Brigham 1988)

$$H(m\Delta f) = \sum_{j=-\infty}^{\infty} H(m\Delta f + jF), \tag{10.166}$$

$$h(k\Delta t) = \sum_{\ell=-\infty}^{\infty} h(k\Delta t + \ell T), \tag{10.167}$$

and can be defined as follows:

$$H(m\Delta f) = \sum_{k=0}^{N-1} h(k\Delta t) e^{-i2\pi k\Delta t m\Delta f} \Delta t = \sum_{k=0}^{N-1} h(k\Delta t) e^{-i2\pi km/N} \Delta t, \tag{10.168}$$

$$h(k\Delta t) = \sum_{m=0}^{N-1} H(m\Delta f) e^{i2\pi k\Delta t m\Delta f} \Delta f = \sum_{m=0}^{N-1} H(m\Delta f) e^{i2\pi km/N} \Delta f. \tag{10.169}$$

In discrete form, the functions have arguments either their wavelengths $t_k = k\Delta t$ or simply their wavenumbers k in the time (space) domain, and $f_m = m\Delta f$ or simply m in the frequency domain. We will use these representations interchangeably, i.e., we will defined the DFT pair in any one of the following three forms:

$$h(k\Delta t) \leftrightarrow H(m\Delta f) \quad \text{or} \quad h(t_k) \leftrightarrow H(f_m) \quad \text{or} \quad h(k) \leftrightarrow H(m). \tag{10.170}$$

The time period T_o, the frequency period F_o, the time spacing Δt, the frequency spacing Δf and the number of discrete points N are related as follows:

$$T_o = \frac{1}{\Delta f} = N\Delta t, \quad F_o = \frac{1}{\Delta t} = N\Delta f. \tag{10.171}$$

The above equations show that there is a certain maximum frequency (shortest wavelength) and a certain minimum frequency (longest wavelength) that can be recovered from the DFT. Frequencies beyond these limits cannot be recovered due to the aliasing and leakage effects. The maximum frequency that can be recovered is $F_o/2$, depends on Δt, and is called the *Nyquist frequency* f_N. From (10.172)

$$|f_N| = \frac{F_o}{2} = \frac{1}{2\Delta t}. \tag{10.172}$$

The aliasing error, say H_e, can be shown mathematically by rewriting (10.166) as

$$H_P(m\Delta f) = \sum_{j=-\infty}^{\infty} H(m\Delta f + jF) = H(m\Delta f) + \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} H(m\Delta f + jF)$$

$$= H(m\Delta f) + H_e(m\Delta f), \tag{10.173}$$

where a subscript P has been added to the left-hand side to indicate the periodic nature of the DFT. Thus to minimize aliasing, the function must be sampled as densely as possible and to eliminate it Δt should be selected such that $1/2\Delta t$ is larger that the highest frequency present in the data. However, the user cannot always select Δt and minimize aliasing, as is the case when gravity or terrain data are only available on regular grids. In such cases, aliasing can be minimized by removing the high-frequency information from the data by, e.g., applying terrain reductions to gravity anomalies.

The minimum frequency that can be recovered depends on T_o, i.e., on both N and Δt, and is $\Delta f = 1/T_o = 1/N\Delta t$. Given Δt, N should be chosen so that it provides the required frequency resolution Δf. In practice, N is much easier to control than Δt but it will always be a finite number and thus leakage will always be present. From Fig. 10.11e, the altered spectrum due to leakage only will be

$$H'(m\Delta f) = T^{-1}H(m\Delta f) * T_o \mathrm{sinc}(T_o m\Delta f). \tag{10.174}$$

This error does nor occur only when T_o is infinite, i.e., when the Π-function becomes a unit constant from $-\infty$ to ∞. In this case, from (10.140) and (10.146) we obtain $\mathbf{F}\{1\} = \delta(f)$, and (10.174), using (10.163), becomes $H'(m\Delta f) = H(m\Delta f) * \delta(m\Delta f) = H(m\Delta f)$. In practice of course this is not possible and, in order to minimize leakage, the truncation of the infinite function is done by functions other than the Π-function, called *window functions* (Harris 1978). These functions have spectra with smaller side lobes than the sinc function, i.e., their spectra are better approximations to an impulse function than the sinc function is. In gravity field applications, leakage can be minimized by removing the low-frequency information from the data by, e.g., removing the contribution of a global geopotential model from gravity anomalies.

A.3.2 Discrete Convolution and Correlation: Circular Convolution and Correlation

Discretization of (10.146) and (10.157) for both functions given at N points results in the following expressions for discrete convolution and correlation:

$$x(k) = \sum_{l=0}^{N-1} g(l)h(k-l)\Delta t = g(k) * h(k), \tag{10.175}$$

$$y(k) = \sum_{l=0}^{N-1} g(l)h(k+l)\Delta t = g(k) \otimes h(k). \tag{10.176}$$

When these equations are evaluated by numerical summation the results are correct and correspond to linear convolution and linear correlation. If, however, the discrete form of (10.160) and (10.161) are used instead, i.e.,

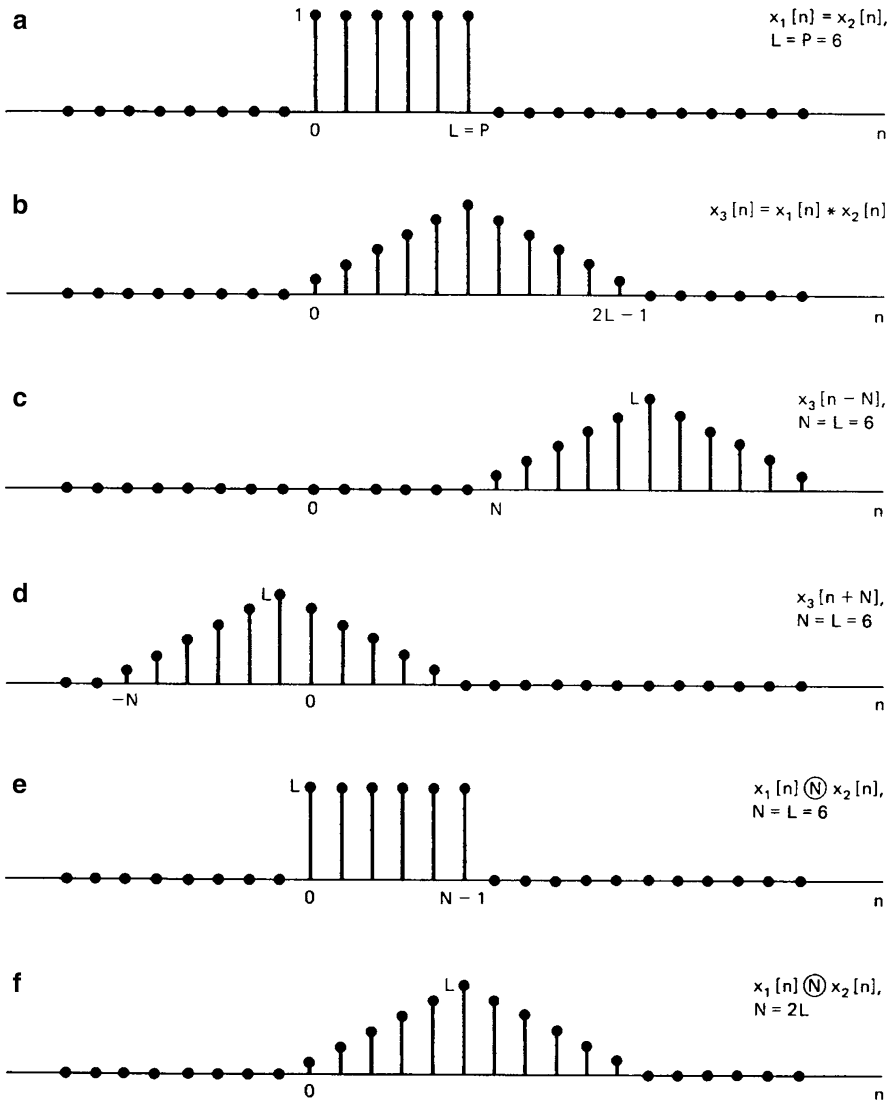


Fig. 10.12 Illustration of circular convolution as linear convolution plus aliasing (After Oppenheim and Schafer 1989)

$$
\begin{aligned}
x_P(k) &= \mathbf{F}^{-1}\{X_P(m)\} = \mathbf{F}^{-1}\{G_P(m)H_P(m)\} \\
&= \mathbf{F}^{-1}\{\mathbf{F}\{g_P(k)\}\mathbf{F}\{h_P(k)\}\}, \qquad (10.177) \\
y_P(k) &= \mathbf{F}^{-1}\{Y_P(m)\} = \mathbf{F}^{-1}\{G_P(m)H_P^*(m)\} \\
&= \mathbf{F}^{-1}\{\mathbf{F}\{g_P(k)\}[\mathbf{F}\{h_P(k)\}]^*\}, \qquad (10.178)
\end{aligned}
$$

both functions are treated as periodic (hence the subscript P), the results are incorrect and correspond to circular convolution and circular correlation. Equations 10.175 and 10.176 indicate that if $g(k)$ and $h(k)$ have N values (or support N) each, then $x(k)$ and $y(k)$ will each have $2N-1$ values. On the other hand, when (10.177) and (10.178) are evaluated by the (periodic) DFT, it is clear that the resulting $x(k)$ and $y(k)$ will each have support N and will be periodic, as well. Mathematically, circular convolution can be viewed as linear convolution contaminated by aliasing (see Fig. 10.12), i.e.,

$$x_P(k) = \sum_{r=-\infty}^{\infty} x(k + rN), \quad 0 \le k \le N-1. \tag{10.179}$$

Circular convolution can be avoided by a procedure called *zero-padding* by which zeros are appended to $g(k)$ and $h(k)$ as follows:

$$g'(k) = \begin{cases} g(k), & 0 \le k \le N \\ 0, & N \le k \le 2N \end{cases}; \qquad h'(k) = \begin{cases} h(k), & 0 \le k \le N \\ 0, & N \le k \le 2N \end{cases}. \tag{10.180}$$

The required steps are: (1) Form $g'(k)$ and $h'(k)$; (2) compute $G'(m)$ and $H'(m)$ via the DFT; (3) compute $X'(m) = G'(m)H'(m)$; and (4) compute $x'(k)$ by applying the inverse DFT to $X'(m)$. Now $x'(k)$ is a $2N-1$ sequence and is exactly the same as $x(k)$ because no aliasing due to overlapping occurs; see again Fig. 10.12. This procedure is the same for computing correlation. For more details on circular convolution and correlation, Oppenheim and Schafer (1989) should be consulted.

A.3.3 Correlation, Covariance, and Power Spectral Density Functions

The discrete correlation function $R_{gh}(t_k)$ of two functions $h(t_k)$ and $g(t_k)$ is defined as

$$R_{gh}(t_k) = \mathbf{E}\{g(t_l)h(t_k - t_l)\} = \lim_{N\to\infty} \frac{1}{N} \sum_{l=0}^{N-1} g(t_l)h(t_k - t_l) = \lim_{T_o\to\infty} \frac{1}{T_o} g(t_k) \otimes h(t_k), \tag{10.181}$$

where we have made use of (10.171) and (10.176) which defines the discrete correlation. When the mean values $\overline{g}$, $\overline{h}$ are subtracted, the formula for the discrete covariance function $C_{gh}(t_k)$ is obtained:

$$C_{gh}(t_k) = \mathbf{E}\{\{g(t_l) - \overline{g}][h(t_k - t_l) - \overline{h}]\} = \lim_{N\to\infty} \frac{1}{N} \sum_{l=0}^{N-1} [g(t_l) - \overline{g}][h(t_k - t_l) - \overline{h}]$$

$$= \lim_{T_o\to\infty} \frac{1}{T_o} g(t_k) \otimes h(t_k) - \overline{g}\,\overline{h} = R_{gh}(t_k) - \overline{g}\,\overline{h}, \tag{10.182}$$

where, using the discrete version of (10.151), $\overline{g}$ (and similarly $\overline{h}$) can be expressed as

$$\overline{g} = \lim_{N\to\infty} \tfrac{1}{N} \sum_{k=0}^{N-1} g(t_k) = \lim_{T_o\to\infty} \tfrac{1}{T_o} G(0). \tag{10.183}$$

When $h(t_k)$ and $g(t_k)$ are the same function, we talk about the auto-covariance and the auto-correlation function. When they are different, we talk about the cross-covariance and the cross-correlation function. The spectrum of the correlation function is called the power spectral density (PSD) function $P_{gh}(f_m)$ and, by (10.178), it has the form

$$P_{gh}(f_m) = \mathbf{F}\{R_{gh}(t_k)\} = \lim_{T_o\to\infty} \frac{1}{T_o} G(f_m) H^*(f_m). \tag{10.184}$$

In practice, of course, we only have a finite number of data and $P_{gh}(f_m)$ is approximated by the biased estimate $\mathbf{F}^{-1}\{G(f_m)H^*(f_m)\}$. If ν records are available each containing N data values, an unbiased estimate for the PSD function is obtained by averaging over all records (Bendat and Piersol 1986) as follows:

$$\hat{P}_{gh}(f_m) = \frac{1}{\nu T_o} \sum_{\lambda=1}^{\nu} G_\lambda(f_m) H_\lambda^*(f_m). \tag{10.185}$$

The normalized standard error ε of $\hat{P}_{gh}$ computed from ν sample records or, more generally, using ν number of averages is

$$\varepsilon = \frac{\sigma(\hat{P}_{gh})}{\hat{P}_{gh}} = \frac{1}{\sqrt{\nu}}, \tag{10.186}$$

where σ denotes the standard error. Thus, 100 averages are required for a 10% error. When only one sample record is available, the estimated PSD is called the *periodogram*. Although very noisy, it might be the only estimate that can be obtained from a single record.

By applying the inverse Fourier transform to the PSD function, an efficient way of estimating correlation and covariance functions of gridded data is obtained:

$$\hat{R}_{gh}(t_k) = \mathbf{F}\{\hat{P}_{gh}(f_m)\}, \tag{10.187}$$

$$\hat{C}_{gh}(t_k) = \mathbf{F}\{\hat{P}_{gh}(f_m) - \overline{g}\,\overline{h}\delta(f_m)\}. \tag{10.188}$$

We end this section by some useful properties of the correlation, covariance, and PSD functions:

$$R_{gh}(-t_k) = R_{hg}(t_k), \quad C_{gh}(-t_k) = C_{hg}(t_k), \tag{10.189}$$

$$R_{gh}(0) = \psi_{gh} = \mathbf{E}[g(t_k)h(t_k)], \quad C_{gh}(0) = \sigma_{gh} = \mathbf{E}[(g(t_k) - \overline{g}) \cdot(h(t_k) - \overline{h})] = \psi_{gh} - \overline{g}\,\overline{h}, \tag{10.190}$$

$$R_{gh}(\infty) = \overline{g}\,\overline{h}, \quad C_{gh}(\infty) = 0, \tag{10.191}$$

$$P_{gh}(-f_m) = P^*_{gh}(f_m) = P_{hf}(f_m), \tag{10.192}$$

$$P_{gh}(0) = T\,\overline{g}\,\overline{h}. \tag{10.193}$$

Note that σ_{gh} is nothing else but the usual covariance while, when $g = h$, σ_{gg} is the variance and ψ_{gg} is the mean square value.

A.3.4 The DFT in Computers

In most computer software for DFTs, such as the FFT subroutines in the IMSL library, the DFT is simply defined by using the wavenumber k instead of the 'wavelength' x_k and also by omitting the period (record length) T_o. This means that the time (space) interval Δt is taken as unit and all other parameters dependent on it are omitted. Thus, in a computer, hence the subscript c, the DFT pair is defined as

$$H_c(m) = \frac{1}{N}\sum_{k=0}^{N-1} h(k)e^{-i2\pi km/N} = \mathbf{F_c}\{h(k)\}, \tag{10.194}$$

$$h_c(k) = \sum_{m=0}^{N-1} H_c(m)e^{i2\pi km/N} = \mathbf{F_c^{-1}}\{H_c(m)\}. \tag{10.195}$$

A comparison of the above equations to (10.168) and (10.169) shows that their exist the following relationships:

$$H_c(m) = \frac{1}{N\Delta x}H(f_m) = \frac{1}{T}H(f_m), \tag{10.196}$$

$$h_c(k) = h(x_k) \tag{10.197}$$

Consequently, when the DFT of $h(x_k)$ is computed by (10.192) the results must be rescaled by T_o to get the correct values. Of practical importance is that

$$H_c(0) = \overline{h} \tag{10.198}$$

and that, for the computation of discrete convolution and correlation, the following relations hold:

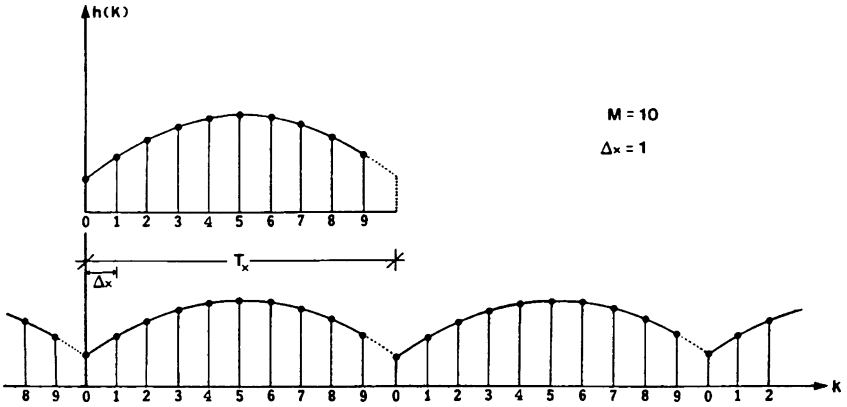


Fig. 10.13 DFT sampling: the end point of a period (After Sideris 1984)

$$x(t_k) = g(t_k) * h(t_k) = T_o x_c(t_k) = T_o \mathbf{F}_c^{-1}\{G_c(m) H_c(m)\}, \quad (10.199)$$

$$y(t_k) = g(t_k) \otimes h(t_k) = T_o y_c(t_k) = T_o \mathbf{F}_c^{-1}\{G_c(m) H_c^*(m)\}. \quad (10.200)$$

Another point that requires attention is the location of the coordinate origin. Usually, computer subroutines consider as origin the first point from the left in both domains. When the points of the sample record are referred to an origin being at the centre of the record, the discrete version of (10.148) must be used to correct the computed spectrum. In such a case, $t_o = N\Delta t/2 = T_o/2$ and thus $e^{-i2\pi m \Delta f T_o/2} = e^{-i\pi m} = \cos(m\pi) = (-1)^m$, which results in

$$h(t_k - T_o/2) \leftrightarrow (-1)^m H(f_m). \quad (10.201)$$

Consequently, when we are after $H(f_m)$, we should multiply the result of the DFT subroutine by $(-1)^m$. Notice that in the product of two spectra obtained by, e.g., (10.199), $(-1)^m$ cancels out. Hence, to avoid the origin shift when we compute convolutions, the product of the two spectra should first be multiplied by $(-1)^m$ and then entered into the inverse DFT subroutine. In the same fashion, special care must be taken for the computation of covariance, correlation and PSD functions.

Finally, in order to avoid extra aliasing errors, no sample should be taken at the end point of the record length. Since the DFT is periodic, the missing end point of a period is considered to be the starting point of the next period. This fact is graphically illustrated in Fig. 10.13, where the function $h(x)$ is sampled at $M = 10$ points per period T_x and is thus represented by the discrete values $h(x_k)$ or simply $h(k)$. It is important to note that the even symmetry of the function is not upset.

A.3.5 The Fast Fourier Transform

The fast Fourier transform (FFT) is an algorithm for computing the DFT much faster (number of required complex multiplications proportional to $N log_2 N$) than by the conventional Fourier transform (number of required complex multiplications proportional to N^2). To illustrate the FFT algorithm, the intuitive development presented in Brigham (1988) for the 1D FFT is explained in the following.

Suppose that the DFT of a function $f(k)$ with $N = 4$ is required. Omitting, for simplicity, the constants in front of the summation symbol, we have

$$H(m) = \sum_{k=0}^{N-1} h(k) e^{-i2\pi km/N} = \sum_{k=0}^{N-1} h(k) W^{km}, \quad m = 0, 1, 2, 3, \tag{10.202}$$

or, equivalently,

$$\begin{pmatrix} H(0) \\ H(1) \\ H(2) \\ H(3) \end{pmatrix} = \begin{pmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{pmatrix} \begin{pmatrix} h(0) \\ h(1) \\ h(2) \\ h(3) \end{pmatrix}. \tag{10.203}$$

Since

$$W^{km} = e^{-i2\pi km/N} = W^{km \bmod(N)}, \tag{10.204}$$

where *km mod(N)* is the remainder of the division of nk by N, (10.203) become

$$\begin{pmatrix} H(0) \\ H(1) \\ H(2) \\ H(3) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & W^1 & W^2 & W^3 \\ 1 & W^2 & W^0 & W^2 \\ 1 & W^3 & W^2 & W^1 \end{pmatrix} \begin{pmatrix} h(0) \\ h(1) \\ h(2) \\ h(3) \end{pmatrix}. \tag{10.205}$$

By bit-reversing the indices of $H(m)$ and by factorizing the matrix of W coefficients into $log_2 N = 2$ matrices, the above system becomes

$$\begin{pmatrix} H(0) \\ H(2) \\ H(1) \\ H(3) \end{pmatrix} = \begin{pmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{pmatrix} \begin{pmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{pmatrix} \begin{pmatrix} h(0) \\ h(1) \\ h(2) \\ h(3) \end{pmatrix}. \tag{10.206}$$

Finally, because $W^2 = -W^0$ and $W^3 = -W^1$, it follows that

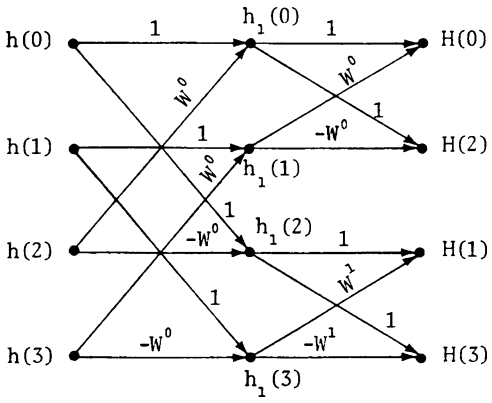


Fig. 10.14 Flow graph of FFT operations for N = 4 (After Sideris 1984)

$$\begin{pmatrix} H(0) \\ H(2) \\ H(1) \\ H(3) \end{pmatrix} = \begin{pmatrix} 1 & W^0 & 0 & 0 \\ 1 & -W^0 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & -W^1 \end{pmatrix} \begin{pmatrix} h_1(0) \\ h_1(1) \\ h_1(2) \\ h_1(3) \end{pmatrix},$$

$$\begin{pmatrix} h_1(0) \\ h_1(1) \\ h_1(2) \\ h_1(3) \end{pmatrix} = \begin{pmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & -W^0 & 0 \\ 0 & 1 & 0 & -W^0 \end{pmatrix} \begin{pmatrix} h(0) \\ h(1) \\ h(2) \\ h(3) \end{pmatrix}. \tag{10.207}$$

From the system of (10.207), a flow graph of operations is constructed and shown in Fig. 10.14.

The matrix factorization introduces zeroes into the sub-matrices and results in an appreciable reduction of multiplications. Figure 10.14 indicates that not only the number of multiplications is reduced but the number of additions is reduced as well, since each $h_1(k)$ is computed only once and then used for the computations of all $H(m)$ in which it takes part. These are the main reasons that the FFT is much faster that the conventional Fourier transform. An extensive discussion of the computational aspects of the FFT algorithm can be found in IEEE (1967) and in Brigham (1988).

A.4 The Two-Dimensional Discrete Fourier Transform

The multi-dimensional continuous Fourier transform pair is defined as follows:

$$G(\underline{f}) = \int_{-\infty}^{\infty} g(\underline{t}) e^{-2\pi i \underline{f}^T \underline{t}} d\underline{t} = \mathbf{F}\{g(\underline{t})\}, \tag{10.208}$$

$$g(\underline{t}) = \int_{-\infty}^{\infty} G(\underline{f}) e^{-2\pi i \underline{f}^T \underline{t}} d\underline{f} = \mathbf{F}^{-1}\{G(\underline{f})\}, \tag{10.209}$$

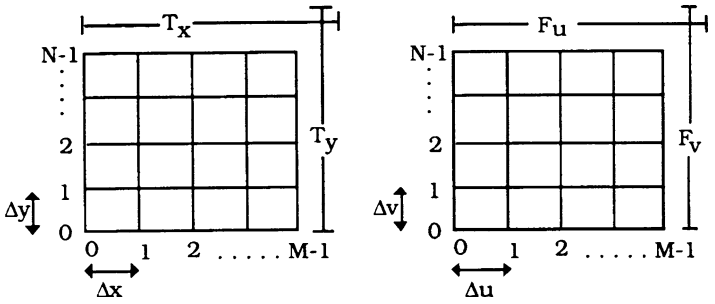


Fig. 10.15 Two-dimensional grids in the space and frequency domain

The vectors $\underline{t}$ and $\underline{f}$ comprise the time (space) and frequency coordinates, respectively. For example, in the case of the three-dimensional (3D) CFT, $\underline{t} = (x, y, z)^T$, $\underline{f} = (u, v, w)^T$, $\underline{f}^T\underline{t} = ux + vy + wz$, $d\underline{t} = dxdydz$ and $d\underline{f} = dudvdw$, where u, v, w are the frequencies corresponding to x, y, z, respectively, and the integrals in (10.208) and (10.209) are triple integrals. Notice that the above definition indicates that the multi-dimensional CFT is separable, i.e. it consists of consecutive applications of the 1D CFT, one for each dimension (or direction), which is of great importance in practical applications. In a similar manner, the properties of the 1D CFT and the convolution and correlation theorems can be extended to many dimensions and will not be repeated here; formulas and more details can be found in Dudgeon and Mersereau (1984), Sideris (1984), Bracewell (1986a), Brigham (1988), Oppenheim and Schafer (1989), and Schwarz et al., (1990). Instead, we will concentrate here on the 2D DFT, which is used in most of the physical geodesy problems.

For a function $h(x_k, y_l)$ given at $M \times N$ gridded points in an area $T_x \times T_y$ with grid spacing Δx and Δy, the two-dimensional discrete Fourier transform pair is

$$H(u_m, v_n) = \sum_{k=0}^{M-1} \sum_{1=0}^{N-1} h(x_k, y_l) e^{-i2\pi(mk/M + nl/N)} \Delta x \Delta y, \qquad (10.210)$$

$$h(x_k, y_l) = \sum_{k=0}^{M-1} \sum_{1=0}^{N-1} H(u_m, v_n) e^{i2\pi(mk/M + nl/N)} \Delta u \Delta v, \qquad (10.211)$$

with (see Fig. 10.15)

$$\Delta u = \frac{1}{T_y} = \frac{1}{M\Delta x}, \qquad \Delta v = \frac{1}{T_y} = \frac{1}{N\Delta y}, \qquad (10.212)$$

$$\Delta x = \frac{1}{F_u} = \frac{1}{M\Delta u} = \frac{1}{2u_N}, \qquad \Delta y = \frac{1}{F_v} = \frac{1}{N\Delta v} = \frac{1}{2v_N}, \qquad (10.213)$$

where u_N and v_N are the Nyquist frequencies corresponding to x and y, respectively. Note that in Fig. 10.15 the end points (here, end row and end column) of a period have not been plotted (see Sect. A.3.5, Chap. 10). Due to the seperability of the 2D Fourier transform, the 2D DFT can be evaluated in computers with limited memory be applying the 1D DFT twice, first along rows and then along columns or vise versa.

A.5 Efficient DFT for Real Functions

In most applications, the data being processed are real but the FFT algorithm is designed for complex functions. Thus, if we only consider a real function, the imaginary part of the algorithm is wasted. In this section, we will give a method to compute the Fourier transform of two real functions via a single DFT and the convolutions of two real functions with the same function simultaneously.

A.5.1 DFT of Two Real Functions Via a Single FFT

If $g(k,l)$ and $h(k,l)$ are two real functions, we can compute their Fourier transform via a single FFT. Let us construct a complex function $y(k,l)$ as the sum of $g(k,l)$ and $h(k,l)$, where one of these is taken to be imaginary, i.e.,

$$y\,(k,l) = g\,(k,l) + i\,h\,(k,l)\,, \tag{10.214}$$

Applying the DFT to (10.214) yields

$$Y(m,n) = \mathbf{F}\{y(k,l)\} = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1}[g(k,l) + i\;h(k,l)]e^{-i2\pi(mk/M+n1/N)}. \tag{10.215}$$

Expanding the right-hand side of (10.215) and denoting $R(m,n)$ and $I(m,n)$ as the real and the imaginary parts of $Y(m,n)$ respectively, we get

$$\begin{aligned} R(m,n) = &\sum_{k=0}^{M-1}\sum_{l=0}^{N-1} g(k,l)\cos 2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right) \\ &- \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} h(k,l)\sin 2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right), \end{aligned} \tag{10.216}$$

$$I(m,n) = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} g(k,l)\sin 2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right) + \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} h(k,l)\cos 2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right), \tag{10.217}$$

and we can easily verify that the DFT of $g(k,l)$ and $h(k,l)$ can be evaluated as

$$G(m,n) = [R(m,n) + R(M-m,N-n)]/2 + i[I(m,n) - I(M-m,N-n)]/2, \tag{10.218}$$

$$H(m,n) = [I(m,n) + I(M-m,N-n)]/2 - i[R(m,n) - R(M-m,N-n)]/2. \tag{10.219}$$

If we divide $R(m,n)$ and $I(m,n)$ into even and odd part as

$$R(m,n) = R_e(m,n) + R_o(m,n), \tag{10.220}$$

$$I(m,n) = I_e(m,n) + I_o(m,n), \tag{10.221}$$

from (10.218) and (10.219) we can see that

$$G(m,n) = R_e(m,n) + iI_o(m,n), \tag{10.222}$$

$$H(m,n) = I_e(m,n) - iR_o(m,n). \tag{10.223}$$

The DFT of the convolution of the two real functions can be directly evaluated from $R(m,n)$ and $I(m,n)$ as

$$\begin{aligned}\mathbf{F}\{g(k,l) * h(k,l)\} = [&R(m,n)I(m,n)\\ &+R(M-m,N-n)I(M-m,N-n)]/2\\ &-i[R^2(m,n) - R^2(M-m,N-n)\\ &-I^2(m,n) + I^2(M-m,N-n)]/2\end{aligned} \tag{10.224}$$

By using this technique, geoid undulations, i.e., (10.23), can be efficiently computed via one direct and one inverse DFT.

A.5.2 Simultaneous Convolution of Two Real Functions with the Same Function

In the computation of terrain corrections, see, e.g., (10.53a), we have to evaluate the convolutions of three real functions with the same kernel function. To save

computer time, two of the three convolutions can be done simultaneously via one convolution as

$$p(k,l) = x(k,l) * y(k,l), \tag{10.225}$$

where $y(k,l)$ is the sum of two real functions as defined by (10.214), $x(k,l)$ represents the kernel function of terrain correction and, as we know, its Fourier transform is an even real function, i.e.

$$X\,(m,n) = Xe\,(m,n)\,. \tag{10.226}$$

The spectrum of $p(k,l)$ is

$$P(m,n) = X(m,n)Y(m,n) = X(m,n)[G(m,n) + iH(m,n)]. \tag{10.227}$$

Considering (10.222), (10.223) and (10.226), we get

$$X(m,n)G(m,n) = X_e(m,n)R_e(m,n) + iX_e(m,n)I_o(m,n), \tag{10.228}$$

$$X(m,n)H(m,n) = X_e(m,n)I_e(m,n) - iX_e(m,n)R_o(m,n). \tag{10.229}$$

By using the properties of the Fourier transform of even and odd functions, we can verify that $\mathbf{F}^{-1}\{X(m,n)G(m,n)\}$ is a real function and equal to the real part of $\mathbf{F}^{-1}\{P(m,n)\}$, $i\mathbf{F}^{-1}\{X(m,n)H(m,n)\}$ is an imaginary function and equal to the imaginary part of $\mathbf{F}^{-1}\{P(m,n)\}$, i.e.,

$$x(k,l) * g(k,l) = \mathbf{F}^{-1}\{X(m,n)G(m,n)\} = real\{\mathbf{F}^{-1}\{P(m,n)\}\}, \tag{10.230}$$

$$x(k,l) * h(k,l) = \mathbf{F}^{-1}\{X(m,n)H(m,n)\} = imag\{\mathbf{F}^{-1}\{P(m,n)\}\}. \tag{10.231}$$

A.6 Use of the Fast Hartley Transform

FFT-based spectral techniques, as standard and indispensable tools for the evaluation of gravity field convolutions, make it possible to perform the computations, such as geoid undulations and terrain reductions, etc., in a large area simultaneously. However, the fact that all signals are real and the FFT is a complex operation makes half of the computer core memory required by an FFT-based program useless, and the complex mathematical operations, such as addition and multiplication, take twice as much time as real operations. To avoid such shortcomings related to the FFT method, this chapter will introduce the use of the fast Hartley transform, and discuss other methods of efficient FFT convolutions.

A.6.1 The Discrete Hartley Transform

Hartley (1942) proposed the use of a new kind of transform that is expressed in a more symmetrical form between the function of the original real variable and its transform, which forms the basis for the present Hartley transform and the fast Hartley transform (FHT). The FHT is as fast as or faster than the FFT, and serves for all the uses, such as the convolution operations and spectral analysis, to which the FFT is at present applied.

This section will discuss the properties of the discrete Hartley transform, and show how the gravity convolutions can be performed by FHT. For more details about the basic principles of the Hartley transform, Hartley (1942) and Bracewell (1986a) can be consulted. For the applications of the fast Hartley transform in physical geodesy, Li and Sideris (1992) is recommended.

A.6.2 Definition of the 1D Discrete Hartley Transform

Hartley (1942) defined a more symmetrical one-dimensional Fourier transform as follows:

$$H(\omega) = \int_{-\infty}^{\infty} h(t)\ \mathrm{cas}2\pi\omega t\ dt, \tag{10.232}$$

$$h(t) = \int_{-\infty}^{\infty} H(\omega)\ \mathrm{cas}2\pi\omega t\ d\omega, \tag{10.233}$$

where

$$\mathrm{cas}\, x = \cos x + \sin x. \tag{10.234}$$

Because the transform pair (10.232) and (10.233) was first defined by Hartley, (10.232) is called the direct Hartley transform and (10.232) is called the inverse Hartley transform (Bracewell 1984, 1986b).

For a real function $h(k)$ given at M gridded points with grid spacing Δx, the one-dimensional discrete Hartley transform pair is defined as

$$H(m) = \Delta x \sum_{k=0}^{M-1} h(k\Delta x)\ \mathrm{cas}\frac{2\pi mk}{M}, \tag{10.235}$$

$$h(k) = \frac{1}{M\Delta x} \sum_{m=0}^{M-1} H(m\Delta u)\ \mathrm{cas}\frac{2\pi mk}{M}, \tag{10.236}$$

where $\Delta u = 1/(M\Delta x)$.

A.6.3 Definition of the 2D Discrete Hartley Transform

For a real function $h(k,l)$ given at $M \times N$ gridded points with grid spacing Δx and Δy, the two-dimensional discrete Hartley transform pair is defined as

$$H(m\Delta u, n\Delta v) = \Delta x \Delta y \sum_{k=0}^{M-1} \sum_{1=0}^{N-1} h(k\Delta x, 1\Delta y)\ \mathrm{cas}\frac{2\pi mk}{M}\ \mathrm{cas}\frac{2\pi nl}{N}, \quad (10.237)$$

$$h(k\Delta x, 1\Delta y) = \frac{1}{M\Delta x N\Delta y} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} H(m\Delta u, n\Delta v)\mathrm{cas}\frac{2\pi mk}{M}\ \mathrm{cas}\ \frac{2\pi nl}{N}, \quad (10.238)$$

where

$$\Delta u = \frac{1}{M\Delta x}, \quad \Delta v = \frac{1}{N\Delta y} \quad (10.239)$$

For convenience, (10.237) and (10.238) can be simply expressed as

$$H(m,m) = \Delta x \Delta x \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} h(k,l)\ \mathrm{cas}mk\ \mathrm{cas}nl, \quad (10.240)$$

$$h(k,l) = \frac{1}{T_x T_y} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} H(m,n)\ \mathrm{cas}mk\ \mathrm{cas}nl. \quad (10.241)$$

The Hartley transform pair is denoted as

$$h\,(k,1) \Leftrightarrow H\,(m,n)\,. \quad (10.242)$$

Similar to the discrete Fourier transform, very efficient operations can be developed for the evaluation of the Hartley transform, which results in the fast Hartley transform.

A.6.4 Properties of the Discrete Hartley Transform

The following properties of the two-dimensional discrete Hartley transform can be derived directly from the definition and, therefore, most of them are listed below without proof.

(a) ***Linearity***

$$a\ h(k,1) + b\ g(k,1) \Leftrightarrow a\ H(m,n) + b\ G(m,n). \quad (10.243)$$

(b) ***Spacing shifting***

$$h(k-\lambda, l-\mu) \Leftrightarrow H(m,n)\cos m\lambda \cos n\mu - H(m,-n)\cos m\lambda \sin n\mu \\ -H(-m,n)\sin m\lambda \cos n\mu + H(-m,-n)\sin m\lambda \sin n\mu. \tag{10.244}$$

If $\lambda = M/2$ and $\mu = N/2$, then

$$h(k-M/2, l-N/2) \Leftrightarrow (-1)^{m+n} H(m,n). \tag{10.245}$$

(c) ***Even function***

If $h(k,l)$ is even with respect to the two variables, i.e.,

$$h_e(k,l) = h_e(-k,l) = h_e(k,-l), \tag{10.246}$$

then,

$$h_e(k,l) \Leftrightarrow H_e(m,n) = \Delta x\ \Delta y \sum_{k=0}^{M-1}\sum_{1=0}^{N-1} h_e(k,l)\cos mk\ \cos nl, \tag{10.247}$$

therefore,

$$H_e(m,n) = H_e(-m,n) = H_e(m,-n) = H_e(-m,-n). \tag{10.248}$$

(d) ***Odd function***

If $h(k,l)$ is odd with respect to the two variables, i.e.,

$$\begin{aligned} &h_o(k,l) = -h_o(-k,l) = -h_o(k,-l), \\ &h_o(0,l) = h_o(M/2,l) = h_o(k,0) = h_o(k,N/2) = 0, \end{aligned} \tag{10.249}$$

then,

$$h_o(k,l) \Leftrightarrow H_o(m,n) = \Delta x \Delta y \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} h_o(k,l)\sin mk\ \sin nl, \tag{10.250}$$

therefore,

$$H_o(m,n) = -H_o(-m,n) = -H_o(m,-n) = H_o(-m,-n). \tag{10.251}$$

(e) ***Odd-even function***

If $h(k,l)$ is odd with respect to one variable and even with respect to the other, i.e.,

$$h_{oe}(k,l) = -h_{oe}(-k,l) = h_{oe}(k,-l) \text{ and } h(0,l) = h(M/2,l) = 0, \quad (10.252)$$

then,

$$h_{oe}(k,l) \Leftrightarrow H_{oe}(m,n) = \Delta x\ \Delta y \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} h_{oe}(k,l) \sin\, mk\, \cos nl, \quad (10.253)$$

therefore,

$$H_{oe}(m,n) = -H_{oe}(-m,n) = H_{oe}(m,-n) = -H_{oe}(-m,-n). \quad (10.254)$$

(f) ***Two-dimensional convolution theorem***

$$\begin{aligned} h(k,1) * g(k,1) \Leftrightarrow\ & G(m,n)H_1(m,n) + G(-m,-n)H_2(m,n) \\ & + G(-m,n)H_3(m,n) + G(m,-n)H_4(m,n), \end{aligned} \quad (10.255)$$

where

$$\begin{aligned} H_1(m,n) &= \frac{1}{4}[H(m,n) + H(-m,-n) + H(m,-n) + H(-m,n)], \\ H_2(m,n) &= \frac{1}{4}[H(m,n) + H(-m,-n) - H(m,-n) - H(-m,n)], \\ H_3(m,n) &= \frac{1}{4}[H(m,n) - H(-m,-n) + H(m,-n) - H(-m,n)], \\ H_4(m,n) &= \frac{1}{4}[H(m,n) - H(-m,-n) - H(m,-n) + H(-m,n)]. \end{aligned} \quad (10.256)$$

If $h(k,l)$ is even, the convolution theorem simplifies to

$$h_e(k,l) * g(k,l) \Leftrightarrow H(m,n)\, G(m,n). \quad (10.257)$$

If $h(k,l)$ is odd, the convolution theorem simplifies to

$$h_o(k,l) * g(k,l) \Leftrightarrow H(m,n)\, G(-m,-n). \quad (10.258)$$

If $h(k,l)$ is odd in k and even in l, the convolution theorem simplifies to

$$h_{oe}(k,l) * g(k,l) \Leftrightarrow H(m,n)\, G(-m,n). \quad (10.259)$$

Proof.
The convolution of $h(k,l)$ with $g(k,l)$ is defined by

$$f(k,l) = h(k,l) * g(k,l) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) g(k-i, l-j) \tag{10.260}$$

and the Hartley transform of $f(k,l)$ is

$$\begin{aligned} F(m,n) &= \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \left(\sum_{i=0}^{M-1} \sum_{j=o}^{N-1} h(i,j) g(k-i, l-j) \right) cas\ mk\ \ cas\ nl \\ &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} g(k,l) \cos m(k+i) \cos n(l+j). \end{aligned} \tag{10.261}$$

With the following identity,

$$\cos(x+y) = \cos x \cos y + \sin(-x) \sin y, \tag{10.262}$$

$F(m,n)$ can be expressed as

$$\begin{aligned} F(m,n) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) [& G(m,n) \cos mi \sin nj + G(-m,-n) \sin mi \sin nj \\ & + G(-m,n) \sin mi \cos nj + G(m,-n) \cos mi \sin nj]. \end{aligned} \tag{10.263}$$

With the following notations,

$$\begin{aligned} H_1(m,n) &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) \cos mi\ \cos nj \\ H_2(m,n) &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) \sin mi\ \sin nj \\ H_3(m,n) &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) \sin mi\ \cos nj \\ H_4(m,n) &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} h(i,j) \cos mi\ \sin nj \end{aligned} \tag{10.264}$$

(10.263) becomes

$$F(m,n) = G(m,n)H_1(m,n) + G(-m,-n)H_2(m,n) + G(-m,n)H_3(m,n) + G(m,-n)H_4(m,n), \tag{10.265}$$

which results in the convolution theorem as expressed in (10.255).

With the equation

$$H(m,n) = H_1(m,n) + H_2(m,n) + H_3(m,n) + H_4(m,n), \tag{10.266}$$

the convolution theorem can be simplified as follows.

If $h(k,l)$ is an even function, then

$$H(m,n) = H_1(m,n) \text{ and } H_2(m,n) = H_3(m,n) = H_4(m,n) = 0. \tag{10.267}$$

Inserting (10.267) into (10.266) gives the Hartley transform pair of (10.257).

If $h(k,l)$ is an odd function, then

$$H(m,n) = H_2(m,n) \text{ and } H_1(m,n) = H_3(m,n) = H_4(m,n) = 0. \tag{10.268}$$

Combining (10.268) with (10.266) yields the Hartley transform pair (10.258).

If $h(k,l)$ is an odd in k and even in l, then

$$H(m,n) = H_3(m,n) \text{ and } H_1(m,n) = H_2(m,n) = H_4(m,n) = 0, \tag{10.269}$$

and, consequently, (10.266) results in the Hartley transform pair of (10.259). □

(g) ***Cross correlation***

$$h(k,l) \otimes g(k,1) \Leftrightarrow G(m,n)H_1(m,n) + G(-m,-n)H_2(m,n) - G(-m,n)H_3(m,n) - G(m,-n)H_4(m,n). \tag{10.270}$$

If $g(k,l)$ is an even function, then

$$h(k,l) \otimes g_e(k,1) \Leftrightarrow G(m,n)H(-m,-n). \tag{10.271}$$

If $h(k,l)$ is an even function, then

$$h_e(k,l) \otimes g(k,1) \Leftrightarrow G(m,n)H(m,n). \tag{10.272}$$

(h) ***DC value***

$$H(0,0) = \frac{T_x T_y}{MN} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} h(k,1) = T_x T_y \mu_h, \tag{10.273}$$

$$h(0,0) = \frac{1}{T_x T_y} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} H(m,n), \tag{10.274}$$

where μ_h is the mean value of $h(k,l)$.

(i) ***The quadratic content theorem***

$$\frac{T_x T_y}{MN} \sum_{k=0}^{M-1} \sum_{1=0}^{N-1} [h(k,1)]^2 = \frac{1}{T_x T_y} \sum_{k=0}^{M-1} \sum_{1=0}^{N-1} [H(m,n)]^2. \tag{10.275}$$

A.7 Relationship Between the DHT and the DFT

A.7.1 Computation of the 1D DFT Via the 1D DHT

Comparing the definition of the one-dimensional discrete Fourier transform

$$H^F(m) = \frac{T_x}{M} \sum_{k=0}^{M-1} h(k) \left[\cos \frac{2\pi mk}{M} - j \sin \frac{2\pi mk}{M} \right], \tag{10.276}$$

and that of the one-dimensional discrete Hartley transform

$$H(m) = \frac{T_x}{M} \sum_{k=0}^{M-1} h(k) \left[\cos \frac{2\pi mk}{M} + \sin \frac{2\pi mk}{M} \right], \tag{10.277}$$

we can see that

$$real\,(H^F(m)) = [H(m) + H(-m)]/2, \tag{10.278a}$$

$$imag\,(H^F(m)) = [H(m) - H(-m)]/2. \tag{10.278b}$$

Equation 10.278 indicates that the real and the imaginary parts of the one-dimensional discrete Fourier transform are equal to the even and the odd parts of the discrete Hartley transform, respectively. If $h(k)$ is an even function, considering that $H(m) = H(-m)$, (10.278) can be simplified as

$$real\,(H^F(m)) = H(m), \tag{10.279a}$$

$$imag\,(H^F(m)) = 0. \tag{10.279b}$$

On the other hand, if $h(k)$ is an odd function, with the relation $H(m) = -H(-m)$, (10.278) becomes

$$real(H^F(m)) = 0, \tag{10.280a}$$

$$imag(H^F(m)) = H(m). \tag{10.280b}$$

When the power spectrum is the desired product, it may be obtained directly from the DHT without first calculating the real and the imaginary part of the DFT as in the usual way of calculating power spectrum, i.e.

$$[H^F(m)]^2 = [H(m)]^2. \tag{10.281}$$

A.7.2 Computation of the 2D DFT Via the 2D DHT

The real and the imaginary part $R(m,n)$ and $I(m,n)$ of the two-dimensional discrete Fourier transform are

$$R(m,n) = \frac{T_x T_y}{M\,N} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} h(k,l) \cos\left(\frac{2\pi mk}{M} + \frac{2\pi nl}{N}\right) \tag{10.282a}$$

$$I(m,n) = \frac{T_x T_y}{M\,N} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} h(k,l) \sin\left(\frac{2\pi mk}{M} + \frac{2\pi nl}{N}\right) \tag{10.282b}$$

Compared with (10.264) and (10.26), and (10.282) becomes

$$R(m,n) = H_1(m,n) - H_2(m,n) = [H(m,-n) + H(-m,n)]/2, \tag{10.283a}$$

$$I(m,n) = H_3(m,n) + H_4(m,n) = [H(m,n) - H(-m,-n)]/2. \tag{10.283b}$$

If $h(k,l)$ is an even function, with (10.267) and (10.282) becomes

$$R(m,n) = H(m,n), \tag{10.284a}$$

$$I\,(m,n) = 0. \tag{10.284b}$$

If $h(k,l)$ is an odd function, with (10.268) and (10.282) becomes

$$R\,(m,n) = -H\,(m,n)\,, \tag{10.285a}$$

$$I\,(m,n) = 0. \tag{10.285b}$$

If $h(k,l)$ is an odd in k and even in l, with (10.269), and (10.282) becomes

$$R(m,n) = 0, \tag{10.286a}$$

$$I(m,n) = H(m,n). \tag{10.286b}$$

So, the real and imaginary parts of the two-dimensional discrete Fourier transform can be easily computed via the discrete Hartley transform.

A.7.3 Advantages Unique to the FHT

The Hartley transform is superior to the Fourier transform with respect to the requirements in both computer time and computer memory. The Hartley transform is symmetric according to the transformation formula and its inverse. The transformation kernel (cas-function) is real; i.e., the Hartley spectrum of a real signal is also real. So, using the Hartley transform instead of the Fourier transform, we can save half of the computer core memory, or, for the same computer system, the Hartley transform can handle an amount of data twice as large as that handled by the Fourier transform. This is very important if we want to compute a large area of geoid undulations or terrain corrections simultaneously.

These properties have led to the use of the Hartley transform for time-efficient Fourier analysis of real signals. For a data length N being an integer power of 2, i.e., $N = 2^n$, the FHT algorithm can be developed in just the same way as the FFT algorithm. As the FHT uses only real operations, it is about twice as fast as the FFT. In practice, typically only 20–30% of the total execution time is consumed in butterfly execution, and the remainder is spent in interpretation, indexing, etc. (Bold 1985). The computer time saved by the FHT may be less than 50% and about one-third. It should be mentioned that very time-efficient real FFT algorithms are available now, but the programming effort increases with increasing speed of computation.

Chapter 11
Combination of Heights

G. Fotopoulos

11.1 Outline of the Chapter

This chapter provides a practical discussion on the combination of heterogeneous height data. Since most of the theory is discussed in detail in the previous chapters, only a brief introduction to the theoretical issues is included along with some insight into why combining geoid, orthometric and ellipsoidal height data is relevant and important on both regional and global scales. The next section is devoted to a detailed outline of a computational methodology that can be used for the optimal combination of heterogeneous height data (via least-squares adjustment). From this approach it is evident that two key elements deserve more attention, in particular the individual accuracy contributions of each of the height types (via variance component estimation) and the role of the parametric model which appears in the general linear functional model used in the adjustment. Modelling options are provided and more importantly an approach for the assessment of selected models is described in detail. These techniques are supported by numerical examples with real data sets (in Canada and Switzerland). Finally, it should be noted that this chapter is an abridged version of Fotopoulos (2003) and readers who are interested in implementing the methodologies shown herein are encouraged to view the aforementioned manuscript for further details.

11.2 Introduction

Observed elevation differences between points on the Earth's surface are traditionally obtained through spirit-levelling (and/or its variants such as trigonometric, barometric levelling, etc.). For over a century the vertical control needs of the geodetic, cartographic, surveying, oceanographic and engineering communities have been well served by this technique. Due to the nature and practical limitations

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_11,

of spirit-levelling most vertical control points are located in valleys and along roads/railways, which restricts the spatial resolution of control networks and confines the representation of the actual terrain. On the other hand, horizontal control networks have historically been established using triangulation and trilateration, which required that points be situated on hilltops or high points (Davis et al. 1981). As a result, most countries have completely separate networks for horizontal and vertical control with few overlapping points. However, with the advent of satellite-based global positioning systems (GPS, GLONASS, upcoming systems such as GALILEO) and space-borne/airborne radar systems (satellite altimetry, LiDAR, InSAR) the ability to obtain accurate heights (and/or height differences/changes) at virtually any point on land or at sea has in fact been revolutionized.

The fundamental relationship, to first approximation, that binds the ellipsoidal heights obtained from global navigation satellite system (GNSS) measurements and heights with respect to a vertical geodetic datum established from spirit-levelling and gravity data as introduced in Chap. 1 is

$$h - H - N = 0 \tag{11.1}$$

where h is the ellipsoidal height, H is the orthometric height and N is the geoidal undulation obtained from a regional gravimetric geoid model or a global geopotential model. The geometrical relationship between the triplet of height types is also illustrated in Fig. 2.2.

For the relative case, where height differences between two points are considered, we simply have

$$\delta h - \delta H - \delta N = 0 \tag{11.2}$$

where δh, δH, and δN refer to the ellipsoidal, orthometric and geoid height differences, respectively.

The inherent appeal of this seemingly simple geometrical relationship between the three height types is based on the premise that given any two of the heights, the third can be derived through simple manipulation of (11.1), or similarly (11.2) for the relative case. In practice, the implementation of the above equation(s) is more complicated due to numerous factors that cause discrepancies when combining the heterogeneous heights (see Sect. 11.6 for more details). Some of these factors include, but are not limited to, the following:

- Random errors in the derived heights h, H, and N
- Datum inconsistencies inherent among the height types, each of which usually refers to a slightly different reference surface
- Systematic effects and distortions primarily caused by long-wavelength geoid errors, poorly modelled GPS errors (e.g., tropospheric refraction), and over-constrained levelling network adjustments
- Assumptions and theoretical approximations made in processing observed data, such as neglecting sea surface topography effects or river discharge corrections for measured tide gauge values

- Approximate or inexact normal/orthometric height corrections
- Instability of reference station monuments over time due to geodynamic effects and land subsidence/uplift

The major part of the aforementioned discrepancies is usually attributed to the systematic errors and datum inconsistencies. The task of dealing with these effects has been designated to the incorporation of a parametric model in the combined adjustment of the heights. Numerous assessments have been conducted using this approach with several different types of parametric models from a simple bias, a bias and a tilt, higher order polynomials with different base functions, finite element models, Fourier series and collocation-based approaches. It is evident that the appropriate type of parameterized surface model will vary depending on the height network data (distribution, density and quality) and therefore a universal model applicable in all cases is not practical. In Sect. 11.6 *valid procedures* for assessing model performance are presented. In this manner, an established general methodology may be implemented that offers the flexibility of being applied with any candidate parametric model and data set. The unknown parameters for a selected surface model are obtained via a common least-squares adjustment of ellipsoidal, orthometric and geoid height data over a network of co-located GPS-levelling benchmarks (points where h, H, and N are known).

A key issue in this type of common adjustment is the separation of errors among each height type, which in turn allows for the improvement of the stochastic model for the observational noise through the estimation of variance components. There are numerous reasons for conducting such variance component estimation (VCE) investigations. For example, consider the case of optimally refining/testing existing gravimetric geoid models using GPS-levelling height data. Such a comprehensive calibration of geoid error models is essential for numerous applications such as, mean sea level studies, connection of different continental height systems, and establishing vertical control independent of spirit-levelling, to name a few. The latter application is especially important in mountainous terrain and remote areas without existing vertical control. The suitability of the stochastic model used in the combined network adjustment of the ellipsoidal, orthometric and geoid height data must also be carefully evaluated. This is an important element for the reliable least-squares adjustment of the geodetic data that is often neglected in practical height-related problems. An additional important area that will benefit from the implementation of VCE methods is the assessment of the a-posteriori covariance matrix for the height coordinates derived from GPS measurements. Specifically, it will allow for a means to test the accuracy values for the ellipsoidal heights provided from post-processing software packages, which are often plagued with uncertainty (and usually overly-optimistic). Furthermore, it will allow for the evaluation of the accuracy information provided for orthometric heights obtained from national/regional adjustments of conventional levelling data.

In the remaining sections of this chapter a detailed analysis of the optimal combination of heterogeneous height data, with particular emphasis on datum inconsistencies, systematic effects and data accuracy. The technique is intended

for vertical control networks consisting of high quality ellipsoidal, orthometric and geoid height data. The problem is strictly treated as a random field with statistical techniques (see Fotopoulos 2003 for more details).

11.3 Why Combine Geoid, Orthometric and Ellipsoidal Height Data?

A number of important geodetic application areas that will benefit from the optimal combination of the heterogeneous height data include (but are not limited to): modernizing regional vertical datums, unifying national/regional vertical datums for a global vertical datum, transforming between different types of height data, and refining and testing existing gravimetric geoid models. As we move towards an increased use (and in some case, exclusive use) of space-based data acquisition technologies for coordinate/and height information the ability to correctly combine traditionally obtained measurements with newer measurements becomes an essential tool. In particular, the study of long-term geodynamic trends requires the use of heterogeneous data to provide the time series for interpolation and extrapolation over time.

11.3.1 Modernizing Regional Vertical Datums

A vertical datum is a reference surface to which the vertical coordinates of points are referred. At a national level, some of the practical uses and benefits of a consistent regional vertical datum include improved coastal/harbour navigation, surveying/engineering applications, accurate elevation models for natural hazards such as floods and coastal erosion, improved management of natural resources (e.g., water), accurate geospatial data, and monitoring of global environmental change. Traditionally, geodesists have used either a geoid, a quasi-geoid, or a reference ellipsoid as a vertical datum. All of these reference surfaces can be defined either globally or regionally, such that they approximate the entire Earth's surface or some specified region, respectively.

With no *official* global vertical datum definition, most countries or regions today use regional vertical datums as a local reference height system. This has resulted in over 100 regional vertical datums being used all over the world. The datums vary due to different types of definitions, different methods of realizations and the fact that they are based on local/regional data. A common approach for defining regional vertical datums is to average sea level observations over approximately 19 years (or more precisely, $\sim$18.6 years, which corresponds to the longest tidal component period) for one or more fundamental tide gauge. This average sea level value is known as mean sea level (MSL) and is used because it was assumed that

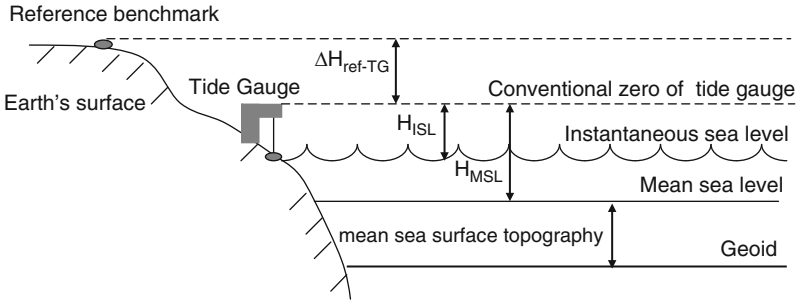


Fig. 11.1 Establishment of a reference benchmark height (After Vaníček and Krakiwsky 1986)

the geoid and MSL coincided (more or less). This assumption is obviously false, as it is known today that the MSL and the geoid differ by approximately ±2 m. Also, the geoid is by definition an equipotential surface, whereas MSL is not, due to numerous meteorological, hydrological, and oceanographic effects.

Figure 11.1 depicts a typical scenario for the establishment of a reference benchmark to define a regional vertical datum. The tide gauge records the instantaneous sea level height H_{ISL} and these values are averaged over a long term in order to obtain the mean value of the local sea level H_{MSL}. The height of the tide gauge is also measured with respect to a reference benchmark that is situated on land a short distance from the tide gauge station. Then the height of the reference benchmark above mean sea level H_{ref} is computed by

$$H_{ref} = H_{MSL} + \Delta H_{ref-TG} \tag{11.3}$$

Levelling begins from this benchmark and reference heights are accumulated by measuring height differences along levelling lines. The accuracy of the reference benchmark height derived in this manner is dependent on the precision of the height difference ΔH_{ref-TG} and the value for mean sea level H_{MSL}. If one assumes that the value for mean sea level is computed over a sufficiently long period of time which averages out all tidal period components and any higher frequency effects such as currents, then the accuracy depends on ΔH_{ref-TG}.

For highly accurate heights such as those needed for a cm-level vertical datum, the tide gauges cannot be assumed to be vertically stable because land motion at tide gauges is a source of systematic error, which causes distortion in the height network if it is not corrected for. Land motion at tide gauges and reference benchmarks may be caused abruptly be earthquakes or by erosion or more subtle changes such as post-glacial rebound and land subsidence. One solution to this problem is to include an independent space-based geodetic technique such as GNSS in order to estimate the land motion at these tide gauges. The challenge for a global solution based on this approach remains as there are still too few measurements available at tide gauges to provide an accurate assessment of the global situation (refer Chaps. 6 and 9).

As new methodologies and techniques evolve to the point where cm-level (and even sub-cm-level) accurate coordinates are needed, the distortions in traditionally-defined regional vertical networks are no longer acceptable. With this in mind, different approaches for realizing a 'modern' regional vertical datum have emerged (Vaníček 1991).

Define the geoid by mean sea level as measured by a network of reference tide gauges situated along the coastlines of the country and fix the datum to zero at these stations. As stated previously, this approach will result in distorted heights as MSL is not an equipotential surface and it varies from the geoid on the order of a few metres. Also, by fixing the datum to zero at these tide gauge stations, one is assuming that the gauge measurements are errorless or any error inherent in the measurements is acceptable. This is also a boldly incorrect assumption. For instance, consider the case in Canada where not only are some tide gauges poorly situated (sites of river discharge), but also the land to which the tide gauges are stationed is moving due to post-glacial rebound. It is known that regions such as Canada and the Scandinavian countries are rebounding or subsiding up to 1–2 mm/year. If these tide gauge motions are neglected, the error propagates into the levelled heights referred to the regional vertical datum and causes distortions and inconsistencies in the final orthometric heights.

Define the vertical datum by performing an adjustment where only one point is held fixed. A "shift" is applied to the resulting heights from the adjustment so that the mean height of all tide gauges equals zero. This approach relies heavily on the measurements from a *single* tide-gauge, while ignoring the observations for MSL made at all other stations.

Use the best model available to estimate sea surface topography at the tide gauge stations and then adjust the network by holding MSL-MSST to zero for all tide gauges. This approach suffers from practical limitations in terms of accuracy of measurements in coastal areas (see, Chap. 9 on the performance of satellite altimetry near the coastlines). Global ocean circulation models derived from satellite altimetry data and hydrostatic models may reach accuracies of 2–3 cm in the open oceans, but the models fall apart in shallow coastal areas giving uncertainties on the order of tens of centimetres. Therefore, with significant problems still looming in the coastal regions, distortions will be evident in heights referred to a vertical datum that is defined with low accuracy SST models.

Define the vertical datum by performing an adjustment with the reference tide gauges allowed to 'float' through the assignment of realistic a-priori weights (estimates of errors). This approach can incorporate all of the information for MSL and SST at the reference tide gauges. With improvements in models obtained from satellite altimetry and a better understanding of the process of tide gauge observations (e.g., reference benchmark stability, changes in position), estimates of the accuracy of the observations can be made.

Use estimates of orthometric heights from satellite-based ellipsoidal heights and precise gravimetric geoidal heights. One of the main advantages of this approach is that it relates the regional vertical datum to a global vertical reference surface (since the satellite-derived heights are referred to a global reference ellipsoid).

11.3.2 Global Vertical Datum

A global vertical datum can be defined as a height reference surface for the whole Earth. The concept of a global vertical datum has been a topic of great research and debate over the past century and has yet to be established as an international standard although numerous proposals from the geodetic community have been made. The establishment of an accurate, consistent and well-defined global vertical datum has many positive implications including the provision of a consistent and accurate method for connecting national and/or regional vertical datums, and the removal of inconsistencies in gravity anomalies and heights resulting from the use of different datums (by referring measurements to a common geopotential surface). Another area where a global vertical datum has been deemed necessary is for global change applications, such as, global change monitoring, mean sea level changes, instantaneous sea surface models, polar ice-cap volume monitoring, post-glacial rebound and land subsidence studies. All of these applications require a global view of the Earth with measurements not only on land, but over the oceans as well.

An accurate datum connection across the globe requires very accurate geoid determination over varying wavelengths (depending on the spatial distance between regional height systems) as well as consistency between regions. Other strategies offered for solving the global vertical datum problem include purely oceanographic approach, the use of satellite altimetry combined with geostrophic levelling, geodetic boundary-value problem, and satellite positioning (GNSS) combined with gravimetry.

11.3.3 GNSS-Levelling

The optimal combination of GPS-derived ellipsoidal heights with gravimetrically-derived geoid undulations for the determination of orthometric heights above mean sea level, or more precisely with respect to a vertical geodetic datum is referred to as GPS-levelling (or more generally GNSS-levelling). The process can be described as follows for the absolute and relative cases (height difference between two points i and j), respectively:

$$H = h - N \tag{11.4}$$

$$H_j - H_i = (h_j - h_i) - (N_j - N_i)\,,\ \Delta H = \Delta h - \Delta N \tag{11.5}$$

This procedure has been a topic of interest over the years and has been demonstrated to provide a viable alternative over conventional levelling methods for lower-order survey requirements. A major limitation of using GPS-levelling as a means for establishing heights or height differences with respect to a local vertical datum is that it is dependant on the achievable accuracy of the ellipsoidal and geoid height data.

In practice, the GPS-levelling technique has become quite common and used often erroneously or with a poor understanding of the transformations between reference surfaces and systematic errors involved. As accuracy requirements increase, the incorrect application of (11.4) has more severe implications. Therefore, it is important to develop proper procedures for combining the heterogeneous height data and a means to convey this information to users.

11.3.4 Refining and Testing Gravimetric Geoid Models

Another common manipulation of (11.1) is the combined use of co-located ellipsoidal and orthometric heights (or height differences) in order to compute geoidal height values at the GPS-levelling benchmarks. This trivial manipulation of (11.4) and (11.5) leads to GPS-derived geoid heights, which are invariably different from the values interpolated from a gravimetrically-derived geoid model. For instance, a gravimetrically computed geoid model, obtained from the *remove-compute-restore* process described in Part I, will (theoretically) refer to the geocentric reference system implicit in the used geopotential model. This reference system will in turn correspond to the adopted coordinate set for the satellite tracking stations used in the global geopotential solution. This coordinate set will not necessarily agree with the adopted reference system for the ellipsoidal heights obtained from the GPS measurements. Furthermore, the local levelling datum to which the orthometric heights refer will not likely correspond to the reference potential value of the geopotential model or the GPS reference system. In practice, the major applications of (11.1) include external independent evaluation of gravimetric geoid accuracy, incorporation of GPS-derived geoid heights into the gravimetric geoid solution as a soft constraint, and densification of networks that have already been positioned by conventional horizontal and vertical methods.

Comparisons between different geoid solutions provide insight into the accuracy of the geoid determination techniques. To date, comparisons of gravimetrically-derived geoid model values interpolated at GPS-on-benchmarks with geometrically computed geoid values provide the best external means of evaluating the geoid model accuracy. In order for this method to provide an indication of the 'accuracy' of the gravimetric geoid model, it is important that the GPS-levelling data used for testing is not incorporated in the original geoid solution (an obvious statement, but often neglected statement).

Long-wavelength errors present in gravimetrically-derived geoid models may be reduced by constraining the geoid solution to observed geoid values at GPS-levelling benchmarks. This is a common procedure implemented in many recent national geoid models through the use of least-squares collocation procedures, and shown to give positive results.

11.4 Least-Squares Adjustment Methodology for Combining Heights

Consider the following general linear functional model:

$$\mathbf{l} = \mathbf{A}\mathbf{x} + \tilde{\mathbf{v}}, \;\; E\{\tilde{\mathbf{v}}\} = 0 \tag{11.6}$$

where the $(m \times 1)$ vector of observations $\mathbf{l}$ is composed of the height "misclosure" at the GPS-levelling benchmark "i" as follows:

$$l_i \;=\; h_i - H_i - N_i \tag{11.7}$$

Using this mathematical model, at *each* GPS-levelling benchmark three independently derived height-types are available. For the purposes of this discussion, all formulations assume absolute height values, however equivalent formulations for baseline information (relative heights: Δh, ΔH, ΔN) can also be applied depending on the form of the available data. The $(m \times u)$ design matrix, $\mathbf{A}$, depends on the parametric model type (u is the number of unknown parameters). A classic four-parameter model introduced in Part I is given by

$$x_1 + x_2 \cos\varphi \cos\lambda + x_3 \cos\varphi \sin\lambda + x_4 \sin\varphi \tag{11.8}$$

where φ, λ are the latitude and longitude, respectively, of the GPS-levelling points. Theis represents a translation (x_2, x_3, x_4) and a change of the reference value of the potential x_1 (Sansò and Venuti 2002b). The full form of the design matrix is

$$\mathbf{A}_{m\times 4} = \begin{pmatrix} 1 & \cos\varphi_1 \cos\lambda_1 & \cos\varphi_1 \sin\lambda_1 & \sin\varphi_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cos\varphi_{m-1} \cos\lambda_{m-1} & \cos\varphi_{m-1} \sin\lambda_{m-1} & \sin\varphi_{m-1} \\ 1 & \cos\varphi_m \cos\lambda_m & \cos\varphi_m \sin\lambda_m & \sin\varphi_m \end{pmatrix} \tag{11.9}$$

This model is often applied naively and it is important to note that numerous studies have revealed more appropriate parametric forms rigorously determined for particular networks (Jiang and Duquenne 1996; Fotopoulos 2003). In (11.6), $\mathbf{x}$ is a $(u \times 1)$ vector containing the unknown parameters corresponding to the selected parametric model (e.g., $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^{\mathrm{T}}$ for the model given in (11.8)) and $\tilde{\mathbf{v}}$ is the zero-mean random error vector composed of a linear combination of $(m \times 1)$ vectors of random errors, denoted by $\mathbf{v}_{(\cdot)}$, for each of the height data types as follows:

$$\tilde{\mathbf{v}} = \mathbf{v}_{\mathrm{h}} - \mathbf{v}_{\mathrm{H}} - \mathbf{v}_{\mathrm{N}} \tag{11.10}$$

An equivalent formula for $\tilde{\mathbf{v}}$ is given by

$$\tilde{\mathbf{v}} = \mathbf{B}\mathbf{v} \tag{11.11}$$

where $\mathbf{B}$ is a block-structured matrix expressed as

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} & -\mathbf{I} \end{bmatrix} \tag{11.12}$$

such that each $\mathbf{I}$ is an $(m \times m)$ unit matrix and $\mathbf{v}$ is a vector of random errors with zero mean, described by

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_h^{\mathrm{T}} & \mathbf{v}_H^{\mathrm{T}} & \mathbf{v}_N^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{11.13}$$

The corresponding CV matrix is given by

$$E\left\{\mathbf{v}\mathbf{v}^{\mathrm{T}}\right\} = \mathbf{C}_{\mathbf{v}} \tag{11.14}$$

where $E\{\cdot\}$ denotes the mathematical expectation operator. The individual CV matrices according to each height type are

$$E\left\{\mathbf{v}_h\mathbf{v}_h^{\mathrm{T}}\right\} = \mathbf{C}_h,\ E\left\{\mathbf{v}_H\mathbf{v}_H^{\mathrm{T}}\right\} = \mathbf{C}_H,\ E\left\{\mathbf{v}_N\mathbf{v}_N^{\mathrm{T}}\right\} = \mathbf{C}_N \tag{11.15}$$

where $\mathbf{C}_{(\cdot)}$ is a (fully-populated) positive-definite symmetric matrix. It will be assumed that no correlation exists between the different height types, which implies that the cross-covariance matrices are formed as follows:

$$\mathbf{C}_{ij} = E\left\{\mathbf{v}_i\mathbf{v}_j^{\mathrm{T}}\right\} = \mathbf{0} \quad \text{where } i \neq j \quad \text{and} \quad i, j = h, H, N \tag{11.16}$$

The effect of such cross-correlations, which exist, for instance, in the case where the gravimetric geoid solution has incorporated height information from the levelling network, are ignored for the purposes of this chapter. This assumption simplifies matters significantly, as well as reduces the computational load. It may also be argued that it is a practical presumption as such reliable cross-covariance matrices are scarcely available in practice with real datasets.

The solution for the unknown parameters is obtained by applying the LS minimization principle of

$$\mathbf{v}^{\mathrm{T}}\mathbf{P}\mathbf{v} = \mathbf{v}_h^{\mathrm{T}}\mathbf{P}_h\mathbf{v}_h + \mathbf{v}_H^{T}\mathbf{P}_H\mathbf{v}_H + \mathbf{v}_N^{\mathrm{T}}\mathbf{P}_N\mathbf{v}_N = \text{minimum} \tag{11.17}$$

where the block diagonal weight matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_h & 0 & 0 \\ 0 & \mathbf{P}_H & 0 \\ 0 & 0 & \mathbf{P}_N \end{bmatrix} = \begin{bmatrix} \mathbf{C}_h^{-1} & 0 & 0 \\ 0 & \mathbf{C}_H^{-1} & 0 \\ 0 & 0 & \mathbf{C}_N^{-1} \end{bmatrix} \tag{11.18}$$

According to the above formulations, one can easily solve for the unknown parameters/coefficients of the parametric surface model by

$$\hat{\mathbf{x}} = [\mathbf{A}^{\mathrm{T}}(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{A}]^{-1}\mathbf{A}^{\mathrm{T}}(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{l} \tag{11.19}$$

The combined adjusted residuals from the adjustment are given by

$$\mathbf{B}\hat{\mathbf{v}} = \hat{\mathbf{v}}_h - \hat{\mathbf{v}}_H - \hat{\mathbf{v}}_N \tag{11.20}$$

It should be noted that the introduction of the **B** matrix (11.11) plays an essential role as it allows for the formulation of **separate** adjusted residuals according to height type, even though, the "observed" input values consist of a **combined** misclosure of all height data (see (11.7)). This being the case, we can explicitly solve for the separate adjusted residuals, according to height data type, by applying the well known formula:

$$\hat{\mathbf{v}} = \mathbf{P}^{-1}\mathbf{B}^{\mathrm{T}}(\mathbf{B}\mathbf{P}^{-1}\mathbf{B}^{\mathrm{T}})\,(\mathbf{w} - \mathbf{A}\hat{\mathbf{x}}) \tag{11.21}$$

where $\mathbf{w} = \mathbf{l}$ and is also explicitly shown in Kotsakis and Sideris (1999) as

$$\hat{\mathbf{v}}_h = \mathbf{C}_h(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{M}\mathbf{l} \tag{11.22a}$$

$$\hat{\mathbf{v}}_H = -\mathbf{C}_H(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{M}\mathbf{l} \tag{11.22b}$$

$$\hat{\mathbf{v}}_N = -\mathbf{C}_N(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{M}\mathbf{l} \tag{11.22c}$$

where the **M** matrix is expressed by

$$\mathbf{M} = \mathbf{I} - \mathbf{A}(\mathbf{A}^{\mathrm{T}}(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}(\mathbf{C}_h + \mathbf{C}_H + \mathbf{C}_N)^{-1} \tag{11.23}$$

This formulation provides us with the instrumental opportunity to evaluate the contribution of each of the height types through the evaluation of $\mathbf{C}_{\mathbf{v}}$, which is represented by the following expression for the case of heterogeneous disjunctive data

$$\mathbf{C}_{\mathbf{v}} = \begin{bmatrix} \mathbf{C}_h & 0 & 0 \\ 0 & \mathbf{C}_H & 0 \\ 0 & 0 & \mathbf{C}_N \end{bmatrix} = \begin{bmatrix} \sigma_h^2\mathbf{Q}_h & 0 & 0 \\ 0 & \sigma_H^2\mathbf{Q}_H & 0 \\ 0 & 0 & \sigma_N^2\mathbf{Q}_N \end{bmatrix} \tag{11.24}$$

where $\mathbf{Q}_h$, $\mathbf{Q}_H$ and $\mathbf{Q}_N$ are known (not necessarily fully-populated) positive-definite cofactor matrices for ellipsoidal, orthometric and geoid height data respectively and σ_h^2, σ_H^2, σ_N^2 are the three corresponding unknown variance components.

The unique perspective obtained by implementing this combined adjustment approach is embedded in two main areas, namely (i) the evaluation of the contribution of the **Ax** component, which refers to the total correction term for the systematic errors and datum inconsistencies in the multi-data test network, and (ii) the separation of residuals according to the height data types (11.22), which allows for the refinement/calibration of data covariance matrices. Estimating the individual σ_h^2, σ_H^2, σ_N^2 values, must be suitably addressed in order to achieve practically useful results.

11.5 Application of MINQUE to the Combined Height Adjustment Problem

There are a number of methods available to perform VCE within the context of LS adjustment. A first solution to the problem was provided by Helmert (1924), who proposed a method for unbiased variance estimates. An independent solution was derived by Rao (1970), who appeared unaware of Helmert's method, and was called the minimum norm quadratic unbiased estimation (MINQUE) method. Under the assumption of normally distributed observations, both Helmert's and Rao's MINQUE approach are equivalent. In Teunissen and Amiri-Simkoowei (2008), VCE by the method of least-squares is rigorously described and demonstrated as being a flexible and relatively simple method to apply in practice. Ultimately, the selection of the appropriate technique should rely on the desired estimator properties, such as translation invariance, unbiasedness, minimum variance, non-negativeness, computational efficiency and ease of implementation, to name a few. In some cases all of these properties cannot be retained for a particular estimator (e.g., the property of unbiasedness may be sacrificed for guaranteed estimation of non-negative variances, see Hartung 1981). In general, the decision for which estimator properties to retain/enforce must be made on a case-by-case basis depending on the data and specific application. The over-riding property that is usually sought after is computational efficiency, which arises due to the massive quantities of data that are used for the estimation of many variance-covariance components. In fact, the main criticism of traditional VCE methods is that they involve repeated inversions of large matrices, intensive computational efforts and large storage requirements for lots of unknowns. For these reasons, one may opt for entirely different estimation procedures, such as the Monte Carlo technique or simplifications to the rigorous algorithm in order to reduce the computational burden involved with inverting large dimensional matrices.

In this chapter, the MINQUE procedure is followed (Rao 1971; Rao and Kleffe 1988). The selection was based on evaluating the utility of criteria such as computational load, balanced versus unbalanced data, ease of implementation, algorithmic flexibility, unbiasedness and non-negative variance factors to the common adjustment of *practical* heterogeneous height data.

The emphasis in this section is to provide the modifications required to the general MINQUE algorithm in order to adopt it to the combined adjustment of ellipsoidal, geoid and orthometric heights.

Given the functional model provided in (11.4) and the selected stochastic model provided in (11.24), the MINQUE problem is reduced to the solution of

$$\mathbf{S}\hat{\boldsymbol{\theta}} = \mathbf{q} \tag{11.25}$$

where $\hat{\boldsymbol{\theta}}$ is a vector containing the three unknown variance components $\sigma_h^2, \sigma_H^2, \sigma_N^2$. The composition of the symmetric matrix $\mathbf{S}$ is

$$\mathbf{S} = \begin{bmatrix} s_{hh} & s_{hH} & s_{hN} \\ s_{Hh} & s_{HH} & s_{HN} \\ s_{Nh} & s_{NH} & s_{NN} \end{bmatrix} \tag{11.26}$$

where each element $\{s_{ij}\}$ in the matrix is computed from

$$s_{ij} = tr\left(\mathbf{RQ}_i\mathbf{RQ}_j\right), \quad i, j = h, H, N \tag{11.27}$$

where $tr(\cdot)$ is the trace operator, $\mathbf{Q}_{(\cdot)}$ is the known positive-definite cofactor matrix for each height type as introduced in (11.24). It should be noted that the matrix $\mathbf{S}$ may not be of full rank and therefore its pseudo-inverse can be used for solving (11.25). $\mathbf{R}$ is a symmetric matrix defined by

$$\mathbf{R} = \mathbf{C}_\mathbf{l}^{-1}\left[\mathbf{I} - \mathbf{A}(\mathbf{A}^\mathrm{T}\mathbf{C}_\mathbf{l}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{C}_\mathbf{l}^{-1}\right] \tag{11.28}$$

where $\mathbf{A}$ is an appropriate design matrix of full column-rank (as in (11.9)) and $\mathbf{C_l}$ is the CV matrix of the observations which is given by the following linearly additive model

$$\mathbf{C_l} = \mathbf{BC_v}\mathbf{B}^\mathrm{T} = \sigma_h^2\mathbf{Q_h} + \sigma_H^2\mathbf{Q_H} + \sigma_N^2\mathbf{Q_N} \tag{11.29}$$

The vector $\mathbf{q}$ contains the quadratic forms

$$\mathbf{q} = \{q_i\}, \quad q_i = \hat{\mathbf{v}}_i^\mathrm{T}\mathbf{Q}_i^{-1}\hat{\mathbf{v}}_i, \quad i = h, H, N \tag{11.30}$$

where $\hat{\mathbf{v}}_i$ is a vector containing the separate residuals for each input height type also denoted by $\hat{\mathbf{v}}_i = \mathbf{Q}_i\mathbf{Rl}$ and easily derived from the combined adjustment scheme described in the previous section and shown analytically by (11.22a–c).

Substituting the appropriate formulations above into the general system given by (11.25), we obtain the explicit expression

$$\begin{bmatrix} tr(\mathbf{RQ}_h\mathbf{RQ}_h) & tr(\mathbf{RQ}_h\mathbf{RQ}_H) & tr(\mathbf{RQ}_h\mathbf{RQ}_N) \\ tr(\mathbf{RQ}_H\mathbf{RQ}_h) & tr(\mathbf{RQ}_H\mathbf{RQ}_H) & tr(\mathbf{RQ}_H\mathbf{RQ}_N) \\ tr(\mathbf{RQ}_N\mathbf{RQ}_h) & tr(\mathbf{RQ}_N\mathbf{RQ}_H) & tr(\mathbf{RQ}_N\mathbf{RQ}_N) \end{bmatrix} \begin{bmatrix} \hat{\sigma}_h^2 \\ \hat{\sigma}_H^2 \\ \hat{\sigma}_N^2 \end{bmatrix} = \begin{bmatrix} \mathbf{l}^\mathrm{T}\mathbf{R}^\mathrm{T}\mathbf{Q}_h\mathbf{Rl} \\ \mathbf{l}^\mathrm{T}\mathbf{R}^\mathrm{T}\mathbf{Q}_H\mathbf{Rl} \\ \mathbf{l}^\mathrm{T}\mathbf{R}^\mathrm{T}\mathbf{Q}_N\mathbf{Rl} \end{bmatrix} \tag{11.31}$$

It is evident from the expression for the $\mathbf{R}$ matrix (11.28) that initial estimates for the unknown variance components must be provided as they are embedded in $\mathbf{C_l}$ that is used to compute $\mathbf{R}$. This introduces one of the main drawbacks or criticisms of the MINQUE approach, which is the fact that it is only a locally best estimator. In other words, it is implied that 'n' users with 'n' different a-priori values for the variance factors have the possibility of obtaining 'n' different estimates, all satisfying the criteria and properties imposed by the MINQUE procedure. This is considered a major obstacle because if good initial estimates were easily obtainable then there would be limited use in performing variance component estimation to begin with! Remedies for overcoming this shortcoming include the use of an iterative

Table 11.1 Effect of correlations on estimated variance factors (Switzerland network)

Type of covariance matrices	$\hat{\sigma}_h^2$	$\hat{\sigma}_H^2$	$\hat{\sigma}_N^2$
Full $\mathbf{Q}_h$, $\mathbf{Q}_H$, $\mathbf{Q}_N$	2.82	5.06	1.02
Diagonal $\mathbf{Q}_h$, $\mathbf{Q}_H$, $\mathbf{Q}_N$	0.71	3.63	1.07

approach in conjunction with ensuring that high redundancy is retained. The iterative implementation of (11.31) is referred to as iterative MINQUE. An elegant closed-form solution of the problem via least-squares as described by Teunissen and Amiri-Simkoowei (2008) would give identical results.

Remark 1. The problem described thus far delimits a rare characteristic for geodetic data, that of balanced data. Normally, when heterogeneous types of geodetic data are used in a combined adjustment, the number of observations per each group of data is not the same. In this case, the problem is pre-designed such that all three height groups/types are available for each network benchmark and subsequent separation of adjusted residuals results in three vectors with the same number of elements. Thus, we can estimate the variance components from balanced data – a less demanding task, in general, than dealing with unbalanced data. An additional advantage offered by the design of this particular problem is a relatively low computational load. Only three variance components are sought. The largest matrix inversion will be on the order of the number of observations, m, which in the absolute height data case described herein is equivalent to the number of levelled benchmarks with GPS data. Thus, the problem here lies in the absence of independently derived and reliable variance estimates for each height type.

Example 1 (Effect of correlation on the estimated variance components). In practice, fully-populated variance-covariance matrices for all types of height data at coincident GPS benchmarks are rarely available. More often than not, cross-correlations are ignored producing diagonal-only CV matrices for the height data. To test the effect of correlations between observations of the same type on the estimated variance components, numerical experiments were conducted with fully-populated and diagonal versions of the same covariance matrices (see Fotopoulos 2005 for complete details). The numerical results are summarized in Table 11.1. The Swiss national test network of GPS-levelling benchmarks distributed throughout an approximately 340×210 km region is used for this numerical example. The original fully populated CV matrix for the ellipsoidal heights was extracted from the results of an undisclosed commercial post-processing software package of GPS data. Typical for GPS, the output CV matrix is overly optimistic, a direct result of neglecting temporal, spatial and physical correlations between GPS phases. The initial fully-populated cofactor matrix for the orthometric heights, $\mathbf{Q}_H$, comes directly from the national adjustment of all first- and second-order levelling measurements. The orthometric heights were obtained from the division of the adjusted geopotential numbers by the mean gravity (computed from surface gravity measurements and a simple 3D density model of the Earth's crust;

see Marti et al. 2000). As expected, the correlation between nearby neighbouring stations is very high. The initial fully-populated CV matrix corresponding to the geoid heights at the GPS-levelling benchmarks was obtained by straightforward application of error propagation to the least-squares collocation equations that were used for the Swiss geoid determination

$$\mathbf{Q}_N = \mathbf{C}_{NN} - \mathbf{C}_{N\Delta g}\,\mathbf{C}_{ZZ}^{-1}\,\mathbf{C}_{N\Delta g}^{\mathrm{T}}$$

where $\mathbf{C}_{NN}$ is the covariance matrix of the true unknown geoid heights, $\mathbf{C}_{N\Delta g}$ denotes the cross-covariance matrix between the computed geoid heights, N, and the measured gravity anomalies, Δg, and $\mathbf{C}_{ZZ} = \mathbf{C}_{\Delta g\Delta g} + \mathbf{C}_{nn}$ where n is noise. The computed CV matrix, in this case, excluded the uncertainty contribution of the global geopotential model as well as other effects such as terrain reductions and assumptions about the density models due to the choice of covariance function. However, it should be noted that in general least-squares collocation can also be used to model the error from the global geopotential model and the topography/density. In practice, regional geoid models are often refined through the incorporation of GPS-levelling data into the gravimetric solution (Tscherning et al. 2001). In such cases, the assumption of disjunctive observations is not strictly valid, which will adversely affect calculations of error models. To rectify this situation, the GPS-levelling observations can be excluded from the computation of the gravimetric geoid, which will ensure independence among the CV matrices, satisfying. Equation 11.16.

By neglecting the off-diagonal elements, we obtain overly optimistic CV matrices compared to the fully-populated case. Results will vary depending on the degree of correlation, however it is clear that unrealistically 'good' results are obtained when correlations are ignored, as expected. An exception is shown in the computed $\hat{\sigma}_N^2$ factor where the estimated values corresponding to the fully-populated and diagonal-only CV matrices are essentially the same, with a slight increase for the diagonal-only CV matrix. The results show that the off-diagonal elements should not be considered insignificant and efforts should be made, when possible, to include all of the available CV information, especially for high precision applications.

11.6 Role of the Parametric Model

The main factors that cause discrepancies when combining the heterogeneous heights include the following (Schwarz et al. 1987):

Random errors in the derived heights h, H, and N

The covariance matrices for each of the height types are usually obtained from separate network adjustments of the individual height types.

Datum inconsistencies inherent among the height types

Each of the triplet of height data refers to a different reference surface. For instance, GPS-derived heights refer to a reference ellipsoid used to determine the satellite orbits. Orthometric heights, computed from levelling and gravity data, refer to a local vertical datum, which is usually defined by fixing one or more tide-gauge stations. Finally, the geoidal undulations interpolated from a gravimetrically-derived geoid model refer to the reference surface used in the global geopotential model, which may not be the same as the one for the gravity anomalies Δg.

Systematic effects and distortions in the height data

These systematic effects have been described in Chap. 6 and are mainly caused by the long wavelength geoid errors, which are usually attributed to the global geopotential model. Biases are also introduced into the gravimetric geoid model due to differences between data sources whose adopted reference systems are slightly different. In addition, systematic effects are also contained in the ellipsoidal heights, which are a result of poorly modelled GPS errors, such as atmospheric refraction and in particular tropospheric errors. Although spirit-levelled height differences are usually quite precise, the derived orthometric heights for a region or nation are usually the result of an over-constrained levelling network adjustment, which introduces distortions.

Assumptions and theoretical approximations made in processing observed data

Common approximations include neglecting sea surface topography (SST) effects or river discharge corrections for measured tide gauge values, which results in a significant deviation of readings from mean sea level. Other factors include the use of approximations or inexact normal/orthometric height corrections and using normal gravity values instead of actual surface gravity values in computing orthometric heights. The computation of regional or continental geoid models also suffers from insufficient approximations in the gravity field modelling method used.

Instability of reference station monuments over time

Temporal deviations of control station coordinates can be attributed to geodynamic effects such as post-glacial rebound, crustal motion and land subsidence. Most GPS processing software eliminate all tidal effects when computing the final coordinate differences. To be consistent, the non-tidal geoid should be used (Chap. 6). More details on the error caused by mixing ellipsoidal heights referring to a non-tidal crust and orthometric heights whose reference surface is the mean or zero geoid is given in Poutanen et al. (1996).

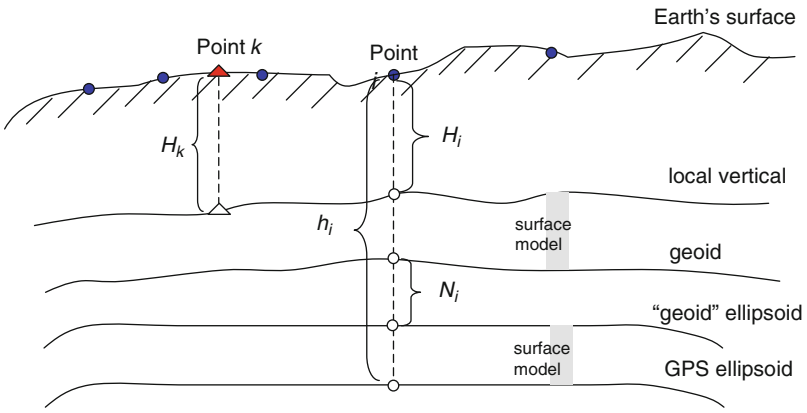


Fig. 11.2 Illustrative view of GPS-levelling and the role of a parametric model

As mentioned previously, thus far, the burden of dealing with most of these factors has been designated to the use of a parametric surface model. Given the theoretical relationship among the three types of height data and the incorporation of an appropriate surface model, the orthometric height for a *new* point which does not belong to the original multi-data network is obtained as follows:

$$H = h - N - a^T \hat{\mathbf{x}} \tag{11.32}$$

The question that must be addressed is *to which vertical reference system does the computed value H refer?* To answer this question, we refer to Fig. 11.2, which provides an illustrative view of the various reference surfaces embedded in the different height data sets.

In this figure, the points on the Earth's surface belong to the multi-data control network and the point denoted by a triangle is the 'new' point for which the orthometric height is to be computed via GPS-levelling. For the sake of this discussion, if one ignores the systematic effects, and concentrates on the datum inconsistencies, one can see from the figure that the role of the surface model is twofold. In general, the datum discrepancies occur between (i) the local vertical datum and the geoid model and (ii) the two ellipsoids to which the GPS measurements and geoid undulations refer to. These discrepancies are typically not constant biases as depicted in the figure, but they may take on a more complicated form.

In order to obtain the orthometric height through GPS-levelling that refers to the local vertical datum for the new point, H_k, a connection between the different height surfaces must be made. This connection is embedded in the $\mathbf{Ax}$ term (11.6) and can take on many forms depending on the selected model. It should be cautioned, however, that the surface model will provide a consistent connection between the heights derived from GPS-levelling and the official local vertical datum, only if the orthometric heights used in the multi-data adjustment also refer to the official local vertical datum.

11.6.1 Modelling Options

It is evident that the parametric model is important in the practical application of the combined height adjustment process. A significant amount of attention in research has been given to this issue, but not without some controversy as to its appropriateness given that the 'model' has no physical meaning. Moving forward with this implies that the user understands and accepts that the model is simply a 'mathematical' means for compensating for the discrepancies between the various heights over a region and cannot be interpreted any further. This being said, it is a practically useful endeavour if the purpose is to 'combine'/merge the heterogeneous height data to accommodate for the easily obtainable ellipsoidal heights via GNSS. Arguably, the selection of the parametric form of the surface model is arbitrary unless some physical reasoning can be applied to the discrepancies between the GPS-derived geoid heights N^{GPS}, and the geoid heights from the gravimetric geoid model N^{grav}, which fulfills

$$\ell_i = h_i - H_i - N_i = a_i^T x + v_i \tag{11.33}$$

$$a_i^T x = N^{GPS} - N^{grav} \tag{11.34}$$

In several cases, a simple tilted plane-fit model satisfies accuracy requirements. However, as the achievable accuracy of GPS and geoid heights improves, the use of such a simple model may not be sufficient. The problem is further complicated because selecting the proper model type depends on the data distribution, density and quality, which varies for each case.

In general, the most common approach to the bilinear term in (11.33) is to employ a parameterized trend with a finite set of unknown parameters represented in its linear form $p = b_1 f_1 + b_2 f_2 + \ldots\ldots + b_q f_q$, where $b_1, b_2, \ldots, b_q$ are the unknown coefficients to be solved for in the combined least-squares approach and $f_1, f_2, \ldots, f_q$ are the base functions, whose type may vary from a simple multiple regression as in:

$$\mathbf{a}_i^T \mathbf{x} = \sum_{m=0}^{M} \sum_{n=0}^{N} (\varphi_i - \overline{\varphi})^n (\lambda_i - \overline{\lambda})^m x_q \tag{11.35}$$

where $\overline{\varphi}, \overline{\lambda}$ are the mean latitude and longitude of the GPS-levelling points, respectively, and x_q contains the q unknown coefficients. Other functions that are trigonometric, harmonic, Fourier series, and wavelets may be used. In some cases, two or more different types of base functions may be merged to model long-wavelengths.

Another option is to adopt a model for the trend surfaces and model the remaining residuals using least-squares collocation where $\tilde{\mathbf{v}} = \mathbf{C}_{sr}(\mathbf{C}_{rr} + \sigma^{-1}\mathbf{I})^{-1}\mathbf{r}$, $\mathbf{r}$ is a vector of known residuals with variance σ, to be predicted at another location, denoted by s. The above equation is usually implemented, although not restricted to, using a second-order Markov covariance model. See Chap. 7 for details.

Thus far, the discussion on the type of model has been based on the use of a single model to represent an entire region. This approach is sometimes limiting as it assumes that a homogeneous set of discrepancies exist over an entire region, regardless of its extent and data distribution. Consider for instance, the task of selecting a single model to adequately model all of the discrepancies across large regions such as Canada, where comparatively sparsely distributed sets of GPS-levelling control points are available. An additional limitation of this approach is that it is typically difficult to model by a single covariance function both long and short wavelength discrepancies.

One way to deal with this is to divide the region into a number of smaller sub-regions and fit the appropriate model to that region noting that the model type/extent may vary for each sub-region. The issues of how to divide the region and how to connect across adjacent sub-regions prevail in these scenarios. One approach to this problem is a global transformation model is applied to deal with the general transformation of reference systems. Several polynomial models are applied to the divided sub-regions in order to deal with local deformations. The combined adjustment employs a set of constraint equations for common points in the neighbouring sub-regions. In the following section a validated procedure is presented for model assessment.

11.6.2 Semi-automated Assessment Procedure

In general, the process applied for selecting the best parametric model in a particular region suffers from a high degree of arbitrariness in both choosing the model type and in assessing its performance. Figure 11.3 provides a suggested semi-automated procedure for parametric model testing, which can be applied to the results of the combined least-squares adjustment of the ellipsoidal/orthometric/geoid heights. The term *semi-automated* is used to describe the procedure as some user intervention is required. It is assumed throughout the process that reliable information for the statistical behaviour of the ellipsoidal, geoid and orthometric height data is available and any gross errors/blunders have been detected and removed from the observational data in order for the results to be meaningful.

The most common method used in practice to assess the performance of the selected parametric model is to compute the statistics for the adjusted residuals after the least-squares fit. The adjusted residuals for each station in the network, $\hat{v}_i$, are computed as follows:

$$\hat{v}_i = h_i - H_i - N_i - \mathbf{a}_i^T \hat{\mathbf{x}} \tag{11.36}$$

The model that results in the smallest set of residuals is deemed to be the most appropriate. If one uses the root of the sums of squares of residuals to compute an RMS, then the RMS decreases as the number of model parameters increases. Thus, this method is valid for testing the precision of a model, but it should not be interpreted as the *accuracy* or the prediction capability of the model. The cross-

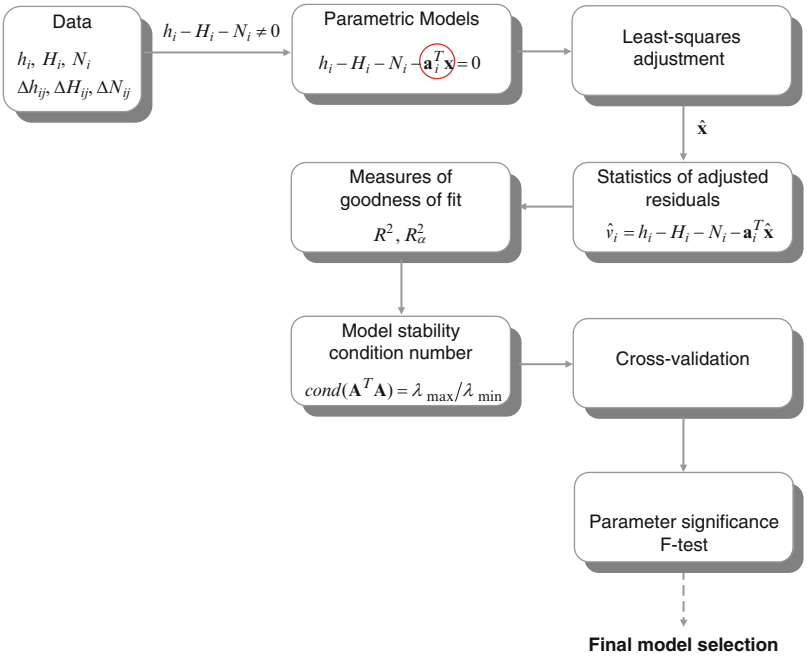


Fig. 11.3 Flowchart of assessment methodology for parametric models

validation approach provides a more realistic indication of the *accuracy* of a selected parametric model and its performance as a prediction surface for a new point. It is the preferred empirical testing scheme, as it does not rely exclusively on the accuracy of a single point or a small subset of points. It also maintains high data redundancy to compute the parameters in the combined least-squares adjustment.

A statistical measure of the goodness of the parametric model fit for a discrete set of points is given by the *coefficient of determination*, denoted by R^2 or the adjusted coefficient of determination, denoted by R_α^2. It can be described as the ratio of the sum of the squares due to the fit, to the sum of the squares about the mean of the observations, as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\ell_i - \hat{v}_i)^2}{\sum_{i=1}^{m} (\ell_i - \overline{\ell}_i)^2}, \quad R_\alpha^2 = 1 - \frac{(m-1)}{(m-u)}(1 - R^2) \tag{11.37}$$

where m observations are given by (11.5) and $\overline{\ell}_i$ is the mean value of the observations. In the extreme case where the parametric model fit is perfect, $\sum_{i=1}^{m} (\ell_i - \hat{v}_i)^2 = 0$ and $R^2 = 1$. The other extreme occurs if one considers the variation from the residuals to be nearly as large as the variation about the mean of

the observations resulting in $R^2 \rightarrow 0$. R^2 is a statistic and as with all statistics its values are somewhat governed by chance and peculiarities in the data. A relevant example to consider is the case where the data redundancy or degrees of freedom is small. In such cases, it is possible to obtain an erroneously large R^2 value, regardless of the quality of the fit. In fact, as the number of explanatory variables in the model increases, so does R^2. To deal with this limitation the alternative formulation for the *adjusted coefficient of determination* where u is the number of parameters in the model may be applied with the caveat that in some cases, it may provide negative values that do not have any meaning. Given the limitations of both measures of fit, it is important to not rely exclusively on these values. Instead, the values should be computed and accompanied by a reasonable interpretation and additional tests, such as the empirical procedures described in the previous two sections. A common result encountered in practice is the result of a low R^2 due to the fact that there was not enough variation in the observations to justify a 'good' or 'bad' fit. Therefore, these statistical measures can be a powerful tool in pointing out inappropriate models rather than establishing the validity of the model, which can be further tested by empirical cross-validation.

The principle of parsimony commonly referred to in statistical literature, where one should not use any more entities, beyond what is necessary, to explain anything is a useful guide in this case and therefore the significance of each parameter in the selected model should be tested. Unnecessary terms may bias other parameters in the model, which will hinder the capability to assess the model performance. Furthermore, over-parameterization may give unrealistic extrema in data voids where control points are missing. This is an important factor for the combined height problem in particular, as one of the most favourable locations to utilize GNSS-levelling are in areas where it is difficult to establish vertical control and therefore data gaps are prevalent. In theory, the decision on the degree of the polynomial/MRE surface should be reached by hypothesis testing (Dermanis and Rossikopoulos 1991). However, the results of such statistical tests are often hindered by the fact that independent coefficients generated by a polynomial series are usually correlated. Therefore, it is worth considering models with orthogonal base functions, which ensures no correlation between coefficients. If the application of these models is not suitable or too complex for practical use, then one can also apply orthogonalization/orthonormalization procedures (e.g., Gram-Schmidt) to decorrelate existing base functions. Finally, a very useful guideline to follow, if possible, is to select a set of nested models. The imposition of such a criterion for a set of models to be tested greatly facilitates the assessment process.

In general, there are three schemes that can be implemented, namely (i) backward elimination: one begins by fitting the data to the most extended or highest order form of the model and tests if a parameter or set of parameters in the model are significant, (ii) forward selection: begin with the lowest order model and test for parameter addition and (iii) stepwise procedure: combination of both aforementioned methods. Statistical testing (F-test) is applied to determine parameter significance. If the procedure is applied properly, statistical tests can be performed without the consequences of multi-colinearity.

Example 2 (Backward elimination). In the backward elimination procedure, one begins by fitting to the data the most extended form of the model. The next step is to test if a parameter or set of parameters in the model are significant. The vector of parameters can be separated and denoted by

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_{(I)} \end{bmatrix}$$

where $\mathbf{x}_I$ is the set of parameters to be tested and $\mathbf{x}_{(I)}$ are the remaining parameters in the model. The test is specified by the null hypothesis ($H_\circ$) that states which parameters are insignificant $H_\circ : \mathbf{x}_I = \mathbf{0}$, versus the alternative hypothesis (H_a) that declares these parameters to be significant, $H_a : \mathbf{x}_I \neq \mathbf{0}$.

The statistic used to test this null hypothesis is the *F-statistic* computed as a function of the observations (Dermanis and Rossikopoulos 1991)

$$\tilde{F} = \frac{\hat{\mathbf{x}}_I^T \, \mathbf{Q}_{\hat{\mathbf{x}}_I}^{-1} \, \hat{\mathbf{x}}_I}{k \hat{\sigma}^2} \tag{11.38}$$

where, $\mathbf{Q}_{\hat{\mathbf{x}}_I}^{-1}$ is the corresponding sub-matrix of the inverse of the normal equations, $\mathbf{Q}_{\hat{\mathbf{x}}} = \mathbf{N}^{-1}$, k is the number of parameters tested, and $\hat{\sigma}^2$ is the a-posteriori variance factor.

The null hypothesis is accepted when $\tilde{F} \leq F_{k,f}^{\alpha}$. $F_{k,f}^{\alpha}$ is computed from standard statistical tables for a confidence level α and degrees of freedom f (see Papoulis 1990 and Koch 1999 for details). If the above condition is fulfilled then the corresponding parameters are deleted from the model. If the contrary is true, $\tilde{F} > F_{k,f}^{\alpha}$, the 'tested' parameters remain in the model. The procedure is repeated until all of the remaining parameters in the model pass the *F-test* or the user is satisfied with the final model.

An alternative equation for computing the *F-statistic* is given by Wesolowsky (1976):

$$\tilde{F} = \frac{\left[\sum (\ell - \hat{v})^2_{partial} - \sum (\ell - \hat{v})^2_{full} \right] / k}{\left[\sum (\ell - \hat{v})^2_{full} \right] / m - u} \tag{11.39}$$

where the subscripts *full* and *partial* denote the values computed using all of the parameters in the model and the *partial* denotes the values computed if the 'tested' parameters are eliminated. This statistic, termed the *partial F-test*, is commonly implemented for testing regression parameters. However, in this case, (11.38) was preferred as it allows for the significance of parameters to be scrutinized and eliminated without the need to repeat the combined least-squares height adjustment.

Applying the described assessment procedure including backward elimination to test parameter significance to the national GPS-levelling benchmark network in Switzerland (acknowledge Swiss Mapping Authority for data here), results in an average RMS after fit of approximately 2.3 m. Figure 11.4 shows the parametric model over the region for an initial six-parameter model and the final nested three parameter (plane-fit) selected.

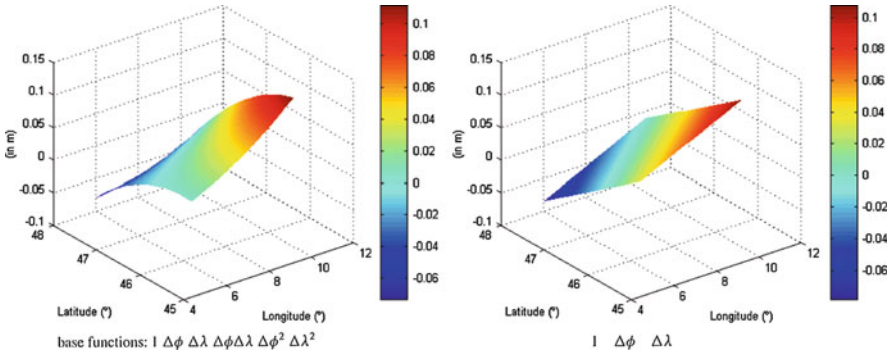


Fig. 11.4 Parametric model fit to GPS-levelling benchmarks for Switzerland network with six parameters (*left*) and three parameters (*right*)

Table 11.2 Description of data

Network	Coverage area	Min (cm)	Max (cm)	μ (cm)	σ (cm)	RMS (cm)
Switzerland (111 pts)	330 × 210 km	−4.9	19.0	1.1	3.8	4.0
Alberta/BC (63 pts)	495 × 335 km	−17.1	25.2	4.5	8.1	9.3

11.6.3 Numerical Example

Two numerical data sets from Switzerland and Canada (labeled Alberta/BC) have been used to demonstrate the selection and assessment process for parametric models. A summary of the basic network characteristics and statistics of the original height residuals computed from the GPS-levelling benchmarks and the corresponding regional gravimetric geoid models is provided in Table 11.2 (for more details on the networks see Fotopoulos 2003; Fotopoulos and Sideris 2005).

Table 11.3 summarizes the analytical models used for testing (note that some of the models in this table have been included to emphasize the numerical pitfalls that can be encountered during the parametric modeling process and it is obvious that these models cannot be interpreted for any "physical" meaning).

In the Swiss case, the GPS-on-benchmark data was well distributed with a small average height misclosure of 1.1 cm, compared to more than 9 cm for the Alberta/BC network. Figure 11.5 shows the computed coefficient of determination and the adjusted coefficient of determination. According to these measures, for the Swiss network, there is only a slight difference between the performance of the different models with a marginal variability in R^2 between 0.56 and 0.66 and the more indicative R^2_α ranging from 0.53 to 0.57. The one outstanding model is the fourth-order bivariate polynomial (model G), which corresponds to the highest

Table 11.3 Summary of parametric models tested

Model	Base functions
A. 1st order polynomial	$1\ \Delta\varphi\ \Delta\lambda$
B. 4-parameter	$1\ \cos\varphi\cos\lambda\ \cos\varphi\sin\lambda\ \sin\varphi$
C. 5-parameter	$1\ \cos\varphi\cos\lambda\ \cos\varphi\sin\lambda\ \sin\varphi\ \sin^2\varphi$
D. 2nd order polynomial	$1\ \Delta\varphi\ \Delta\lambda\ \Delta\varphi\Delta\lambda\ \Delta\varphi^2\ \Delta\lambda^2$
E. differential similarity	$\cos\varphi\cos\lambda\ \cos\varphi\sin\lambda\ \sin\varphi\ \frac{\sin\varphi\cos\varphi\sin\lambda}{W}\ \frac{\sin\varphi\cos\varphi\cos\lambda}{W}$ $\frac{1-f^2\sin^2\varphi}{W}\ \frac{\sin^2\varphi}{W}$; $W=\sqrt{1-e^2\sin^2\varphi}$
F. 3rd order polynomial	$1\delta\varphi\ \delta\lambda\ \Delta\varphi\Delta\lambda\ \Delta\varphi^2\ \Delta\lambda^2\ \Delta\varphi^2\Delta\lambda\ \Delta\varphi\Delta\lambda^2\ \Delta\varphi^3\ \Delta\lambda^3$
G. 4th order polynomial	$1\ \Delta\varphi\ \Delta\lambda\ \Delta\varphi\Delta\lambda\ \Delta\varphi^2\ \Delta\lambda^2\ \Delta\varphi^2\Delta\lambda\ \Delta\varphi\Delta\lambda^2\ \Delta\varphi^3\ \Delta\lambda^3\ \Delta\varphi^2\ \Delta$ $\Delta\varphi^3\Delta\lambda\ \Delta\varphi\Delta\lambda^3\ \Delta\varphi^4\ \Delta\lambda^4$

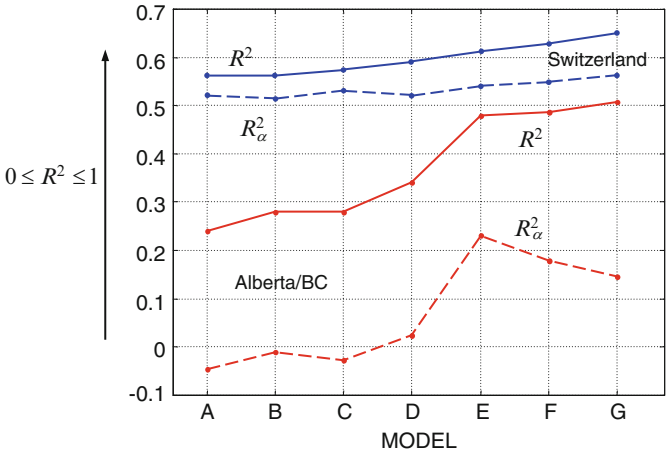


Fig. 11.5 Statistical measures of goodness of fit for various parametric models

measures of goodness of fit. Therefore, based on the combined results of these first tests, the fourth-order polynomial would be identified as being suitable for the Swiss region. However, further tests will show that a very different conclusion can be drawn. Perhaps the most revealing test results are given in Fig. 11.6, which summarizes the results of the empirical cross-validation at independent control points. The models are arranged according to the number of parameters increasing from top to bottom. From these values it is evident that the fourth-order polynomial fit is not the best choice, for either the Swiss or Canadian network. Furthermore, the results reveal that the optimal choice for the Swiss network would be the classic four-parameter fit (model B) with an overall RMS of 2.4 cm. Evidently, two very different conclusions on the optimal model can be drawn from the same data depending on the criteria for testing. It should be mentioned however, that

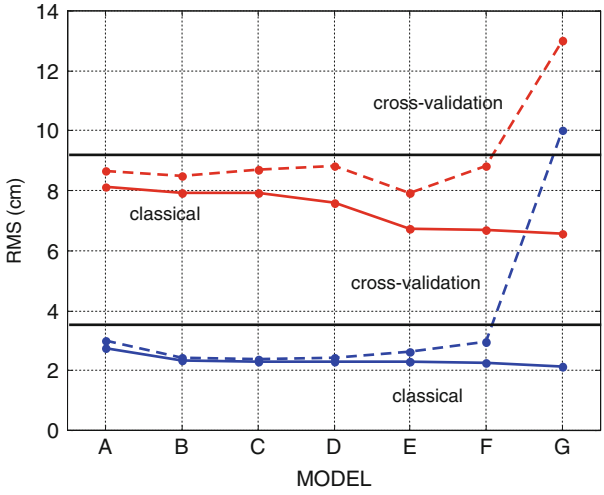


Fig. 11.6 Results of classical empirical testing and cross-validation procedures

in this case there is only a minor difference in performance between each model. This is most likely due to the fact that the data is consistently distributed and exhibits rather small variations from point-to-point. Nonetheless, the use of a parametric model does reduce the original RMS of 4 cm by ∼1 cm to just over 2 cm. It is clear that the four-parameter model is the best choice for reducing the overall range to between −6 and 7 cm.

The performance of the parametric model can also be gauged on the numerical stability over the region of interest. Since, there is not a prominent variation in the achievable fit for each model, the condition number may provide some insight into the overall performance of the model. The computed values reveal that in general, the most stable models are those of the lowest order with fewer unknown parameters. The higher order models tend to be less stable and less accurate when applied at independent control points.

The next step in the assessment process would be to determine if any of the model parameters are insignificant using the procedure described in Sect. 11.6.2 The procedure was carried out, however, it soon became obvious that it was not necessary. The selected model is of low order (2nd) and consists of only four terms. Based on the collective results presented above, it was deemed appropriate to make the final decision of the classic four-parameter fit (model B) for the Swiss network, shown in Fig. 11.7.

In the case of the Alberta/BC network the corresponding measures of goodness of fit, R^2 and R^2_α, depicted in Fig. 11.5, clearly indicates the differential similarity model as a better choice than the third or fourth-order polynomial models according to R^2_α. The inflation caused in R^2 by an increase in the number of parameters from 7, 10 to 15 for the differential similarity, third, and fourth-order polynomial models, respectively, is insignificant compared to the jump experienced from

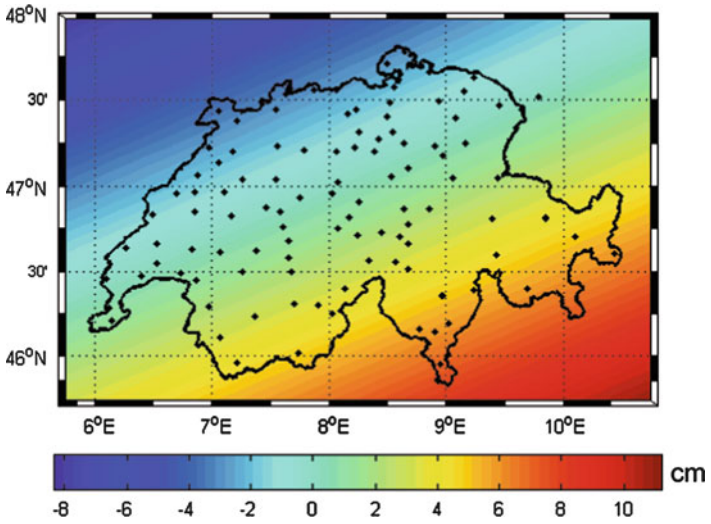


Fig. 11.7 Classic four-parameter surface fit for the Swiss test network

the second-order polynomial (six terms) to the seven-term differential similarity model. The inconclusive negative values obtained for the adjusted coefficient of determination for the lower-order models should also be noted.

Figure 11.6 emphasizes the visible effects of over-parameterization exhibited by the behaviour of the fourth-order polynomial trend surface, which provides a high RMS of 13 cm during cross-validatin. This is even an inferior result to not applying any parametric model. The third-order polynomial model performs close to the original misclosure RMS at the 9 cm-level. The model that gives the best prediction results for the Alberta/BC network is the differential similarity with an RMS of 6.7 cm.

Results after the model fit (classical) and cross-validation empirical tests are illustrated, which clearly identifies the differences in performance between the models and the various conclusions that can be drawn regarding the best model depending on the type of test used. Therefore, the importance of using an independent empirical test such as cross-validation cannot be stressed enough. Often users are wary of such a practice because they would prefer to use as much data as possible for computing the spatial model. However, there is no substitute for model validation. Given the eminent economic benefits, high efficiency and improved achievable accuracy of instituting satellite-based vertical control points, it is only expected that the number of GPS-on-benchmarks will increase over time.

In most practical cases, the truncated/approximated four- or five-parameter versions of the differential similarity model are implemented. To test the significance of the additional parameters in the extended seven-parameter version of the model (model E), the backward elimination procedure was used. The first statistical test is implemented to determine if the fourth, fifth, sixth and seventh parameters are

significant. The hypothesis is set up as follows:

$$H_\circ : \left[x_4 \; x_5 \; x_6 \; x_7 \right]^T = \mathbf{0} \text{ vs. } H_a : \left[x_4 \; x_5 \; x_6 \; x_7 \right]^T \neq \mathbf{0}$$

where,

$$x_4 = \frac{\sin\varphi \cos\varphi \sin\lambda}{W}$$
$$x_5 = \frac{\sin\varphi \cos\varphi \cos\lambda}{W}$$
$$x_6 = \frac{1 - f^2 \sin^2\varphi}{W}$$
$$x_7 = \frac{\sin^2\varphi}{W}.$$

The computed $\tilde{F}$-value was (6.44) compared to the critical value obtained from the statistical tables of $F_{4,56}^{0.05} = 2.54$ and $F_{4,56}^{0.01} = 3.68$ for different levels of significance. In both cases, $\tilde{F} > F_{k,f}^{\alpha}$, and therefore the null hypothesis is rejected suggesting that all of the tested terms are significant. Additional *F-tests* were conducted, testing each of the seven parameters individually, i.e. $H_\circ : x_i = 0, i = 1, \ldots, 7$, and the results indicated that all seven parameters are statistically significant.

One must be cautious with the interpretation of these results as correlation among the model parameters may distort results. Consequently, the model was re-formulated with a *new* set of orthonormal base functions using the Gram-Schmidt process, which gives a *new* set of uncorrelated parameters. Each new parameter was tested for significance by applying the same procedure as above. Surprisingly, it was found that for the orthonormal form of the model, only **two** of the total seven parameters were significant at the 99% confidence level ($\alpha = 0.01$). Using a 95% confidence level ($\alpha = 0.05$), **four** of the seven terms were deemed significant. Table 11.3 summarizes the statistics after the fit for the three versions of the orthonormalized parametric models (i.e., seven, four, two terms). The RMS of fit is on the same level as those achieved using the models given in Table 11.3.

11.7 Remarks

Although not directly used in this chapter, it is important to provide a brief overview of the normal height system as it is the basis of heights in many regions worldwide. The equivalent form of (11.1), which gives the geometrical relationship between the ellipsoidal height, h, normal height, H^*, and height anomaly, ς, is given by

$$h - H^* - \varsigma = 0 \tag{11.40}$$

In this case, the geoid surface is replaced by the quasi-geoid, which is closely related to the geoid and in fact coincides with the geoid in the open seas. An important distinction between the geoid and the quasi-geoid is that the latter is not considered to be an equipotential surface of the Earth's gravity field. Since orthometric, normal and dynamic heights are linked through the geopotential number, it is theoretically possible to convert between any of the three height types. The described combined height adjustment procedure and VCE methodology can be used with normal heights and the quasi-geoid through a straightforward replacement of orthometric heights and the geoid.

Another important remark is on the accuracy of the ellipsoidal, orthometric and geoid heights based on the intrinsic errors that affect the 'initial' CV matrices that are used for the adjustment. Although not directly dwelled on herein, there are several well established references on the accuracy of these height components. For the most part, the computation of geoid heights has been covered in Chap. 6 in this book. The orthometric heights primarily suffer from post-adjustment biases as the random and systematic errors inherent during the leveling process can be rectified. The ellipsoidal heights obtained from GPS measurements (although they can also be acquired through VLBI, SLR, DORIS, other GNSS) is in general more challenging than estimating horizontal coordinates. The most limiting factor remains the very high correlation of receiver clock corrections and tropospheric zenith delay parameters with the ellipsoidal height. The estimation of these effects significantly hinders the achievable accuracy of the height component, even in the absence of other errors and biases. A suggested means for partially decorrelating the height from the receiver clock and tropospheric delay is to take advantage of the zenith dependence and process GPS data at low elevation cut-off angles. Of course, lowering the elevation cut-off introduces other problems with data processing as the noise level increases significantly. Therefore, due to the nature of the satellite configuration and the need to estimate receiver clocks (even differences), the height component is inherently less accurate than the horizontal positions.

Part III
Advanced Analysis Methods
Fernando Sansò

Chapter 12
Hilbert Spaces and Deterministic Collocation

12.1 Outline of the Chapter

It is impossible to study modern geodesy and its computational methods without having at least some basic concepts on Hilbert spaces and the calculus based on the related mathematical theory. In this chapter we collect some definitions and theorems according to what we need in the book.

We shall restrict ourselves to describe separable real Hilbert spaces (HS) and we shall add some basic notions on the theory of HS endowed with reproducing kernels (RKHS), to end up, as main goal, with the so-called deterministic collocation theory, i.e. the theory of optimally estimating a function, in a RKHS, when a certain number of observations on this function are given.

In particular in Sect. 12.2 we run through the standard notions of linear spaces, Banach spaces and their dual, Hilbert spaces, supplying a number of examples useful in the sequel. In Sect. 12.3 we describe more closely the geometry of Hilbert spaces, in particular the concept of orthogonality and orthogonal projection, with its implications on the possibility of representing the dual of the space by the scalar product with its own elements (Riesz theorem (Riesz and Nagy 1965)). We pass then to describe bases, and in particular orthonormal bases. Finally, in Sect. 12.4 we introduce Hilbert spaces with reproducing kernels and their basic properties. The section culminates with the definition of the "best" approximation problem in deterministic sense, or deterministic collocation, giving its functional solution, illustrated by several examples.

We warn the readers not acquainted with HS methods to try to look upon them as a simple generalization of the Euclidean spaces R^n when the dimension n tends to infinity so that in many formulas simple sums become series and only the problem of their convergence has to be taken care of. In this way the basic geometry of HS is perfectly accounted for. What is left out is the intricacy of the functional interpretation of this geometry when the HS is made up of functions, like the famous HS of square integrable functions on some set T, namely $L^2(T)$, requiring Lebesgue theory of measure and integration.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_12,

The lack of this knowledge though, is alleviated here because we shall deal mainly with RKHS, the elements of which are generally better behaved functions and where convergence of sequences implies pointwise convergence too.

The material of this Chapter is covered by any standard text book on functional analysis. In particular one can look at Riesz and Nagy (1965) and Yosida (1978). Reproducing kernel Hilbert spaces of harmonic functions are illustrated in Axler et al. (2001).

12.2 An Introduction to Hilbert Spaces

Definition 1 (Linear space). A real linear space X is a collection of elements, $X \equiv \{x\}$, on which the two operation, $+$, sum of two elements, and $\cdot$, product of an element of the space by a real number, are defined in such a way that

$$
\begin{array}{ll}
(a)\ \forall x, y \in X & \exists z = x + y = y + x \in X \\
(b)\ \forall x, y, z \in X & w = (x + y) + z = x + (y + z) \\
(c)\ \forall \lambda \in R, x \in X & \exists v = \lambda \cdot x = x \cdot \lambda \\
(d)\ \lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y\ , & (\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x
\end{array}
$$

In particular there is an element $0 \in X$ and numbers $(0, 1) \in R$ such that

$$
\begin{array}{l}
(e)\ \ \forall x \in X\ , x + 0 = x\ ;\ 0 \cdot x = 0 \\
(f)\ \forall x \in X\ \ 1 \cdot x = x\ .
\end{array}
$$

Due to the third of (e) we can use the same symbol for the null element in X and the zero in R.

Definition 2 (Linear independence). The n elements of a finite set $\{x, x_2 \ldots, x_n\} \in X$ are said to be linearly independent if

$$
\sum_{k=1}^{n} \lambda_k x_k = 0 \Rightarrow \lambda_k = 0,\ k = 1, \ldots n\ . \tag{12.1}
$$

It is obvious that none of the x_i in (12.1) can be the null element.

Definition 3 (Span). Fix n independent elements $\{x_k, k = 1 \ldots n\}$, then the set of elements

$$
\left\{ X = \sum_{k=1}^{n} \lambda_k x_k\ ;\ \{\lambda_k\} \in R^n \right\} \tag{12.2}
$$

is called the Span $\{x_1 \ldots x_n\}$. It is obvious that (12.2) is itself a linear space, in fact it is a subspace of X.

Remark 1 (Tensor notation). In order to use a more synthetic notation we shall put

$$\mathbf{x} = \begin{vmatrix} x_1 \\ \vdots \\ x_n \end{vmatrix}, \ (x_i \in X), \mathbf{x} \in X^{(n)} \ ; \ \boldsymbol{\lambda} = \begin{vmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{vmatrix}, (\lambda_i \in R) \ \boldsymbol{\lambda} \in R^n \tag{12.3}$$

and work with the ordinary algebraic rules with these objects, so Definition (3) will be represented also by

$$\text{Span}\{\mathbf{x}\} \equiv \{\boldsymbol{\lambda}^t \mathbf{x}; \ \boldsymbol{\lambda} \in R^n\}. \tag{12.4}$$

Definition 4 (Dimension of X). We say that X has dimension N if there is $\mathbf{x} \in X^{(N)}$, a tensor constituted of N independent vectors, such that

$$X \equiv \text{Span}\{\mathbf{x}\} \ , \ (\mathbf{x} \in X^{(N)}); \tag{12.5}$$

if $\forall \mathbf{x} \in X^{(N)}, \forall N$, with $\mathbf{x}$ an independent N-tuple,

$$\text{Span}\{\mathbf{x}\} \subset X, \tag{12.6}$$

the inclusion being in strict sense, we say that X is infinite dimensional.

Definition 5 (Basis). $\{\mathbf{x}\}$ in (12.5) is called a basis of X. Any other $\{\mathbf{y}\} \in X^{(N)}$ such that

$$\mathbf{y} = \Lambda \mathbf{x}; \ Y_i = \sum_{k=1}^{N} \lambda_{ik} x_k, \tag{12.7}$$

with Λ an $N \times N$ invertible matrix, is a basis of X.

Proposition 1. *Any N-tuple of independent vectors $\mathbf{y}$ in a linear space of dimension N is a basis, i.e. there is an invertible Λ such that (12.7) holds.*

Proof. In fact $\forall \mathbf{y} \in X^{(N)}$ one can write $\mathbf{y} = \Lambda \mathbf{x}$ because $\mathbf{x}$ is a basis by definition. If Λ is not invertible, the same is true for Λ^t, and there is a vector $\boldsymbol{\mu} \neq 0 \in R^n$ such that $\Lambda^t \boldsymbol{\mu} = 0$. But then $\boldsymbol{\mu}^t \mathbf{y} = \boldsymbol{\mu}^t \ \Lambda \mathbf{x} = (\Lambda^t \boldsymbol{\mu})^t \ \mathbf{x} = 0$ without being $\boldsymbol{\mu} = 0$, contrary to the hypothesis of independence of $\mathbf{y}$. □

Proposition 2. *Every $N + 1$ elements in a linear space X of dimension N are linearly dependent.*

Example 1. The set of the real polynomials of degree N in the real variate t, $\{P_n(t)\} \equiv \mathcal{P}_n^1$ is a linear space with dimension

$$N = n + 1.$$

In fact for any $P_n(t)$ we can write

$$P_n(t) = a_0 + a_1 t + \ldots + a_n t^n,$$

which shows that

$$\mathbf{x} = \begin{vmatrix} 1 \\ t \\ \vdots \\ t^n \end{vmatrix}$$

is indeed a basis of $\mathcal{P}_n^1$.

Example 2. The space $\mathcal{P}_n^2$ of polynomials of degree n in 2 variables (x, y) has a more complicated structure.

Since every $P_n(x, y)$ is the sum of monomials of degrees $0, 1 \ldots n$, we can first decompose $\mathcal{P}_n^2 = H_0^2 + H_1^2 + \ldots H_n^2$, where each H_k^2 contains only polynomials homogeneous of degree k.

So

$$\begin{aligned} &P_0(x, y) \in H_0^2 \rightarrow P_0(x, y) = a_0 \\ &P_1(x, y) \in H_1^2 \rightarrow P_1(x, y) = a_1 x + b_1 y \\ &P_2(x, y) \in H_2^2 \rightarrow P_2(x, y) = a_2 x^2 + b_2 xy + c_2 y^2 \end{aligned}$$

and so forth.

We easily see that

$$\dim H_k^2 = k + 1$$

so that we have

$$\dim \mathcal{P}_n^2 = \sum_{k=0}^{n} (k + 1) = \frac{(n + 1)(n + 2)}{2}$$

Definition 6 (Linear functional). A function $L : X \rightarrow R$ such that

$$\forall x, y \in X,\ \forall \lambda, \mu \in R,\ L(\lambda x + \mu y) = \lambda L(x) + \mu L(y) \tag{12.8}$$

is a linear functional on X.

Definition 7 (Algebraic dual). The set of linear functionals on X, is also a linear space X' with the linear combination rule

$$\forall L, M \in X',\ \forall a, b \in R,\ (aL + bM)(x) = aL(x) + bM(x); \tag{12.9}$$

X' is called the algebraic dual of X.

Proposition 3. *X' has the same dimensionality as X.*

Example 3. Let $X \equiv C_b(T)$ be the space of all bounded, continuous functions on T; then the so-called evaluation functional

$$\forall x \in X,\ \mathrm{ev}_{\bar{t}}\{x(t)\} = x(\bar{t})$$

is a linear functional on X.

Let $X \equiv L_1(T)$ be the space of all measurable functions, integrable over T, then $\forall A$ measurable $\subset T$, $\forall x \in X,\ I_A(x) = \int_A x(t)dt$, is a linear functional on X.

Definition 8 (Norm). A norm on a linear space X is a (non-linear) functional $X \to R^+;\ \|x\|$, such that

$$\begin{aligned}&(a)\ \|x\| \geq 0\ ,\ \|x\| = 0 \ \Leftrightarrow\ x = 0\\ &(b)\ \|\lambda x\| = |\lambda|\|x\|\\ &(c)\ \|x + y\| \leq \|x\| + \|y\|;\end{aligned}$$

due to (a)–(c), $\|x - y\|$ has the properties of a distance.

Definition 9 (Cauchy sequence). A sequence $\{x_n\}$ in X is said to be Cauchy if it satisfies the Cauchy condition

$$\forall \varepsilon > 0, \exists N_\varepsilon;\ \forall n, m > N_\varepsilon\ ;\ \|x_n - x_m\| < \varepsilon$$

or, said in another way

$$\lim_{n\to\infty} \|x_n - x_{n+p}\| = 0$$

uniformly in p.

Definition 10 (Banach space). A normed space X is called complete if $\forall\{x_n\}$ that is Cauchy there is a limit in the space, namely if $\exists x \in X$ such that

$$\lim_{n\to\infty} \|x_n - x\| = 0;$$

a complete normed space is called a Banach space (BS).

Definition 11 (Bounded linear functionals). In X' we can introduce a norm, also called the dual norm of $\|\ \|$,

$$L \in X'\ ,\ \|L\| = \sup_{\|x\|=1} |L(x)|; \tag{12.10}$$

a linear functional L is said to be bounded if

$$\|L\| < +\infty; \tag{12.11}$$

the set of bounded linear functionals

$$X^* \equiv \{L \in X';\ \|L\| < +\infty\} \tag{12.12}$$

is a subspace of X'.

Proposition 4. *X^* is always complete, i.e. it is a BS, whether X is Banach (i.e. complete) or not.*

Proposition 5. *If X is an N-dimensional BS, then it enjoys the Weiestrass property, i.e. every bounded sequence $\{x_n\}$ ($\|x_n\| \le c < +\infty$) has a convergent subsequence. Viceversa if X is a BS and has the Weiestrass property, then it is necessarily finite dimensional.*

Definition 12 (Scalar product). Let X be a real linear space, then a real scalar product on X is a bilinear functional, $< , >$, of $X^{(2)} \to R$ with the following properties:

$$\begin{aligned}
&(a) < x, y > = < y, x > \\
&(b) < \lambda x + \mu y, z > = \lambda < x, z > + \mu < y, z > \\
&(c) < x, x > \ge 0\ ;\ < x, x > = 0 \iff x = 0
\end{aligned}$$

Proposition 6 (Schwarz). *One has the inequality*

$$| < x, y > | \le \sqrt{< x, x >< y, y >}; \tag{12.13}$$

equality holds only if $y = \lambda x$ for some $\lambda \in R$.

Proof. Since $\forall \lambda \in R$

$$\begin{aligned}
O \le P(\lambda) &= < x - \lambda y, x - \lambda y > \\
&= < x, x > -2\lambda < x, y > + \lambda^2 < y, y >,
\end{aligned} \tag{12.14}$$

then the discriminant of $P(\lambda)$ has to be negative, i.e.

$$< x, y >^2 - < x, x >< y, y > \le 0;$$

in particular, from Definition 12(c), $P(\lambda)$ is zero for some λ only if $x = \lambda y$, for the same λ. □

Proposition 7. *If X is endowed with the scalar product $< , >$ then it is normed with*

$$\|x\| = \sqrt{< x, x >}. \tag{12.15}$$

Definition 13 (Hilbert space). If X, endowed with the scalar product $<\ ,>$, is a BS with respect to the norm (12.15), then it is called a Hilbert Space (HS). In particular a HS is always complete.

Example 4. Let W be a $n \times n$ square, strictly positive, symmetrical matrix. Then R^n is a HS with the scalar product

$$\forall \boldsymbol{\lambda}, \boldsymbol{\mu} \in R^n,\ <\boldsymbol{\lambda}, \boldsymbol{\mu}> \equiv \boldsymbol{\lambda}^t W \boldsymbol{\mu}.$$

Remark 2. The example above shows that the same space can bear many equivalent HS structures in the sense that, denoting $<\ ,>_W$ the scalar product with weight W, we have, for some positive α and β,

$$\alpha <\boldsymbol{\lambda}, \boldsymbol{\lambda}>_I = \alpha \boldsymbol{\lambda}^t \boldsymbol{\lambda} \leq <\boldsymbol{\lambda}, \boldsymbol{\lambda}>_W = \boldsymbol{\lambda}^t W \boldsymbol{\lambda} \leq \beta \boldsymbol{\lambda}^t \boldsymbol{\lambda} = \beta <\boldsymbol{\lambda}, \boldsymbol{\lambda}>_I . \quad (12.16)$$

Condition (12.16) guarantees that Cauchy sequences and limits with weight W or with weight I are always the same.

Example 5. let $\boldsymbol{\lambda} \in R^\infty$, i.e. the element $\boldsymbol{\lambda} \equiv \{\lambda_1, \lambda_2 \ldots\}$ is just a sequence of real numbers. We define the space ℓ^2 as the subspace of R^∞ of those vectors for which

$$\boldsymbol{\lambda} \in \ell^2\ ;\ |\boldsymbol{\lambda}|^2 = \sum_{n=1}^{+\infty} \lambda_n^2 < +\infty. \quad (12.17)$$

We define in ℓ^2 the scalar product

$$\boldsymbol{\lambda}, \boldsymbol{\mu} \in \ell^2 \quad <\boldsymbol{\lambda}, \boldsymbol{\mu}>_{\ell^2} = \boldsymbol{\lambda}^t \boldsymbol{\mu} = \sum_{n=1}^{+\infty} \lambda_n \mu_n \quad (12.18)$$

so that we have as corresponding norm

$$\|\boldsymbol{\lambda}\|_{\ell^2} = |\boldsymbol{\lambda}|. \quad (12.19)$$

ℓ^2 is a HS, as you are invited to prove in Exercise 4.

Example 6. Let T denote any set (for instance $T \equiv R$ or the surface of the unit sphere S_1), $d\mu$ a measure on T (for instance $d\mu$ = the Lebesgue measure on R, or $d\mu = d\sigma = \sin\vartheta d\vartheta d\lambda$ on S_1 in spherical coordinates (ϑ, λ)) and with $\mathcal{M}(T)$ the linear space of measurable real functions $f(t)$ defined on T (e.g. $f(\vartheta, \lambda)$ on S_1). Remark that a function $f(t)$ is measurable on T if the sets $\{t;\ f(t) \leq a\}$ are measurable $\forall a \in \mathcal{R}$. Consider the subspace of those functions $f(t)$ such that

$$\int_T f^2(t) d\mu(t) < +\infty, \quad (12.20)$$

with the usual identification that f and g are the same "functions" if they coincide μ-almost everywhere

$$f(t) \equiv g(t) \Longleftrightarrow \mu\{t\ ;\ f(t) \neq g(t)\} = 0\ . \tag{12.21}$$

Then this subspace of $\mathcal{M}(T)$, that we denote $L^2(T)$, is a HS with the scalar product

$$< f, g >_{L^2(T)} = \int_T f(t)g(t)d\mu(t). \tag{12.22}$$

For instance, when $T \equiv S_1$, we use, systematically trough the book,

$$< f, g >_{L^2(S_1)} = \frac{1}{4\pi} \int_{S_1} f(\vartheta, \lambda)g(\vartheta, \lambda)d\sigma. \tag{12.23}$$

Example 7. In principle this example is a particular case of the previous one, however it is so relevant to the matter of the book that we prefer to present it in explicit form.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, where $\mathcal{A}$ is the σ-algebra of the (measurable) events and P a probability measure on $\mathcal{A}$. Let $\mathcal{M}(\Omega)$ be the space of random variables defined on $(\Omega, \mathcal{A}, P)$ i.e. of functions $X(\omega)$ measurable with respect to P; we denote as $\mathcal{L}^2(\Omega)$ the subspace of $\mathcal{M}(\Omega)$ of those variables which have finite second moment, i.e.

$$\forall X(\omega) \in \mathcal{L}^2(\Omega);\ E\{X^2\} = \int X^2(\omega)dP(\omega) < +\infty; \tag{12.24}$$

this is indeed a HS with scalar product

$$\forall X(\omega), Y(\omega) \in \mathcal{L}^2(\Omega); \quad < X, Y >_{\mathcal{L}^2(\Omega)} = E\{XY\}. \tag{12.25}$$

Notice that the use of the average operator avoids using integral symbols. Observe also that if one restricts $X(\omega)$ to the subspace of $\mathcal{L}^2(\Omega)$ constituted by random variables with zero average, that we call $\mathcal{L}^2_0(\Omega)$, scalar product and norm are related to covariance and variance in the sense that, since $E\{X\} = E\{Y\} = 0$,

$$\forall X, Y \in \mathcal{L}^2_0(\Omega); \quad < X, Y >= E\{XY\} = \sigma_{XY} = C(X, Y)$$
$$\|X\|^2 = E\{X^2\} = \sigma^2_X.$$

Notice as well that convergence in $\mathcal{L}^2(\Omega)$ of a sequence $\{X_n\}$ to some random variable X is just the usual convergence in mean square sense, i.e.

$$\left(X_n \underset{\mathcal{L}^2(\Omega)}{\longrightarrow} X\right) \Longleftrightarrow (E\{(X_n - X)^2\} \to 0).$$

12.3 Orthogonality, Duality, Bases

Definition 14 (Orthogonality). Let X be a HS; two elements $x, y \in X$ are said to be orthogonal if

$$< x, y >= 0. \tag{12.26}$$

Moreover let S be a linear subspace of X; we say that x is orthogonal to S, $(x \perp S)$, if

$$< x, y >= 0, \quad \forall y \in S. \tag{12.27}$$

The following results are so fundamental in approximation theory and its applications, illustrated in the book, that we state them in the form of a Theorem and its Corollary.

Theorem 1 (Orthogonal projection). *Let X be a HS and S a closed subspace strictly contained in X; then, given any $x \in X$ there is one and only one $\widehat{x} \in S$ such that the orthogonal decomposition*

$$\begin{cases} x = \widehat{x} + v \\ \widehat{x} \in S, \ v \perp S \end{cases} \tag{12.28}$$

is verified. Furthermore $\widehat{x}$ is the point of S which turns out to be closest to x, i.e.

$$\widehat{x} = \arg\min_{y \in S} \|x - y\|^2. \tag{12.29}$$

Proof. Let us put

$$d^2 = \inf_{y \in S} \|x - y\|^2. \tag{12.30}$$

Since a norm is never a negative number, the Inf in (12.30) exists and it is $d^2 \geq 0$. If $d^2 = 0, x \in S$ (remember that S is closed) and there is nothing to prove. If $d^2 > 0$, let $\{y_n\}$ be an extremizing sequence, i.e. one for which

$$\lim_{n \to \infty} \|x - y_n\|^2 = d^2 \ , \ \{y_n\} \in S. \tag{12.31}$$

We shall prove that $\{y_n\}$ is Cauchy, therefore $\exists \lim_{n \to \infty} y_n = \widehat{x}$ and it has to be $\widehat{x} \in S$ because S is closed. At this point (12.29) will be proved.

Then we have to prove that $\{y_n\}$ is Cauchy. First note that if $\{y_n\}$ satisfies (12.31), then also

$$\forall p, \quad \left\| x - \frac{1}{2}(y_n + y_{n+p}) \right\| \to d; \tag{12.32}$$

in fact

$$\begin{aligned} d &\le \left\| x - \frac{1}{2}(y_n + y_{n+p}) \right\| = \left\| \frac{1}{2}(x - y_n) + \frac{1}{2}(x - y_{n+p}) \right\| \\ &\le \frac{1}{2}\|x - y_n\| + \frac{1}{2}\|x - y_{n+p}\|. \end{aligned} \tag{12.33}$$

Since the RHS of (12.33) tends to d when $n \to \infty$, (12.32) must be true. Note that this implies that the limit (12.32) is uniform in p.

Now let us apply the result (12.140) of Exercise 6; substituting x with $x - y_n$ and y with $x - y_{n+p}$, we obtain

$$\begin{aligned} &\|2x - (y_n + y_{n+p})\|^2 + \|y_n - y_{n+p}\|^2 \\ &= 2\|x - y_n\|^2 + 2\|x - y_{n+p}\|^2. \end{aligned} \tag{12.34}$$

The first term at the left tends to $4d^2$ because of (12.32); the two terms at the right end tend both to $2d^2$ because of (12.31). Then we have too

$$\|y_n - y_{n+p}\|^2 \underset{n\to\infty}{\longrightarrow} 0, \tag{12.35}$$

uniformly in p, i.e. $\{y_n\}$ is Cauchy.

Therefore it is true that, setting $\widehat{x} = \lim y_n \in S$, $\forall y \in S$, $\|x - y\|^2 \ge \|x - \widehat{x}\|^2$.

Put

$$y = \widehat{x} + th \ \in S, \quad \forall h \in S,$$

and compute

$$P_2(t) = \|x - y\|^2 = \|x - \widehat{x}\|^2 - 2t < x - \widehat{x}, h > +t^2\|h\|^2; \tag{12.36}$$

note that $P_2(t)$ must have a minimum at $t = 0$, so that

$$-P_2'(0) =< x - \widehat{x}, h >= 0\,, \quad \forall h \in S. \tag{12.37}$$

This proves (12.28), namely that $v = x - \widehat{x}$ is orthogonal to S.

Finally we argue that the decomposition (12.28) is unique, i.e. if there are $\widehat{x}_1, \widehat{x}_2 \in S$ and $v_1, v_2 \perp S$ such that

$$x = \widehat{x}_1 + v_1 = \widehat{x}_2 + v_2, \tag{12.38}$$

then it has to be

$$\widehat{x}_1 = \widehat{x}_2 \ , \ v_1 = v_2.$$

In fact from (12.38) we have

$$\widehat{x}_1 - \widehat{x}_2 = v_2 - v_1; \tag{12.39}$$

but this is obviously possible only if $\widehat{x}_1 - \widehat{x}_2 = 0$ and $v_1 - v_2 = 0$, as otherwise the two vectors are in orthogonal spaces. □

Corollary 1. *Assume that S is a finite dimensional subspace of X, with an N dimensional basis* $\mathbf{x}$*,*

$$S \equiv \{\boldsymbol{\lambda}^t \mathbf{x};\ \boldsymbol{\lambda} \in R^N\}; \tag{12.40}$$

then, denoting with $\widehat{x}$ the orthogonal projection of x on S, with

$$G \equiv < \mathbf{x}, \mathbf{x}^t > \equiv \{< x_i \ , \ x_k >\} \tag{12.41}$$

the Gramian of the basis $\mathbf{x}$*, with*

$$\mathbf{w} = G^{-1}\mathbf{x} \tag{12.42}$$

the so-called dual basis of $\mathbf{x}$*, we have, $\forall x \in X$,*

$$\begin{aligned} \widehat{x} &= < x, \mathbf{x}^t > G^{-1}\mathbf{x} \equiv < x, \mathbf{x}^t > \mathbf{w} = < x, \mathbf{w}^t > \mathbf{x} \\ &= \Sigma_i < x, x_i > w_i \equiv \Sigma_i < x, w_i > x_i. \end{aligned} \tag{12.43}$$

In addition the squared approximation error $\mathcal{E}^2$, i.e. the square of the norm of the residual $v = x - \widehat{x}$ can be computed by

$$\mathcal{E}^2 = \|x - \widehat{x}\|^2 = \|x\|^2 - < x, \mathbf{x}^t > G^{-1} < \mathbf{x}, x > \tag{12.44}$$

Proof. We first prove that G is invertible, so that (12.42) is well-defined.

Remember that the components of $\mathbf{x}$ are linearly independent by hypothesis, then

$$\begin{aligned} &G\boldsymbol{\lambda} = 0 \Rightarrow \boldsymbol{\lambda}^t G \boldsymbol{\lambda} = \boldsymbol{\lambda}^t < \mathbf{x}, \mathbf{x}^t > \boldsymbol{\lambda} = 0 \\ &\Rightarrow < \boldsymbol{\lambda}^t \mathbf{x}, \mathbf{x}^t \boldsymbol{\lambda} > = \|\boldsymbol{\lambda}^t \mathbf{x}\|^2 = 0 \rightarrow \boldsymbol{\lambda} = 0; \end{aligned}$$

therefore G is invertible.

Now put

$$\widehat{\mathbf{x}} = \boldsymbol{\lambda}^t \mathbf{x} = \mathbf{x}^t \boldsymbol{\lambda} \in S, \tag{12.45}$$

for some $\boldsymbol{\lambda}$, and

$$h = \boldsymbol{\mu}^t \mathbf{x}\,,\ \forall \boldsymbol{\mu}$$

in (12.37) to find the identity in $\boldsymbol{\mu}$

$$\begin{aligned} \forall \boldsymbol{\mu}, \qquad 0 &= < h, x - \widehat{x} > = < \boldsymbol{\mu}^t \mathbf{x}\,,\ x - \mathbf{x}^t \boldsymbol{\lambda} > \\ &= \boldsymbol{\mu}^t \{ < \mathbf{x}, x > - < \mathbf{x}, \mathbf{x}^t > \boldsymbol{\lambda} \} \end{aligned}$$

so that

$$G\boldsymbol{\lambda} = < \mathbf{x}, x > \Rightarrow \boldsymbol{\lambda} = G^{-1} < \mathbf{x}, x > . \tag{12.46}$$

Recall that G is a symmetric matrix; hence using (12.46) in (12.45) we receive

$$\begin{aligned} \widehat{x} &= < x, \mathbf{x}^t > G^{-1} \mathbf{x} = < x, (G^{-1}\mathbf{x})^t > \mathbf{x} \\ &= \Sigma_{jk} < x, x_j > \{ < x_j, x_k > \}^{(-1)} x_k \end{aligned} \tag{12.47}$$

coinciding with the two forms of (12.43).

Finally, observing that, due to the orthogonality of the vectors $\widehat{x}$ and v, it holds

$$\|x\|^2 = \|\widehat{x}\|^2 + \|v\|^2,$$

we find, using (12.47),

$$\begin{aligned} \mathcal{E}^2 = \|v\|^2 &= \|x\|^2 - < x, x^t > G^{-1} < \mathbf{x}, \mathbf{x}^t > G^{-1} < \mathbf{x}, x > \\ &= \|x\|^2 - < x, \mathbf{x}^t > G^{-1} < \mathbf{x}, x > \end{aligned}$$

as it was to be proved. □

Definition 15 (Orthogonal complement). Let S be a subspace of X; the set of elements y which are orthogonal to S is a closed linear subspace of X which is called the orthogonal complement of S and denoted $S^{\perp}$. The – possibly finite – dimension of $S^{\perp}$ is called the co-dimension of S.

Remark 3. Note that when S is closed, Theorem 1 can be re-frased as: given any $x \in X$ there is one and only one decomposition

$$x = \widehat{x} + v\ \ \widehat{x} \in S,\ v \in S^{\perp};$$

furthermore, since

$$\|x\|^2 = \|\widehat{x}\|^2 + \|v\|^2 \Rightarrow \|\widehat{x}\| \leq \|x\|,\ \|v\| \leq \|x\|$$

we claim that both $\widehat{x}, v$ depend with continuity on x.

Proposition 8. *Given a bounded (and then continuous) and not null linear functional L on X, its null space, i.e. the set*

$$S_0 \equiv \{x\ ;\ L(x) = 0\},$$

is a closed subspace of codimension 1.

Proof. That S_0 is closed is obvious from the continuity of L. Then take any x such that $L(x) \neq 0$; put $x = x_0 + h$ with $x_0 \in S_0$ and $h \in S_0^{\perp}$, so that $L(x) = L(h) \neq 0$. We have to prove that h is unique, up to a multiplicative constant. Assume there is another $h' \in S_0^{\perp}$ so that $L(h') \neq 0$. Then any not-null linear combination $ah+bh' \in S_0^{\perp}$ and therefore it has to be $L(ah+bh') \neq 0$, unless $ah+bh' = 0$. Take $a = L(h')$ and $b = -L(h)$ to find that

$$aL(h) + bL(h') = 0 \Rightarrow L(ah + bh') = 0;$$

but then $h' = -\frac{a}{b}h$. □

Theorem 2 (Riesz representation theorem). *Given any bounded linear functional L on the HS, X, there is one and only one vector y_L that allows the representation of L in terms of scalar product*

$$L(x) =< y_L, x >, \quad \forall x \in X \tag{12.48}$$

Proof. That y_L is unique is obvious as

$$< y_L, x >=< y'_L, x >, \quad \forall x \ \Rightarrow < y_L - y'_L, x >= 0, \quad \forall x \ \Rightarrow y_L = y'_L.$$

To find y_L call S_0 the null space of L and h a vector of unit norm defining $S_0^{\perp}$; since $\forall x,\ x = x_o + < x, h > h,\ (x_0 \in S_0)$, we have $L(x) = L(< x, h > h) = =< x, h > L(h) =< x, L(h)h >$. Put $y_L = L(h)h$. □

Definition 16 (Total family). A family $\mathcal{T} \equiv \{y\} \subset X$ is said to be total in X if

$$< x, y >= 0\ ,\ \ \forall y \in \mathcal{T} \Rightarrow x = 0. \tag{12.49}$$

Definition 17 (Separability). We say that a HS, X is separable if there is a sequence $\mathbf{y} = \{y_n\}$ which is total. It is clear that we can always assume that each y_n is linearly independent of $y_1, y_2 \ldots y_{n-1}$.

Definition 18 (Span of a sequence). Call R_0^{∞} the subspace of R^{∞} constituted by sequences $\boldsymbol{\lambda} \equiv \{\lambda_1, \lambda_2 \ldots\}$, which have null components $\forall n > N$; here N is varying with $\boldsymbol{\lambda}$.

Let a sequence $\mathbf{y} = \{y_n\}\ n = 1, 2, \ldots$ be given; we define

$$\text{Span}\{\mathbf{y}\} = \{y = \boldsymbol{\lambda}^t \mathbf{y};\ \boldsymbol{\lambda} \in R_0^{\infty}\}. \tag{12.50}$$

Observe also that if we define

$$Y^N \equiv \{y = \boldsymbol{\lambda}^t \mathbf{y},\ \forall \boldsymbol{\lambda} \in R_0^\infty\ ,\ \lambda_n = 0,\ n > N \text{ fixed}\} \tag{12.51}$$

we have

$$\text{Span}\{\mathbf{y}\} = \bigcup\nolimits_N Y^N. \tag{12.52}$$

In general Span $\{\mathbf{y}\}$ is not a closed subspace of X.

Proposition 9. *The sequence* $\mathbf{y} \equiv \{y_n\}$ *is total if and only if Span*$\{\mathbf{y}\}$ *is everywhere dense in* X *or, said in another way, if and only if the closure of Span* $\{\mathbf{y}\}$ *is* X.

Proof. Assume $\mathbf{y}$ to be total and call $\widetilde{X} = \overline{\text{Span}\{\mathbf{y}\}} = \overline{\bigcup_N Y^N}$; it is clear that $\widetilde{X}$ is a closed subspace of X. If $\widetilde{X} \subset X$ strictly, then there is $x \in X, x \notin \widetilde{X}$ and we can put, in view of the Theorem 1,

$$x = \widetilde{x} + h,\ \widetilde{x} \in \widetilde{X},\ h \in \widetilde{X}^\perp,\ h \neq 0;$$

but then

$$< h, y_n >= 0,\ \forall n \tag{12.53}$$

because $y_n \in \widetilde{X}$. If $\{y_n\}$ is total, (12.53) implies $h = 0$, contrary to $\widetilde{X} \subset X$ strictly. On the other hand if $\overline{\text{Span}\{\mathbf{y}\}} \equiv X$, then $\forall x \in X$ fixed we can find a sequence $x_\ell \in \text{Span}\{\mathbf{y}\}$ such that $x_\ell \to x$. Therefore if $< x, y_n >= 0, \forall n$, it has to be $< x, x_\ell >= 0, \forall \ell$ too; but then

$$< x, x >= \lim_{\ell \to \infty} < x, x_\ell >= 0,$$

namely $x = 0$ and $\{\mathbf{y}_n\}$ is total as it was to be proved. □

Remark 4. We notice that, owing to Proposition 9, if $\{y_n\}$ is total, $\forall x$ we can find a sequence

$$y^N = \sum_{n=1}^{N} \lambda_n^N y_n \tag{12.54}$$

such that

$$\|x - y^N\| \underset{N \to \infty}{\longrightarrow} 0. \tag{12.55}$$

This is obviously an important fact in approximation theory.

Note that in (12.54) it is clearly stated that in general the λ coefficients depend on N too; so the existence of the limit (12.55) does not mean at all that there are series of $\{y_n\}$ by which we can represent all x.

Definition 19 (Orthonormal Complete Sequence-ONC). A sequence $\{y_n\}$ is said to be ONC, if

$$\begin{aligned} &\text{(a)} \ < y_n, y_j > = \delta_{nj} \\ &\text{(b)} \ \{y_n\} \text{ is total.} \end{aligned}$$

The sequence $\{y_n\}$ in this case is also called an *orthonormal* (ON) basis of X.

Let us observe that property (a) in Definition 19 means that the Gramian G of an ON basis $\mathbf{y} = \{y_n\}$ is

$$G = I \Rightarrow G^{-1} = I. \tag{12.56}$$

Accordingly, the vector $\widehat{x}^N$, which is the projection of x on $Y^N = \text{Span}\{\mathbf{y}^N\} = \text{Span}\{y_n, n \leq N\}$, is given by (cf. (12.47))

$$\widehat{x}^N = \left\langle x, \left(\mathbf{y}^N\right)^t \right\rangle \mathbf{y}^N = \sum_{k=1}^{N} < x, y_k > y_k \tag{12.57}$$

Proposition 10 (Fourier). *Let $\{y_n\}$ be an ON basis of X, then*

$$\forall x \in X \ , \quad x = \sum_{n=1}^{+\infty} < x, y_n > y_n \equiv \left\langle x, \mathbf{y}^t \right\rangle \mathbf{y} \tag{12.58}$$

the series being convergent in the sense of X; furthermore (Parseval's identity)

$$\|x\|^2 = \sum_{n=1}^{+\infty} < x, y_n >^2 = < x, \mathbf{y}^t > < \mathbf{y}, x > \tag{12.59}$$

and $\forall x, \forall z \in X$

$$< x, z > = \sum_{n=1}^{+\infty} < x, y_n > < z, y_n > = < x, \mathbf{y}^t > < \mathbf{y}, z > . \tag{12.60}$$

Proof. Recall that according to Remark 4, $\exists y^N \in Y^N$ such that $y^N \to x$ in X; but then, since $\widehat{x}^N$ is the projection of x on Y^N,

$$\|x - \widehat{x}^N\| \leq \|x - y^N\| \to 0.$$

together with (12.57), this proves (12.58). Equations 12.59 and 12.60 are easily proved considering that even for the full sequence $\mathbf{y}$ we have $< \mathbf{y}, \mathbf{y}^t >= I$, this last being the identity matrix in R^∞. □

Remark 5. Note that any separable HS has an ONC basis. In fact if X is separable there is a sequence $\{v_n\}$ which is total. Let us call $V_N = \text{Span}\{v_n, n \leq N\}$. Then we can define P_N, an orthogonal projection operator on V_N; further put

$$x_n = v_n - P_{n-1}v_n \ , \ y_n = \frac{x_n}{\|x_n\|};$$

since $v_n \notin V_{n-1}$ by hypothesis, $\|x_n\| \neq 0$ always. Then $\{y_n\}$ is ON; moreover, since

$$\text{Span}\{y_n, n \leq N\} \equiv V_N,$$

$\{y_n\}$ is also complete.

Remark 6. Fourier's theorem (Proposition 10) generalizes to any separable HS, X, the Euclidean vector calculus of R^n. Moreover there is another very important consequence of (12.59), i.e. of Parseval's identity; we see that given any $x \in X$ we can define an infinite vector $\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda} =< x, \mathbf{y} >; \ |\boldsymbol{\lambda}|^2_{\ell^2} = \|x\|^2_X. \tag{12.61}$$

In other words any ON basis in X determines an isometric map of X into ℓ^2; viceversa given any $\boldsymbol{\lambda} \in \ell^2$ we can define the inverse mapping

$$x = \boldsymbol{\lambda}^t \mathbf{y}; \ \|x\|^2_X = |\boldsymbol{\lambda}|_{\ell^2} \tag{12.62}$$

This basically means that all separable HS have ℓ^2 as an isometric image, so that any property can be proved by looking at its representation in ℓ^2. Another useful remark related to Proposition 10 is that if $\{y_n\}$ is just ON but not necessarily complete in X, then it will be by definition complete in $Y \equiv \left[\bigcup_N Y^N\right] \subset X$, which is a closed subspace of X. Therefore $\forall x \in X$ we can write

$$x = \widehat{x} + h, \ \widehat{x} \in Y, \ h \in Y^\perp$$

to the effect that

$$\widehat{x} = \sum_{k=1}^{+\infty} < \widehat{x}, y_k > y_k = \sum_{k=1}^{+\infty} < x, y_k > y_k. \tag{12.63}$$

But then

$$\|\widehat{x}\| = \sum_{k=1}^{+\infty} < \widehat{x}, y_k >^2 \leq \|x\|^2, \tag{12.64}$$

known as *Parseval's inequality*.

Proposition 11. *We give here a useful sufficient condition for the ON sequence* $\{y_n\}$ *to be complete in* X. *Let* $\widetilde{X}$ *be a linear subspace densely embedded in* X. *This property can be verified by proving that* $\widetilde{X}$ *is total for* X. *Assume further that* $\{y_n\}$ *is an ON basis for* $\widetilde{X}$, *in the sense that* $\forall \widetilde{x} \in \widetilde{X}$

$$\widetilde{x} = \sum_{k=1}^{+\infty} < \widetilde{x}, y_k > y_k, \tag{12.65}$$

the series being convergent in the X *norm. Then* $\{y_n\}$ *is total and hence complete in* X.

Proof. Let $\{\widetilde{x}_n\} \in \widetilde{X}$ be such that $\widetilde{x}_n \to x$. Let further $x \in X$ be such that

$$< x, y_k >= 0, \ \forall k \tag{12.66}$$

But then, recalling (12.65), we find

$$\begin{aligned} \|\widetilde{x}_n\|^2 = \sum_{k=1}^{+\infty} < \widetilde{x}_n, y_k >^2 &= \sum_{k=1}^{+\infty} < \widetilde{x}_n - x, y_k >^2 \\ &\leq \|\widetilde{x}_n - x\|^2 \end{aligned} \tag{12.67}$$

because of Parseval's inequality (12.64).

Equation 12.67 implies $\widetilde{x}_n \to 0$, i.e. $x = 0$. So (12.66) implies $x = 0$, i.e. $\{y_n\}$ is total. □

Proposition 12. *(this is just a specification of Proposition 11). Assume that* $X \equiv L^2(T)$ *and* T *is bounded set in* R *or* R^2 *or* R^3. *Take* $\widetilde{X}$ *to be* $\mathcal{D}(T)$, *the space of* C^∞ *functions, with compact support in* T. *Define the Dirichlet kernels* $D_N(t, t')$ *as*

$$D_N(t, t') = \sum_{k=1}^{N} y_k(t) y_k(t'),$$

Where $\{y_k(t)\}$ *is ON in X.*

If one can prove that

$$t \in T \lim_{N \to \infty} \int_T D_N(t, t') \varphi(t') dt' \equiv \varphi(t), \tag{12.68}$$

or, said in another way, that

$$\lim_{N \to \infty} D_N(t, t') = \delta(t, t'),$$

then

$$\forall(t) = \sum_{k=0}^{+\infty} < \varphi, y_k > y_k,$$

the series being convergent in $X \equiv L^2$*, so that Proposition 11 applies and* $\{y_k\}$ *is complete in* X*.*

Proof. We want to prove that, setting

$$\varphi_N(t) = \int_T D_N(t,t')\varphi(t)dt = \\ = \sum_{k=1}^{N} < \varphi, y_k > y_k(t),$$

one has

$$\lim_{N\to\infty} \| \varphi - \varphi_N \|_X^2 = \lim_{N\to\infty} \int_T [\varphi(t) - \varphi_N(t)]^2 dt = 0. \tag{12.69}$$

However (12.69) dose not allow to pass directly to the limit under the integral, so making the proof elementary. Nevertheless we observe that D_N is just the kernel of the orthogonal projector P_N, on Span $\{y_k, k = 1 \dots N\}$, i.e. $\varphi_N = P_N\varphi$.

Accordingly

$$\| \varphi - \varphi_N \|^2 = < \varphi, \varphi > -2 < \varphi, \varphi_N > + < \varphi_N, \varphi_N > = \\ = \| \varphi \|^2 - \| \varphi_N \|^2 = \int_T \varphi^2 dt - \int_T \varphi_N^2(t)dt$$

so that to prove (12.69) we need only to show that

$$\lim_{N\to\infty} \int_T \varphi_N^2(t)dt = \int_T \varphi^2(t)dt \tag{12.70}$$

On the other hand $\{\varphi_N(t)\}$ is an L^2 convergent sequence so that we must have $\varphi_N(t) \underset{L^2}{\to} \bar{\varphi}(t)$, for some $\bar{\varphi}(t)$, as well as

$$\int_T \varphi_N^2(t)dt \to \int_T \bar{\varphi}^2(t)dt. \tag{12.71}$$

Moreover, we know that from $\{\varphi_N(t)\}$ we can always find a subsequence $\{N_j\}$ such that

$$\lim_{j\to\infty} \varphi_{N_j}(t) = \bar{\varphi}(t)$$

almost everywhere. But, on account of (12.68), we see that

$$\varphi(t) = \bar{\varphi}(t)$$

almost everywhere, so that (12.71) is (12.70) and (12.69) is proved. □

We remark explicitely here that Proposition 12 can be applied to the case that T is the unit circle and $\{y_k\}$ is the ordinary Fourier basis (see Example 8 below) and that T is the unit sphere S_1 and $\{y_k\}$ coincides with the sequence of spherical harmonics (see Theorem 3 in Section 13.4).

Example 8. The probably most classical example of a HS with an ON basis is that of $L^2([0,1])$, i.e. L^2-functions $f(t)$ defined on [0,1] (or equivalently functions $f(\vartheta), \vartheta = 2\pi t$, defined on the unit circle) with scalar product

$$< f, g >= \int_0^1 f(t)g(t)dt. \tag{12.72}$$

In this case one ON basis is the original Fourier basis, namely

$$Y_{2n}(t) = \sqrt{2-\delta_{n0}}\cos 2\pi nt\ ,\ Y_{2n+1}(t) = \sqrt{2}\sin 2\pi nt.$$
$$n = 0, 1, 2 \ldots$$

To verify the orthonormality relations

$$< Y_{2n}, Y_{2m} >= \delta_{nm},\ < Y_{2n+1}, Y_{2m+1} >= \delta_{nm},\ < Y_{2n}, Y_{2m+1} >= 0$$

is a simple integration exercise that everybody should make at least once.

That $\{Y_n\}$ is total, and then complete, in $L^2[0,1]$ derives from an application of Proposition 11, if we observe that the classical theory of Fourier series guarantees that

$$\widetilde{x}(t) = \sum_{k=0}^{+\infty}\left(\int_0^1 \widetilde{x}(\tau)Y_k(\tau)d\tau\right)Y_k(t),$$

with a uniform convergence of the series, when $\widetilde{x}(t)$ is continuous with its derivative on [0,1]. That the space $\widetilde{X}$, of functions continuous with their first derivative, is densely embedded in $L^2[0,1]$ is a fundamental lemma of the calculus of variations, that we don't prove here.

Accordingly we conclude that the Fourier series representation

$$x(t) = \left(\int_0^1 x(\tau)d\tau\right) + 2\sum_{k=1}^{+\infty}\left[\left(\int_0^1 \cos 2\pi k\tau \cdot x(\tau)d\tau\right)\cos 2\pi kt + \right.$$
$$\left. + \left(\int_0^1 \sin 2\pi k\tau \cdot x(\tau)d\tau\right)\sin 2\pi kt\right] \tag{12.73}$$

is valid $\forall x \in L^2([0,1])$, the series being convergent in L^2 norm.

Example 9. We consider the class of functions $f(t)$ such that $f \in L^2([0,1])$, $f'(t) \in L^2([0,1])$. We call $H^{1,2}([0,1])$ this space; from its definition it is clear that $H^{1,2}([0,1]) \subset L^2([0,1])$. Let us introduce in $H^{1,2}$ the scalar product

$$< f, g >_{H^{1,2}} = \int_0^1 \{f(t)g(t) + f'(t)g'(t)\}dt; \tag{12.74}$$

we want to prove that $H^{1,2}$ is a HS.

It is enough that we represent $f \in H^{1,2}$ as

$$f(t) = f_o + 2\sum_{k=1}^{+\infty}(a_k \cos 2\pi k t + b_k \sin 2\pi k t)$$

$$\|f\|^2_{L^2} = a_0^2 + 2\sum_{k=1}^{+\infty}\left(a_k^2 + b_k^2\right)$$

to find

$$f'(t) = 2\sum_{k=1}^{+\infty}(2\pi k)(-a_k \sin 2\pi k t + b_k \cos 2\pi k t)$$

$$\|f'\|^2_{L^2} = 2\sum_{k=1}^{+\infty}(2\pi k)^2\left(a_k^2 + b_k^2\right)$$

and so

$$\|f\|^2_{H^{1,2}} = a_0^2 + 2\sum_{k=1}^{+\infty}(1 + 4\pi^2 k^2)\left(a_k^2 + b_k^2\right). \tag{12.75}$$

As such, this case enters into that studied in Exercise 5 and therefore $H^{1,2}$ is a HS.

Example 10. For future use we consider a subspace of $H^{1,2}$ namely that of $f(t)$ such that $f(0) = f(1) = 0$. We denote by $H_0^{1,2}$ such space. Due to the above conditions we can define in $H_0^{1,2}$ an equivalent norm derived by the scalar product

$$< f, g >_{H_0^{1,2}} = \int_0^1 f'(t)g'(t)dt. \tag{12.76}$$

This is a true norm, in fact $\|f\|^2_{H_0^{1,2}} = 0 \rightarrow \int_0^1 f'(t)^2 dt = 0 \rightarrow f'(t) = 0$ almost everywhere in [0,1] an then $f(t) = c$ and $c = 0$ because $f(0) = f(1) = 0$.

We notice also that, since $f(t)$ can be continued on $[-1, 0]$ as an odd function, so that the extended $f'(t)$ is again square integrable over $[-1, 1]$, we can put, owing to Exercise 10,

$$f(t) = \sum_{k=1}^{+\infty} a_k \sin \pi k t \ , \ a_k = 2 \int_0^1 f(\tau) \sin \pi k \tau d\tau .$$

We have in this case

$$\| f \|^2_{H_0^{1,2}} = \frac{1}{2} \sum_{k=1}^{+\infty} (\pi k)^2 a_k^2 .$$

The corresponding scalar product with

$$g(t) = \sum_{k=1}^{+\infty} b_k \sin \pi k t ,$$

is given by

$$< f, g >_{H_0^{1,2}} = \frac{1}{2} \sum_{k=1}^{+\infty} (\pi k)^2 a_k b_k . \tag{12.77}$$

Example 11. We pick up again the spaces of polynomials of Example 2 and Exercise 1; namely we concentrate on $\mathcal{P}_n^3$ by endowing it with a suitable scalar product. First let us introduce the multi-index notation: we put, with $\alpha_1, \alpha_2, \alpha_3$ integer numbers,

$$\xi = (x, y, z); \ r = |\xi|; \ \alpha = (\alpha_1, \alpha_2, \alpha_3); \ \alpha_1, \alpha_2, \alpha_3 \geq 0,$$
$$|\alpha| = \alpha_1 + \alpha_2 + \alpha_3; \ \alpha! = \alpha_1! \alpha_2! \alpha_3! ; \ \xi^\alpha = x^{\alpha_1} y^{\alpha_2} z^{\alpha_3}$$

so that any polynomial $P_n(\xi)$ is naturally represented in terms of monomials $\xi^\alpha \in H^3_{|\alpha|}$ by

$$P_n(\xi) = \sum_{k=0}^{n} \sum_{|\alpha|=k} a_\alpha \xi^\alpha . \tag{12.78}$$

If we substitute in (12.78) the vector $\xi = (x, y, z)$ with $\partial_\xi = (\partial_x, \partial_y, \partial_z)$ we obtain the differential operator

$$P_n(\partial_\xi) = \sum_{k=0}^{n} \sum_{|\alpha|=k} a_\alpha \partial_\xi^\alpha . \tag{12.79}$$

We now define for every two polynomials $P_n(\xi), Q_n(\xi)$

$$< P_n, Q_n >= P_n(\partial_\xi) Q_n(\xi) \big|_{\xi=0} . \tag{12.80}$$

Due to the representations (12.78) and (12.79), to compute (12.80) we need the following formula

$$\begin{aligned} < \xi^{\alpha}, \xi^{\beta} > &= \left(\partial_x^{\alpha_1} x^{\beta_1}\right)\left(\partial_y^{\alpha_2} y^{\beta_2}\right)\left(\partial_z^{\alpha_3} z^{\beta_3}\right)\Big|_{xyz=0} \\ &= \delta_{\alpha_1\beta_1}\beta_1!\delta_{\alpha_2\beta_2}\beta_2!\delta_{\alpha_3\beta_3}\beta_3! \\ &= \delta_{\alpha\beta}\beta! \,. \end{aligned} \tag{12.81}$$

If we put

$$Q_n(\xi) = \sum_{k=0}^{n} \sum_{|\alpha|=k} b_\alpha \xi^\alpha,$$

taking the product (12.80) we get

$$< P_n, Q_n >= \sum_{k=0}^{n} \sum_{|\alpha|=k} a_\alpha \beta_\alpha \alpha! \tag{12.82}$$

All that shows that the decomposition

$$\mathcal{P}_n^3 = H_0^3 \oplus H_1^3 \oplus \ldots \oplus H_n^3 \tag{12.83}$$

is in fact an orthogonal decomposition such that

$$\text{Span}\left\{ \frac{1}{\sqrt{\alpha!}}\xi^\alpha \,;\, |\alpha| = k \right\} = H_k^3 \tag{12.84}$$

and $\frac{1}{\sqrt{\alpha!}}\xi^\alpha, |\alpha| = k$ is an ON basis in H_k^3.

12.4 Hilbert Spaces with Reproducing Kernel

Definition 20 (Kernel). Take a real function $K(t,t')$ defined on $(T \times T)$ and a separable HS, X, of functions on T; further assume that $\forall t$ fixed $K(t,t') \in X$ as a function of t'; then $\forall x(t') \in X$ and every fixed t you can form the scalar product

$$< K(t,t'), x(t') >= y(t); \tag{12.85}$$

this gives rise to a new function $y(t)$ and the set of these functions will, generally, be in some other linear space Y. If you assume that $\forall x \in X$, y, given by (12.85), is again in X, i.e. $Y \subseteq X$, then you say that $K(t,t')$ is a kernel on X.

In order to specify on which variable the scalar product is acting we shall adopt, depending on the case, the notation

$$< K(t,t'), x(t') >=< K(t,t'), x(t') >_{t'}=< K(t,\ \cdot), x(\ \cdot) > . \quad (12.86)$$

Example 12. The most classical example of kernels is probably that of an integral operator on $L^2(T)$; the kernel of the operator $K(t,t')$, when for instance $K \in L^2(T \times T)$, is then a kernel on L^2, with

$$< K(t,t'), x(t') >_{L^2}= \int_T K(t,t')x(t')dt' = y(t).$$

Definition 21 (Reproducing kernel). A kernel $K(t,t')$ on the HS, X is called a *reproducing kernel*, (RK), if

$$\forall x(t) \in X\ ,\ \forall t \in T;\ < K(t,t'), x(t') >\equiv x(t); \quad (12.87)$$

if X has a reproducing kernel, then we say that it is a *Reproducing Kernel Hilbert Space* (RKHS).

We observe that $< K(t,\cdot), \cdot >, X \to X$, is in fact the identity operator of X, so we can say that the RK, with the scalar product $< , >$, represents the identity of X.

Theorem 3. *Let X be a RKHS with RK, $K(t,t')$; then $K(t,t')$ is unique, symmetric, i.e. $K(t,t') = K(t',t)$, and given any ONC $\{y_n(t)\}$ in X we have*

$$\sum_{n=1}^{+\infty} y_n(t)y_n(t') = K(t,t'), \quad (12.88)$$

Proof. We prove the theorem in reverse order.

By definition we have

$$< K(t,t'), y_n(t') >= y_n(t), \quad (12.89)$$

i.e. $\{y_n(t)\}$ for fixed t, is the vector of Fourier coefficients of $K(t,t')$, so that, by Parseval's identity (12.59), the following series has to converge

$$\|K(t,t')\|^2_{t'} = \Sigma y_n^2(t) < +\infty, \quad (12.90)$$

and we are allowed to write

$$\sum_{n=1}^{+\infty} y_n(t)y_n(t') = K(t,t'), \quad (12.91)$$

the series being convergent in X for $\forall t$. Then we observe that $K(t,t')$ is indeed symmetric.

Since (12.91) holds for a specific, fixed $\{y_n(t)\}$ whatever is the kernel $K(t,t')$ with a reproducing property, $K(t,t')$ must be unique. Since $K(t,t')$ is unique, (12.91) holds, whatever is the ONC $\{y_n(t)\}$. □

Corollary 2. *Let $K(t,t')$ be the RK of a HS, X, then it is a positive definite function and in particular*

$$\sum_{i,j=1}^{N} \lambda_i \lambda_j K(t_i,t_j) = \left\| \sum_{i=1}^{N} \lambda_i K(t_i,\cdot) \right\|^2 \tag{12.92}$$

Proof. That the quadratic form in (12.92) is positive, is straightforward already because of the representation (12.88) of $K(t,t')$. As for the second implication of (12.92), we can just compute

$$\left\| \sum_{i=1}^{N} \lambda_i K(t_i,\cdot) \right\|^2 = < \sum_{i=1}^{N} \lambda_i K(t_i,\cdot), \sum_{j=1}^{N} \lambda_j K(t_j,\cdot) >$$

$$= \sum_{i,j=1}^{N} \lambda_i \lambda_j < K(t_i,\cdot), K(t_j,\cdot) > = \sum_{i,j=1}^{N} \lambda_i \lambda_j K(t_i,t_j).$$

□

Proposition 13. *Let X be a RKHS and $K(t,t')$ be such that*

$$\sup_{t\in T} K(t,t) \le c < +\infty \tag{12.93}$$

$$\lim_{t,t'\to\tau} K(t,t') = K(\tau,\tau)\ ,\ \forall \tau \in T, \tag{12.94}$$

then

(a) Every $x(t) \in X$ is bounded,
(b) Every $x(t) \in X$ is continuous,
(c) The evaluation functional (cf. Example 3)

$$ev_{\bar{t}}\{x(t)\} = x(\bar{t})$$

is continuous on X at any point $\bar{t} \in T$.

Proof. (a) Use Schwarz inequality (see Exercise 11 and (12.93)) in

$$x(t) = < K(t,\cdot), x(\cdot) >$$

(b) Use Schwarz inequality (see Exercise 11 and (12.94)) in

$$|x(t+h) - x(t)|^2 = | < K(t+h,\cdot) - K(t,\cdot), x(\cdot) > |^2$$

(c) Already from (a) one has

$$|ev_{\bar{t}}\{x\}| \leq \sqrt{K(\bar{t},\bar{t})} \cdot \|x\| \ . \tag{12.95}$$

□

Proposition 14. *This Proposition is in a sense the inverse of Proposition (13). Let X be a HS of functions bounded on T; then:*

(a) If $ev_{\bar{t}}\{x\}$ is a bounded functional $\forall \bar{t} \in T$,

$$|ev_{\bar{t}}\{x\}| = |x(\bar{t})| \leq c\|x\| \ , \tag{12.96}$$

then X is a RKHS

(b) If X is continuously embedded in $C(T)$ (i.e. every $x(t)$ is continuous and $(x_n \to 0$ in $X) \Rightarrow (x_n \to 0$ in $C(T))$),
then X is a RKHS.

Proof. (a) By using Riesz Theorem 2, $\forall \bar{t} \in T$, there is an element $K(\bar{t},\cdot) \in X$ such that

$$\mathrm{ev}_{\bar{t}}\{x\} = x(\bar{t}) =< K(\bar{t},\cdot), x(\cdot) >;$$

$K(t,t')$ is then the RK,

(b) We note that in this case we must have

$$\|x\|_{C(T)} = \sup_{t\in T} |x(t)| \leq c\|x\|_X, \ (c < +\infty) \tag{12.97}$$

for otherwise there is x_n such that $\|x_n\|_C / \|x_n\|_X \to +\infty$, i.e. putting

$$\xi_n = \frac{x_n}{\|x_n\|_C}$$

we find that $\|\xi_n\|_X \to 0$ while $\|\xi_n\|_C = 1$ contrary to the hypothesis in (b). But (12.97) implies that $\mathrm{ev}_{\bar{t}}\{x\}$ is continuous $\forall \bar{t} \in T$.

□

Remark 7. Note that (b) in Proposition 14 is more stringent than (a), because (12.97) implies (12.96). This is important when the functions in X might not be bounded, as it can happen if they are continuous on a set T which is open. The case of functions harmonic in an open set and square integrable on the boundary, has in fact this characteristic. Yet, even if T is open, (12.96) is sufficient to claim the existence of a RK in X. We also note explicitly that when (b) is satisfied, i.e. (12.97) is satisfied, any sequence $x_n \to x$ in X is also such that $x_n(t) \to x(t)$ uniformly on T.

Example 13. Any N-dimensional HS, X, whose elements are function, has a RK, because it has an ONC sequence $\{y_n\}, n = 1, 2 \dots N$, and the sum

$$K(t, t') = \sum_{n=1}^{N} y_n(t) y_n(t')$$

has no convergence problems.

For instance $\mathcal{P}_N^3$ of Example 11 has the reproducing kernel $K(\boldsymbol{\xi}, \boldsymbol{\eta})$ (note that $\boldsymbol{\xi}, \boldsymbol{\eta}$ are 3D vectors)

$$K(\xi, \eta) = \sum_{n=0}^{N} \frac{1}{n!} (\boldsymbol{\xi}^t \boldsymbol{\eta})^n \tag{12.98}$$

In fact, remember that the following multinomial formula holds

$$(\boldsymbol{\xi}^t \boldsymbol{\eta})^n = \sum_{|\alpha|=n} \frac{n!}{\alpha!} \boldsymbol{\xi}^\alpha \boldsymbol{\eta}^\alpha,$$

so that

$$K_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{1}{n!} (\boldsymbol{\xi}^t \boldsymbol{\eta})^n = \sum_{|\alpha|=n} \frac{\boldsymbol{\xi}^\alpha}{\sqrt{\alpha!}} \frac{\boldsymbol{\eta}^\alpha}{\sqrt{\alpha!}}. \tag{12.99}$$

Remember that (cf. (12.84)) $\left\{ \dfrac{\boldsymbol{\xi}^\alpha}{\sqrt{\alpha!}}, |\alpha| = n \right\}$ is ONC in H_n^3, so that (12.99) is the RK of H_n^3. Due to the orthogonal decomposition (12.83), it is obvious that, under the scalar product (12.81),

$$< K_n(\boldsymbol{\xi}, \cdot), K_m(\boldsymbol{\xi}', \cdot) >= \delta_{nm} K_n(\xi, \xi') \tag{12.100}$$

and therefore, for $P_N(\boldsymbol{\eta}) = \sum_{k=0}^{N} \sum_{|\alpha|=k} a_\alpha \boldsymbol{\eta}^\alpha$ and $K(\boldsymbol{\xi}, \boldsymbol{\eta})$ given by (12.98)

$$< K(\boldsymbol{\xi}, \boldsymbol{\eta}), P_N(\boldsymbol{\eta}) >= \sum_{n=0}^{N} \sum_{k=0}^{N} \sum_{|\alpha|=k} a_\alpha < K_n(\boldsymbol{\xi}, \boldsymbol{\eta}), \boldsymbol{\eta}^\alpha >$$

$$= \sum_{n=0}^{N} \sum_{|\alpha|=n} a_\alpha < K_n(\boldsymbol{\xi}, \boldsymbol{\eta}), \boldsymbol{\eta}^\alpha >= \sum_{n=0}^{N} \sum_{|\alpha|=n} a_\alpha \boldsymbol{\xi}^\alpha = P_N(\boldsymbol{\xi}).$$

Example 14. Take $X = H^{1,2}([0,1])$ as in Example 9: put

$$y_0(t) = 1,\ y_{2n}(t) = \sqrt{\frac{2}{1+4n^2\pi^2}}\cos 2\pi nt,$$
$$y_{2n+1}(t) = \sqrt{\frac{2}{1+4n^2\pi^2}}\sin 2\pi nt.$$

First verify directly that

$$< y_k(t), y_\ell(t) >_{H^{1,2}} = \int_0^1 [y_k(t)y_\ell(t) + y_k'(t)y_\ell'(t)]dt = \delta_{k\ell}.$$

Observe then that the series

$$K(t,t') = 1 + \sum_{n=1}^{+\infty} \frac{\cos 2\pi nt \cos 2\pi nt' + \sin 2\pi nt \sin 2\pi nt'}{1+4\pi^2 n^2}$$
$$= 1 + \sum_{n=1}^{+\infty} \frac{\cos 2\pi n(t-t')}{1+4\pi^2 n^2} < +\infty$$

is in fact convergent $\forall t, t' \in [0,1]$ and

$$K(t,t) = 1 + 2\sum_{n=1}^{+\infty} \frac{1}{1+4n^2\pi^2} < +\infty;$$

then $K(t,t')$ is the reproducing kernel of $H^{1,2}([0,1])$.

Example 15. The HS, $X = L^2([0,1])$ is not a RKHS. It is enough to observe that already the condition (12.95) is not satisfied and that in fact the evaluation functional cannot be defined in L^2 because two functions which differ only for a value at a point $\overline{t}$ are one and the same element of L^2. As a further check of our statement one can observe that

$$y_0(t) = 1\ ,\ y_{2n}(t) = \sqrt{2}\cos 2\pi nt\ ,\ y_{2n+1}(t) = \sqrt{2}\sin 2\pi nt$$

is an ONC system in L^2, but

$$\sum_{n=0}^{+\infty} y_n(t)^2 = 1 + 2\sum_{n=1}^{+\infty}(\cos^2 2\pi nt + \sin^2 2\pi nt) = +\infty$$

contrary to condition (12.90) that any RK has to satisfy.

Proposition 15. *Let X be a RKHS with a continuous kernel $K(t,t')$, so that all $x(t) \in X$ are continuous functions too; take any sequence $\{t_i\}$ which is dense in T, then the set $\{K(t_i,\cdot)\}$ is total in X.*

Proof. If $x \in X$ is such that

$$< K(t_i, \cdot), x(\cdot) >= x(t_i) = 0 \ \forall i \ ,$$

then $x(t) \equiv 0$ on T, because $x(t)$ is continuous. □

Proposition 16. *A RKHS, X is functionally completely identified by its RK; in other words given a symmetric positive definite $K(t, t')$ we can build a RKHS, X, of which K is the RK.*

Proof. In fact, due to Proposition 15 the linear space

$$V \equiv \text{Span}\{K(t_i, \cdot)\}$$

is densely embedded in X, $\overline{V} \equiv X$. Furthermore the scalar product of two elements of V can always be computed without knowing explicitly its form, because

$$< \sum_{i=1}^{N} \lambda_i K(t_i, \cdot), \ \sum_{j=1}^{M} \mu_j K(t_j, \cdot) > \tag{12.101}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} \lambda_i \mu_j K(t_i, t_j) \ .$$

Then we can also compute norms and evaluate limits of Cauchy sequences, which are general elements of X, i.e. we build X as the closure of V under the norm implied by (12.101). □

Remark 8. Associated with a random function $X(t, \omega), t \in T, \omega \in \Omega$ in $\mathcal{L}_0^2(\Omega)$ (cf. Example 7) there is the so-called covariance function of $X(t, \omega)$, i.e.

$$C(t, t') = E\{X(t, \omega) X(t', \omega)\} \ ;$$

as it is well-known, this is a symmetric and positive definite function, which is continuous in $T \times T$ if it is continuous on the diagonal $t' = t$.

Therefore we can define a RKHS, that we can denote H_C, which has C as RK; such a space is known in stochastic literature (Kallianpur 1980) as the Cameron-Martin space of the random function.

It is interesting to note that it is exactly in this space that one obtains the "best" prediction of $X(t, \cdot)$ given some known sample values $X(t_i), i = 1, 2 \ldots N$, as discussed in Part I, Sect. 5.4.

Proposition 17. *Let $L(\ )$ be a bounded linear functional on a RKHS, X, with RK, $K(t, t')$, and let y_L be its Riesz representer; then*

$$y_L(t') = L_t[K(t, t')]; \tag{12.102}$$

furthermore

$$L_t[< K(t, t'), x(t') >] =< L_t[K(t, t')], x(t') > \ . \tag{12.103}$$

Proof. We just observe that

$$L_t[K(t,t')] =< y_L(t), K(t,t') >=< K(t',t), y_L(t) >= y_L(t').$$

Now (12.102) means also that scalar product and linear functional can be exchanged, in fact

$$\begin{aligned} L_t[x(t)] &= L_t < K(t,t'), x(t') > \\ &= < y_L(t'), x(t') >=< L_t K(t,t'), x(t') > . \end{aligned}$$

□

Definition 22 (Notation for Riesz representers). We define the following notation

$$L_t[K(t,t')] = K(L,t') \tag{12.104}$$

$$M_{t'}\{L_t[K(t,t')]\} = L_t\{M_{t'}[K(t,t')]\} = K(L,M) \tag{12.105}$$

$$\mathbf{L}_t = \{L_{it},\ i = 1,2\ldots n\}\ ,\ \mathbf{M}_{t'} \equiv \{M_{it'}\ ,\ i = 1,2\ldots m\}$$

$$\mathbf{L}_t\{\mathbf{M}^t_{t'}[K(t,t')]\} = K(\mathbf{L},\mathbf{M}^t)\ ; \tag{12.106}$$

Remark 9. Note that $K(\mathbf{L},\mathbf{M}^t)$ is an $n \times m$ matrix and that if you put $\mathbf{M} = \mathbf{L}$ then it becomes a symmetric definite positive matrix. In fact

$$K^t(\mathbf{L},\mathbf{M}^t) = K(\mathbf{M},\mathbf{L}^t)$$

and if you put $\mathbf{M} = \mathbf{L}$ you see that $K(\mathbf{L},\mathbf{L}^t)$ is symmetric. Moreover (see also the Corollary of the Theorem 3)

$$\begin{aligned} &\boldsymbol{\lambda}^t K(\mathbf{L},\mathbf{L}^t)\boldsymbol{\lambda} = (\boldsymbol{\lambda}^t\mathbf{L}_t)(\boldsymbol{\lambda}^t\mathbf{L}_{t'})K(t,t') \\ &= (\boldsymbol{\lambda}^t\mathbf{L}_t)(\boldsymbol{\lambda}^t\mathbf{L}_{t'}) < K(t,\cdot), K(t',\cdot) > \\ &=< \boldsymbol{\lambda}^t K(\mathbf{L},\cdot)\ ,\ \boldsymbol{\lambda}^t K(\mathbf{L},\cdot) >= \|\boldsymbol{\lambda}^t K(\mathbf{L},\cdot)\|^2\ . \end{aligned} \tag{12.107}$$

In the same way you can prove that in general

$$\boldsymbol{\lambda}^t K(\mathbf{L},\mathbf{M})\boldsymbol{\mu} =< \boldsymbol{\lambda}^t K(\mathbf{L},\cdot)\ ,\ \boldsymbol{\mu}^t K(\mathbf{M},\cdot) >\ . \tag{12.108}$$

We are now in a position to state the main result of this chapter from the point of view of the approximation theory which, in the context of geodesy, is also known as *deterministic collocation theory*. Conceptually this result is just re-stating Theorem 1 and its Corollary in the context of RKHS theory; yet its physical interpretation is so important that it is worth expressing it in a separate theorem.

Definition 23 (Formulation of the collocation problem). Assume the following to hold

(a) We have an unknown function (field) $x(t)$ of which we know a-priori that it is a member of RKHS, X with known RK, $K(t,t')$,

(b) We have the results of N observations performed on the field, which can be expressed in terms of linear functionals of X,

$$L_i(x) = c_i \quad i = 1, 2 \ldots N$$

or

$$\mathbf{L}(x) = \mathbf{c}; \tag{12.109}$$

$\{c_i\}$ are assumed to be known without error and for this reason this problem is sometimes called the problem of "exact" collocation,

(c) We shall assume that L_i are continuous functionals on X as otherwise it makes no sense to perform the corresponding measurement. This is in reality a constraint on X, i.e. on $K(t,t')$, since we must have $\forall i,\ K(L_i,\cdot) \in X$, i.e. $K(L_i, L_i) < +\infty$, to the effect that (12.109) can be written as

$$< K(\mathbf{L},\cdot), x(\cdot) >= \mathbf{c}. \tag{12.110}$$

Given the hypotheses (a)–(c) we can formulate a direct and a dual "optimal" approximation problem.

(d) *Direct formulation or smoothing*: we consider (12.110) as an underdetermined equation with unknown x in X and known term $\mathbf{c}$ in R^N, then we may expect it to have an infinite number of solutions and among them we look for the "smoothest", i.e. for $\widehat{x}$ such that

$$\begin{cases} \widehat{x} = Arg \min \|x\| \\ < K(\mathbf{L},\cdot), \widehat{x} >= \mathbf{c}; \end{cases} \tag{12.111}$$

furthermore we want to know the magnitude of the error $x - \widehat{x}$, i.e.

$$\mathcal{E}_{\text{tot}}(x) = \|x - \widehat{x}\|, \tag{12.112}$$

(e) *Dual formulation*: we give an arbitrary bounded functional on X, $L(\cdot)$, represented by $y_L \in X$; we want to approximate $< y_L, x >$ from what we know, i.e. the vector $\mathbf{c}$ of (12.110). We define

$$V_{\mathbf{L}} = \text{Span}\{K(\mathbf{L},\cdot)\}$$

and we search for a $\widehat{y}_L \in V_{\mathbf{L}}$ so that the relative error

$$(\widehat{y}_L \in V_{\mathbf{L}}), \quad \mathcal{E}(L, x, \widehat{y}_L) = \frac{| < y_L, x > - < \widehat{y}_L, x > |}{\|x\|} \tag{12.113}$$

becomes uniformly minimum with respect to x. Namely we set up a Minimax criterion as follows: first we define the uniform relative error as (note that when x sweeps X, $\frac{x}{\|x\|}$ sweeps the unit sphere)

$$\begin{aligned} \mathcal{E}_r(L, \widehat{y}_L) &= \sup_{x \in X} \frac{| < y_L - \widehat{y}_L, x > |}{\|x\|} \\ &= \|y_L - \widehat{y}_L\| \end{aligned} \tag{12.114}$$

and then we look for $\widehat{y}_L \in V_{\mathbf{L}}$ which minimizes this error, i.e.

$$\mathcal{E}_r(L) = \min_{\widehat{y}_L \in V_{\mathbf{L}}} \|y_L - \widehat{y}_L\|; \tag{12.115}$$

in particular we want to find both $\widehat{y}_L$ and $\mathcal{E}_r(L)$.

Notice that in (12.112) the index tot stems for total, while in (12.114) and (12.115) the index r stems for relative, because $\mathcal{E}_{\text{tot}}(x)$ does depend on the entire vector x while $\mathcal{E}_r(L)$ expresses an error independently of x because of the previous extremization (12.114).

Theorem 4. *Given the hypotheses (a)–(c) under Definition 23, the following holds:*

(a) Let us call $\widehat{x}$ the orthogonal projection of x on $V_{\mathbf{L}}$, $\widehat{x} = P_{\mathbf{L}}x$, given by (cf. (12.41) and (12.43))

$$\begin{aligned} \widehat{x} &= < x, K(\cdot, \mathbf{L}^t) > K(\mathbf{L}, \mathbf{L}^t)^{-1} K(\mathbf{L}, \cdot) \\ &= \mathbf{c}^t K(\mathbf{L}, \mathbf{L}^t)^{-1} K(\mathbf{L}, \cdot)\ , \end{aligned} \tag{12.116}$$

where (12.110) has been taken into account; (12.116) means that given the data $\mathbf{c}$, $\widehat{x}$ is fixed; then the solution of the smoothing collocation problem (12.111) is exactly $\widehat{x} = P_{\mathbf{L}}x$ and (cf. (12.44))

$$\begin{aligned} \mathcal{E}^2_{\text{tot}}(x) &= \|x\|^2 - \|\widehat{x}\|^2 \\ &= \|x\|^2 - \mathbf{c}^t K(\mathbf{L}, \mathbf{L}^t)^{-1} \mathbf{c} \end{aligned} \tag{12.117}$$

(b) Let us call $\widehat{y}_L = P_{\mathbf{L}} y_L$ the orthogonal projection of y_L on $V_{\mathbf{L}}$, i.e.

$$\begin{aligned} \widehat{y}_L &= < y_L, K(\cdot, \mathbf{L}^t) > K(\mathbf{L}, \mathbf{L}^t)^{-1} K(\mathbf{L}, \cdot) \\ &= K(L, \mathbf{L}^t) K(\mathbf{L}, \mathbf{L}^t)^{-1} K(\mathbf{L}, \cdot)\ ; \end{aligned} \tag{12.118}$$

then the solution of the dual collocation problem (12.115) is given by (12.118), so that one can write

$$\begin{aligned}\widehat{< y_L, x >} &= < \widehat{y}_L, x > \\ &= K(L, \mathbf{L}^t) K(\mathbf{L}, \mathbf{L}^t)^{-1} < K(\mathbf{L}, \cdot), x(\cdot) >, \\ &= K(L, \mathbf{L}^t) K(\mathbf{L}, \mathbf{L})^{-1} \mathbf{c}\end{aligned} \tag{12.119}$$

and the relative error is

$$\mathcal{E}_r^2 = K(L, L) - K(L, \mathbf{L}^t) K(\mathbf{L}, \mathbf{L}^t)^{-1} K(\mathbf{L}, L) \,, \tag{12.120}$$

(c) *The two solutions are equivalent in the sense that, choosing $L = ev_{\bar{t}}$ in (12.119), we get exactly*

$$\widehat{< ev_{\bar{t}}, x >} = \widehat{x}(\bar{t}) \,. \tag{12.121}$$

Proof. (a) Since $\widehat{x}$ is fixed by (12.116), one can put

$$x = \widehat{x} + h, \; h \perp V_{\mathbf{L}}$$

so that

$$\|x\|^2 = \|\widehat{x}\|^2 + \|h\|^2; \tag{12.122}$$

the minimum of (12.122) is attained at $h = 0$. Formula (12.117) is just a specification of (12.44),

(b) This is a straightforward application of the Corollary to Theorem 1,

(c) Put $L = \mathrm{ev}_{\bar{t}}$ in (12.119) and, noting that in this case $K(L, \mathbf{L}^t) = K(\bar{t}, \mathbf{L}^t)$, compare it with (12.116).

□

It should be no wonder that one and the same $\widehat{x}$ is simultaneously solution of two optimization problems, namely

$$\widehat{x} = \arg\min_{\eta \in V_{\mathbf{L}}} \|x - \eta\|, \; \widehat{x} = \arg \min_{\xi = \widehat{x} + h, h \perp V_{\mathbf{L}}} \|\xi\| \tag{12.123}$$

as it is clarified by Fig. 12.1 and readers can verify by themselves.

Note that (12.117) can never be used in practice, unless one has a-priori a guess on the value of $\|x\|^2$.

Example 16. Take $X = H_0^{1,2}$. Assume that the exact observation $x(t_0)$, $(0 < t_0 < 1)$, is taken and we want to estimate $L(x) = \widehat{x}(t)$ for all $t \in [0, 1]$, together with $\mathcal{E}_r^2(t)$.

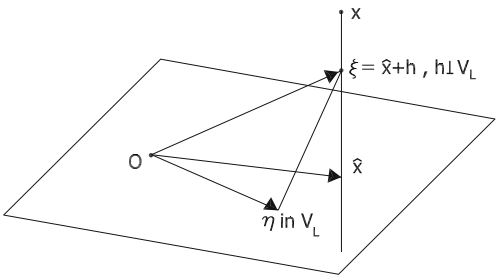


Fig. 12.1 Equivalence of the two minimum principles (12.123)

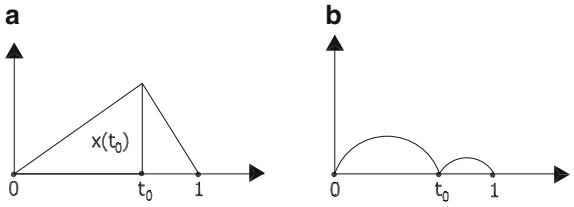


Fig. 12.2 The plot of (**a**) $\widehat{x}(t)$, (**b**) $\mathcal{E}_r^2(t)$

Recalling that the RK, $K(t,t')$, is given in Exercise 12, we can put (cf. (12.119))

$$\widehat{x}(t) = K(t,t_0)K^{-1}(t_0,t_0)x(t_0) = \begin{cases} \frac{t}{t_0}x(t_0) & t \le t_0 \\ \frac{1-t}{1-t_0}x(t_0) & t \ge t_0 \end{cases};$$

the relative error is

$$\mathcal{E}_r^2(t) = K(t,t) - \frac{K^2(t,t_0)}{K(t_0,t_0)} = \begin{cases} \frac{t(t_0-t)}{t_0} & t \le t_0 \\ \frac{(1-t)(t-t_0)}{1-t_0} & t \ge t_0 \end{cases}$$

The plot of these two functions shows that the interpolation is exact, so that the error goes to zero at t_0 as well as at 0 and 1 where we must have $x(t) \equiv 0$ (Fig. 12.2).

The last case we are going to treat in this chapter is still deterministic, to what refers to our unknown $x \in X$, but it allows to introduce a random noise in the observations, in the sense that now our observation equations become

$$\mathbf{c} = \mathbf{L}(x) + \boldsymbol{\nu} =< K(\mathbf{L},\cdot), x(\cdot) > + \boldsymbol{\nu} \tag{12.124}$$

where $\boldsymbol{\nu}$ is a random vector in R^N, with

$$E\{\boldsymbol{\nu}\} = 0,\ E\{\boldsymbol{\nu}\boldsymbol{\nu}^t\} = C_{\boldsymbol{\nu}}. \tag{12.125}$$

It is clear in this context that it makes no sense to impose a pure smoothing condition as in (12.111) where, through its second relation, we impose to $\widehat{x}$ to satisfy exactly the observation equations. In fact, imagine that $x = 0$ in (12.124); then $\mathbf{c} = \boldsymbol{\nu}$, and if we force exactly $\mathbf{L}(\widehat{x}) = \mathbf{c} = \boldsymbol{\nu}$ we get some non-smooth solution, where we would like that an optimization concept would help us in smoothing such a solution. Thus we are led to find a criterion that compromises between going close to the observations, i.e. keeping $[\mathbf{c} - < K(\mathbf{L}, \cdot), \widehat{x} >]^t C_{\boldsymbol{\nu}}^{-1} [\mathbf{c} - < K(\mathbf{L}, \cdot), \widehat{x} >]$ small, and smoothing, i.e. keeping $\|\widehat{x}\|$ small. This is the meaning of the next definition.

Definition 24 (Tikhonov or hybrid norm optimization). Let us put for the sake of brevity $\mathbf{y} = K(\mathbf{L}, \cdot)$ and $G = K(\mathbf{L}, \mathbf{L}^t) =< \mathbf{y}, \mathbf{y}^t >$, as in (12.41).

We say that $\widehat{x}$ is an α-Tikhonov smoother if it satisfies the optimization criterion

$$\begin{cases} \widehat{x}_\alpha = \arg\min_{\xi \in X} Q(\xi, \alpha) \\ Q(\xi, \alpha) = [\mathbf{c} - < \mathbf{y}, \xi >]^t C_\nu^{-1} [\mathbf{c} - < \mathbf{y}, \xi >] + \alpha \|\xi\|^2; \end{cases} \tag{12.126}$$

since we expect that $\widehat{x}$ will depend on the random variable $\boldsymbol{\nu}$, we shall take as an index of the goodness of the approximation

$$\mathcal{E}^2(x, \alpha) = E_{\boldsymbol{\nu}} \{ \|x - \widehat{x}_\alpha\|^2 \} . \tag{12.127}$$

Theorem 5. *The α-Tikhonov smoother is given by*

$$\begin{aligned} \widehat{x}_\alpha &= \mathbf{y}^t (G + \alpha C_{\boldsymbol{\nu}})^{-1} \mathbf{c} \\ &= K(\cdot, \mathbf{L}) [K(\mathbf{L}, \mathbf{L}^t) + \alpha C_{\boldsymbol{\nu}}]^{-1} \mathbf{c}; \end{aligned} \tag{12.128}$$

for this optimal estimator we have

$$\begin{aligned} \mathcal{E}^2(x, \alpha) = \|x\|^2 &+ < x, \mathbf{y}^t > (G + \alpha C_{\boldsymbol{\nu}})^{-1} G (G + \alpha G_{\boldsymbol{\nu}})^{-1} < \mathbf{y}, x > + \\ &-2 < x, \mathbf{y}^t > (G + \alpha C_{\boldsymbol{\nu}})^{-1} < \mathbf{y}, x > \\ &+ Tr (G + \alpha C_{\boldsymbol{\nu}})^{-1} G (G + \alpha C_{\boldsymbol{\nu}})^{-1} C_{\boldsymbol{\nu}} \end{aligned} \tag{12.129}$$

Proof. First of all we note that (12.126) implies that $\forall \xi \in V_{\mathbf{L}}, \forall h \in V_{\mathbf{L}}^{\perp}$

$$Q(\xi + h, \alpha) = Q(\xi, \alpha) + \alpha \|h\|^2. \tag{12.130}$$

Then assume $\widetilde{x}_\alpha$ to be solution of (12.126) and $\widehat{x}_\alpha$ its orthogonal projection on $V_{\mathbf{L}}$; since

$$Q(\widetilde{x}_\alpha, \alpha) = Q(\widehat{x}_\alpha, \alpha) + \alpha \|\widetilde{x}_\alpha - \widehat{x}_\alpha\|^2$$

we see that $\widetilde{x}_\alpha$ is a minimum point only if $\widetilde{x}_\alpha = \widehat{x}_\alpha$.

Therefore any solution of (12.126) has to lie in $V_{\mathbf{L}}$. So we can put $\xi = \boldsymbol{\lambda}^t \mathbf{y}$ and look for the minimum with respect to $\boldsymbol{\lambda}$ of

$$Q = [\mathbf{c} - G\boldsymbol{\lambda}]^t C_{\boldsymbol{\nu}}^{-1} [\mathbf{c} - G\boldsymbol{\lambda}] + \alpha \boldsymbol{\lambda}^t G \boldsymbol{\lambda}. \tag{12.131}$$

To minimize (12.130) is a standard problem in l.s. theory; its result is given by the solution of the normal equation (remember that $G^t = G$)

$$G C_{\boldsymbol{\nu}}^{-1} G\boldsymbol{\lambda} + \alpha G\boldsymbol{\lambda} = G C_{\nu}^{-1} \mathbf{c}.$$

Now, recalling that $\mathbf{y}$ is an independent basis of $V_{\mathbf{L}}$ so that G^{-1} exists, we multiply this equation by $C_\nu G^{-1}$ and we find

$$(G + \alpha C_{\boldsymbol{\nu}})\boldsymbol{\lambda} \equiv G_\alpha \boldsymbol{\lambda} = \mathbf{c}$$

where we have set $G_\alpha = G + \alpha C_{\boldsymbol{\nu}}$. This gives

$$\boldsymbol{\lambda} = G_\alpha^{-1} \mathbf{c} \Rightarrow \widehat{x}_\alpha = \mathbf{y}^t G_\alpha^{-1} \mathbf{c}, \tag{12.132}$$

as it was to be proved.

Now we have

$$\|x - \widehat{x}_\alpha\|^2 = \|x\|^2 - 2 < x, \mathbf{y}^t > G_\alpha^{-1} \mathbf{c} + \mathbf{c}^t G_\alpha^{-1} G G_\alpha^{-1} \mathbf{c} \tag{12.133}$$

and recalling that

$$E\{\mathbf{c}\} = < \mathbf{y}, x >, C_{\mathbf{c}} = C_\nu$$

we derive from (12.133),

$$\begin{aligned} \mathcal{E}^2(x, \alpha) = E\{\|x - \widehat{x}_\alpha\|^2\} = \|x\|^2 - 2 < x, \mathbf{y}^t > G_\alpha^{-1} < \mathbf{y}, x > \\ + < x, \mathbf{y}^t > G_\alpha^{-1} G G_\alpha^{-1} < \mathbf{y}, x > + Tr G_\alpha^{-1} G G_\alpha^{-1} C_\nu \end{aligned} \tag{12.134}$$

which proves (12.129). □

Remark 10. If we let $\alpha \to \infty$ in (12.128) and (12.129) we immediately see that $G_\alpha \to 0$ and

$$\widehat{x}_\alpha \to 0\,, \ \mathcal{E}^2(x, \alpha) \to \|x\|^2$$

this is because for large α the smoothing term in $Q(\xi, \alpha)$ prevails and indeed $\xi = 0$ is the "smoothest" solution, leaving the whole x as an "error". If, on the contrary, we take $\alpha = 0$ we have no smoothing at all, so that

$$\widehat{x}_0 = \mathbf{y}^t G^{-1} \mathbf{c};\ \mathcal{E}^2(x, 0) = \|x\|^2 - < x, \mathbf{y}^t > G^{-1} < \mathbf{y}, x > + Tr G^{-1} C_\nu;$$

in this case we reproduce exactly the observations, because $< \mathbf{y}, \widehat{x}_0 >= \mathbf{c}$ and the error is in fact the total error (12.117),

$$\mathcal{E}_t^2 = \|x\|^2 - < x, \mathbf{y}^t > G^{-1} < \mathbf{y}, x >,$$

with the addition of the effect of the noise represented by $TrG^{-1}C_\nu$. Between these two extreme behaviours one would like to find an optimal value of α which minimizes $\mathcal{E}^2(x, \alpha)$.

One can observe that formula (12.134) is of no use for this purpose, if we don't have any prior knowledge on x. Nevertheless, if we use $\mathbf{c}$ as a guess for the value of $< \mathbf{y}, x >$, (12.134) can be utilized to try to optimize $\mathcal{E}^2$ with respect to α, namely we can minimize the approximate expression

$$\mathcal{E}^2 \sim \|x\|^2 - 2\mathbf{c}^t G_\alpha^{-1}\mathbf{c} + \mathbf{c}^t G_\alpha^{-1} G G_\alpha^{-1}\mathbf{c} + TrG_\alpha^{-1} G G_\alpha^{-1} C_\nu; \qquad (12.135)$$

indeed we don't need to know $\|x\|^2$, which is constant, to look for the minimum of (12.135).

Remark 11. We expect naturally that $\mathcal{E}^2$ in (12.134) is always larger than $\mathcal{E}_t^2$ given by (12.117), because in the case of Theorem 5 we have the further error introduced by the noise $\boldsymbol{\nu}$. As a matter of fact, using (12.117) we see that

$$\begin{aligned}
\mathcal{E}^2 &= \mathcal{E}_{\text{tot}}^2 + < x, \mathbf{y}^t > G^{-1} < \mathbf{y}, x^t > -2 < x, \mathbf{y}^t > G_\alpha^{-1} < \mathbf{y}, x > \\
&\quad + < x, \mathbf{y}^t > G_\alpha^{-1} G G_\alpha^{-1} < \mathbf{y}, x > + TrG_\alpha^{-1} G G_\alpha^{-1} C_{\boldsymbol{\nu}} \\
&= \mathcal{E}_{\text{tot}}^2 + < x, \mathbf{y}^t > [G^{-1} - G_\alpha^{-1}] G [G^{-1} - G_\alpha^{-1}] < \mathbf{y}, x > \\
&\quad + TrG_\alpha^{-1} G G_\alpha^{-1} C_\nu = \mathcal{E}_t^2 + \alpha^2 < x, \mathbf{y}^t > G_\alpha^{-1} C_\nu G^{-1} C_\nu G_\alpha^{-1} < \mathbf{y}, x > \\
&\quad + TrG_\alpha^{-1} G G_\alpha^{-1} C_\nu, \qquad (12.136)
\end{aligned}$$

where the identity

$$G^{-1} - G_\alpha^{-1} = G_\alpha^{-1}(G_\alpha - G)G^{-1}$$

has been used. Since the second and third term in the right hand side of (12.136) are positive, we have indeed

$$\mathcal{E}^2 \geq \mathcal{E}_{\text{tot}}^2.$$

We note also that the use of the approximation $< \mathbf{y}, x > \sim \mathbf{c}$ in (12.136) gives the comfortable formula

$$\mathcal{E}^2(x, \alpha) \sim \mathcal{E}_{\text{tot}}^2 + \alpha^2 \mathbf{c}^t G_\alpha^{-1} C_\nu G^{-1} C_\nu G_\alpha^{-1}\mathbf{c} + TrG_\alpha^{-1} G G_\alpha^{-1} \qquad (12.137)$$

which can be used to find (approximately) the optimal value of α.

12.5 Exercises

Exercise 1. Prove that, going to polynomials in three variables (x, y, z), one has

$$\dim H_k^3 = \frac{(k+1)(k+2)}{2}$$

(**Hint**: notice that the spaces of homogeneous polynomials in three variables H_k^3 can be decomposed according to the formula

$$H_k^3 = H_k^2 + H_{k-1}^2 \cdot z + \ldots + H_0^2 \cdot z^k$$

where each subspace is clearly linearly independent of the others).

Exercise 2. Prove that if the n-dimensional linear space X has the basis $\mathbf{x}$, then the components of $\boldsymbol{\lambda}$ in

$$x = \boldsymbol{\lambda}^t \mathbf{x} = \sum_{i=1}^{n} \lambda_i(x) x_i$$

do depend linearly on x, i.e. $\lambda_i(x) \in X'$ ($\boldsymbol{\lambda} \in X'^{(N)}$).

Prove also that $\boldsymbol{\lambda}_i(x)$ is a basis of X'.

(**Hint**: decompose as above x on $\mathbf{x}$ and apply $L(x)$, then use linearity of L).

Exercise 3. Use the property (c) in Definition (12) and (12.14) to verify that (12.15) is a norm, i.e. it satisfies properties (a)–(c) of Definition (8).

Exercise 4. Prove that ℓ^2 is a HS, i.e. that the definition (12.18) is consistent in the sense that the series is convergent because both $\boldsymbol{\lambda}, \boldsymbol{\mu}$ belong to ℓ^2.

(**Hint**: use Schwarz inequality, and the fact that ℓ^2 is complete, i.e. that if $\{\boldsymbol{\lambda}^k\}$ is Cauchy then it has limit $\boldsymbol{\lambda}$ in ℓ^2).

Exercise 5. Take any positive sequence $q \equiv \{q_n\}, q_n > 0$ and define in R^∞ the squared norm

$$|\boldsymbol{\lambda}|_q^2 \equiv \sum_{n=1}^{+\infty} \lambda_n^2 q_n, \tag{12.138}$$

which is clearly related to the scalar product

$$< \boldsymbol{\lambda}, \boldsymbol{\mu} >_{\ell_q^2} = \sum_{n=1}^{+\infty} \lambda_n \mu_n q_n, \tag{12.139}$$

Prove that the space ℓ_q^2 of vectors $\boldsymbol{\lambda}$ satisfying (12.138) is a HS.

(**Hint**: you can take advantage of the equivalence between the two conditions

$$\{\sqrt{q_n}\lambda_n\} \in \ell^2 \Longleftrightarrow \{\lambda_n\} \in \ell^2_q,$$

i.e. by showing that ℓ^2_q is an isometric image of ℓ^2. Isometric means that

$$|\{\sqrt{q_n}\lambda_n\}|_{\ell^2} \equiv |\{\lambda_n\}|_{\ell^2_q}).$$

Prove that:

(a) If

$$0 < \alpha \leq q_n \leq \beta < +\infty$$

then the norm of ℓ^2_q is equivalent to that of ℓ^2, i.e.

$$\alpha|\boldsymbol{\lambda}|^2 \leq |\boldsymbol{\lambda}|^2_{\ell^2_q} \leq \beta|\boldsymbol{\lambda}|^2,$$

(b) That if $q_n \to \infty$ when $n \to \infty$ convergence in ℓ^2_q implies convergence in ℓ^2
(c) That if $q_n \to 0$ when $n \to \infty$ convergence in ℓ^2 implies convergence in ℓ^2_q.

Exercise 6. Prove the following relations, $\forall x, y$ belonging to the HS, X,

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \tag{12.140}$$

$$\|x + y\|^2 - \|x - y\|^2 = 4 < x, y > . \tag{12.141}$$

Notice that (12.141) clarifies that only if $\|x + y\|^2 - \|(x - y)\|^2$ is linear in x (and then also in y) the norm $\|\ \|$ in question can be derived from a scalar product.

Use this statement to prove that the BS of functions continuous on [0,1], i.e. $C([0,1])$, with norm

$$\|f\|_{C[0,1]} = \max_{t\in[0,1]} |f(t)|$$

is not a HS.

(**Hint**: use $f_1(t) = t$, $f_2(t) = 1 - t$, on [0,1] to prove that

$$\|f_1 + f_1\|^2 + \|f_1 - f_2\|^2 = 2$$

while

$$\|f_1\|^2 = 1,\ \|f_2\|^2 = 1$$

so that (12.140) is violated).

Exercise 7 (Deterministic least squares). Consider the case that X is finite dimensional, $X = R^m$, and that S is an n-dimensional subspace

$$S \equiv \text{Span}\{\mathbf{a}_1 \dots \mathbf{a}_n\};$$

namely

$$\widehat{\mathbf{x}} \in S \leftrightarrow \widehat{\mathbf{x}} = \sum_{i=1}^{n} \lambda_i \mathbf{a}_i \equiv A\boldsymbol{\lambda}$$

where A has $\mathbf{a}_i$ as i-th columns.

Introduce in X a W-scalar product as in Example 3.

Prove that the best approximation $\widehat{\mathbf{x}}$ in S (cf. (12.47)) of any $\mathbf{x}$ is just the ordinary l.s. estimator

$$\widehat{\mathbf{x}} = A(A^t W A)^{-1} A^t W \mathbf{x}.$$

In order to avoid confusion between vectors in R^m and the basis, prove this formula developed in terms of components, namely

$$\widehat{\mathbf{x}} = \Sigma_{i,k}\ \mathbf{a}_i (\mathbf{a}_i^t W \mathbf{a}_k)^{(-1)} \mathbf{a}_k^t W \mathbf{x}$$

Exercise 8. Prove that the application of (12.9) and (12.44) to the case $\mathcal{L}_0^2(\Omega)$ (cf. Example 7) provides as $\widehat{X}(\omega)$ the linear regressor of $X(\omega)$ on the basis $\mathbf{X}(\omega)$ and its (squared) estimation error.

Exercise 9. Prove that $S^\perp$ is closed even if S is not.

(**Hint**: use the fact that, due to Schwarz inequality, $< x, y >$ is continuous in y for fixed x).

Exercise 10. By translating and dilating the variables t of Example 8 we obtain the space $L^2([-1, 1])$ with its Fourier representation

$$x(t) = \frac{1}{2}a_0 + \sum_{k=1}^{+\infty}(a_k \cos \pi k t + b_k \sin \pi k t)$$

$$a_k = \int_{-1}^{1} x(\tau) \cos \pi k \tau d\tau,\ b_k = \int_{-1}^{1} x(\tau) \sin \pi k \tau d\tau.$$

Note that in L^2 you find two closed subspaces of even and odd functions

$$f_e(-t) = f_e(t)\ ,\ f_o(-t) = -f_o(t).$$

Prove that $\forall f \in L^2$ there is a decomposition

$$f(t) = f_e(t) + f_o(t)$$

and find the expression of f_e, f_o by using their properties; prove that

$$L_e^2 = \{f \in L^2([-1,1]), f = f_e\}$$
$$L_o^2 = \{f \in L^2([-1,1]), f = f_o\}$$

are two orthogonal complements.

Prove that $\{\cos \pi k\tau,\ k = 0,1,2\ldots\}, \{\sin \pi k\tau, k = 1,2\ldots\}$ are respectively orthogonal bases of L_e^2 and L_o^2.

Exercise 11. By exploiting the reproducing property of $K(t,t')$ and Schwarz inequality prove that

$$\|K(t,\cdot)\|^2 = K(t,t), \tag{12.142}$$
$$|K(t,t')|^2 = |< K(t,\cdot), K(t',\cdot) >|^2 \leq K(t,t)K(t,t') \tag{12.143}$$
$$\|K(t+h,\cdot) - K(t,\cdot)\|^2 = K(t+h,t+h) +$$
$$-2K(t+h,t) + K(t,t). \tag{12.144}$$

Note also that the first relation shows that $K(t,t)$ is always positive.

Exercise 12. Using the Example 10, prove that

$\{y_n(t)\} = \left\{\frac{\sqrt{2}}{\pi n} \sin \pi nt\right\},\ n = 1,2\ldots$ is an ONC system in $H_0^{1,2}$ on the interval [0,1]. Prove that the RK in $H_0^{1,2}$ is

$$K(t,t') = 2\sum_{n=1}^{+\infty} \frac{\sin \pi nt \sin \pi nt'}{\pi^2 n^2} = \begin{cases} t'(1-t)\ ,\ t' \leq t \\ t(1-t')\ ,\ t' \geq t. \end{cases}$$

(**Hint**: remember that the RK is unique, so, going directly to the definition, you need only to prove that $\forall f(t) \in H_0^{1,2}$ you have

$$\int_0^1 D_{t'}K(t,t') \cdot D_{t'}f(t')dt' \equiv f(t),$$

recalling that $f(0) = f(1) = 0$).

Exercise 13. Let $X = H_K$, with

$$K(t,t') = e^{-|t-t'|}\ ,\ t,t' \in R^1.$$

Assume that the observations are

$$\mathbf{x} = \mathbf{L}(x) = \begin{vmatrix} x(t_1) \\ x(t_2) \end{vmatrix} = \begin{vmatrix} x_1 \\ x_2 \end{vmatrix}; \left(-t_1 = t_2 = \frac{1}{2}\log 2\right)$$

take $L() = \mathrm{ev}_t()$ for any $t \geq 0$; prove that

$$\widehat{x}(t) = \begin{cases} \frac{\sqrt{2}}{3}[(2x_2 - x_1)e^t + (2x_1 - x_2)e^{-t}] \,, & 0 \leq t \leq t_2 \\ \sqrt{2}x_2 e^{-t} & t_2 \leq t \end{cases};$$

verify that $\widehat{x}(t_2) = x_2$. For this second case compute $\mathcal{E}_r^2(0)$ and prove that $\mathcal{E}_r^2(0) = \frac{1}{3}$.

Exercise 14. Prove that, taking $T \equiv R^1$,

$$L_{\bar{t}}(x) = \mathrm{ev}_{\bar{t}}(D_t x) = \dot{x}(\bar{t})$$

is not a bounded functional if $X = H_K$ and $K(t,t') = e^{-|t-t'|}$, while it is bounded if $X = H_K$ and $K(t,t') = e^{-(t-t')^2}$. Compute $K(L,L)$ in this second case and prove that

$$K(L_t, L_{t'}) = 2e^{-(t-t')^2} - 4(t-t')^2 e^{-(t-t')^2}.$$

Exercise 15. Take $T \equiv R^2$ and put $r^2 = (t_1 - t_1')^2 + (t_2 - t_2')^2$, $X = H_K$, $K(t,t') = e^{-r^2}$; define also

$$\mathbf{L}(x) = \begin{vmatrix} L_0(x) \\ L_1(x) \\ L_2(x) \end{vmatrix} = \begin{vmatrix} \mathrm{ev}_0(x) \\ \mathrm{ev}_0(D_{t_1}x) \\ \mathrm{ev}_0(D_{t_2}x) \end{vmatrix} = \begin{vmatrix} x(0,0) \\ D_{t_1}x(0,0) \\ D_{t_2}x(0,0) \end{vmatrix}$$

and prove that

$$K(\mathbf{L}, \mathbf{L}^t) = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{vmatrix}.$$

Exercise 16. Following the Example 13, $K(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{1}{n!}(\boldsymbol{\xi}^t \boldsymbol{\eta})^n$ is a RK in H_n^3 (i.e. the space of homogeneous polynomials, in R^3, of degree n). Assume that $P_n(\boldsymbol{\xi})$ is our unknown and that

$$\mathbf{L}(P_n) = \begin{vmatrix} c_1 \\ c_2 \end{vmatrix} = \begin{vmatrix} P_n(\boldsymbol{\xi}_1) \\ P_n(\boldsymbol{\xi}_2) \end{vmatrix}.$$

prove that

$$\widehat{P_n(\xi)} = \lambda_1 \frac{1}{n!}(\boldsymbol{\xi}_1^t \boldsymbol{\xi})^n + \lambda_2 \frac{1}{n!}(\boldsymbol{\xi}_2^t \boldsymbol{\xi})^n$$

where

$$\left|\begin{matrix}\lambda_1\\ \lambda_2\end{matrix}\right| = \frac{n!}{|\boldsymbol{\xi}_1|^2|\boldsymbol{\xi}_2|^2 - (\boldsymbol{\xi}_1^t\boldsymbol{\xi}_2)^2}\left|\begin{matrix}c_1|\boldsymbol{\xi}_1|^2 - c_2\boldsymbol{\xi}_1^t\boldsymbol{\xi}_2\\ -c_1\boldsymbol{\xi}_1^t\boldsymbol{\xi}_2 + c_2|\boldsymbol{\xi}_2|^2\end{matrix}\right|.$$

Exercise 17. Let X be as in Exercise 13, with $K(t,t') = e^{-|t-t'|}$,

$$\mathbf{c} = \left|\begin{matrix}c_1\\ c_2\end{matrix}\right| = \left|\begin{matrix}x(t_1)+\nu_1\\ x(t_2)+\nu_2\end{matrix}\right|$$

$$-t_1 = t_2 = \frac{1}{2}\log 2.$$

Put $q = \alpha\sigma_\nu^2$ and, by applying the Tikhonov optimization with smoothing factor α, prove that

$$\widehat{x}_\alpha(t) = \frac{\sqrt{2}}{4(1+q)^2-1}\{[2(1+q)c_1 - c_2]e^{-t} + [2(1+q)c_2 - c_1]e^t\};\ 0 \le t \le t_2$$

$$\widehat{x}_\alpha(t) = \frac{\sqrt{2}}{4(1+q)^2-1}e^{-t}[2qc_1 + (3+4q)c_2]\,;\quad t_2 \le t.$$

In particular verify that, $\forall\alpha > 0$, i.e. $\forall q > 0$,

$$\widehat{x}(t_2) = \frac{2qc_1 + (3+4q)c_2}{4(1+q)^2-1} \neq c_2$$

and that the following limit relations hold

$$\lim_{\alpha\to 0}\widehat{x}(t_2) = c_2$$

$$\lim_{\alpha\to\infty}\widehat{x}(t_2) = 0$$

in accordance with Remark 10.

Exercise 18. Let X be a RKHS with a general RK, $K(t,t')$. Assume that one value only of $x(t)$ is observed, i.e. putting $K_0 = K(t_0,t_0)$, we have

$$\mathbf{x} = [x(t_0) + \nu] = c$$

$$\mathbf{y} = [K(t_0,\cdot)]$$

$$G = [K_0]$$

$$G_\alpha = [K_0 + \alpha\sigma_\nu^2];$$

verify that the solution $\widehat{x}_\alpha$ and its predicted error $\mathcal{E}^2(x,\alpha)$ are (cf. (12.136))

$$\widehat{x}_\alpha(t) = \frac{K(t_0,t)}{K_0} \cdot c$$

$$\mathcal{E}^2(x,\alpha) = \left(\|x\|^2 - \frac{x(t_0)^2}{K_0} \right) + \alpha^2 x(t_0)^2 \cdot \frac{\sigma_\nu^4}{K_0(K_0+\alpha\sigma_\nu^2)^2} + \frac{\sigma_\nu^2 K_0}{(K_0+\alpha\sigma_\nu^2)^2}.$$

Use in the above expression the approximation (12.137), i.e.

$$\mathcal{E}^2(x,\alpha) = \mathcal{E}_t^2 + \alpha^2 \frac{c^2\sigma_\nu^4}{K_0(K_0+\alpha\sigma_\nu^2)^2} + \frac{\sigma_\nu^2 K_0}{(K_0+\alpha\sigma_\nu^2)^2}$$

to prove that an optimal value for α is

$$\alpha = \frac{K_0}{c^2}.$$

Chapter 13
On Potential Theory and HS of Harmonic Functions

13.1 Outline of the Chapter

The Newton potential of the earth as well as its anomalous gravity potential are harmonic functions outside the earth body B, therefore the interest of geodesy in spaces of harmonic functions is quite justified. More precisely, from the mathematical point of view we are interested in a situation in which B is an open, simply connected bounded set, with a relatively smooth boundary S and $\overline{B}^c = \Omega$ (the complement of the closure of B) is simply connected too. Let us note explicitly that this prevents B from having holes in it or even single points removed.

We start in Sect. 13.2 building some Hilbert spaces of harmonic polynomials, which, being embedded into polynomial spaces, are indeed finite dimensional. In particular it is proved that these have their own reproducing kernels and, by transforming Cartesian into spherical coordinates, a fundamental relation is found between such reproducing kernels and the sequence $P_n(\cos\psi_{xy})$. Each Legendre polynomial multiplied by $|x|^n \cdot |y|^n$ turns out to be the reproducing kernel of the subspace of harmonic polynomials homogeneous of degree n. By using the properties of reproducing kernel Hilbert spaces, illustrated in Part III, Sect. 12.5, then one finds the famous *summation theroem*.

The approach follows the idea in Krarup (2006), though departing form them in some important steps. For other approaches one can consult (Nikiforov and Uvarov 1988).

In Sect. 13.3 all the machinery of Sect. 13.2 is translated into properties of spherical harmonics. When these are considered as a sequence in $L^2(\sigma)$ (space of functions square integrable on the unit sphere S_1) they are proved not only to be orthonormal but also complete. Thus they are an orthonormal basis in $L^2(S_1)$. Going into the matter of more general spaces of harmonic functions, some classical properties are proved like the maximum principle or the principle of identity of harmonic functions.

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_13,

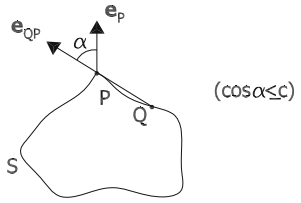


Fig. 13.1 A surface S satisfying the cone condition; $c = \cos\vartheta$

A fundamental result is then established, namely that the sequencies of *internal* as well as that of *external* spherical harmonics, when restricted to any closed bounded and smooth surface S form a complete basis of $L^2(S)$. This implies for instance that any function $f \in L^2(S)$ can be approximated as well as we like, by a finite sum of (external) solid spherical harmonics, i.e. by a *global model*.

The proof that, when we approximate $f \in L^2(S)$ with a sequence of functions harmonic in Ω, i.e. outside S, we also approximate a function u, harmonic in Ω, and that this function, suitably restricted to S, becomes equal to f, is the main purpose of Sect. 13.5. On related matters one can usefully read Fichera (1948).

To do that, the concept of Green's function is introduced and some of its properties are studied. In doing so we create a prototype Hilbert space of harmonic functions, namely that in which potentials in Ω have boundary values in $L^2(S)$.

13.2 Harmonic Functions and Harmonic Polynomials

Recall that in this chapter B is a simply connected open set, as specified at the beginning of Sect. 13.1.

We shall put in the sequel $\overline{B} = B \cup S$, the closure of B and $\Omega = (\overline{B})^c$. We shall assume that S is relatively smooth meaning at least that Gauss' theorem applies to $\overline{B}$, for instance that S satisfies a so-called cone condition, i.e. there is a positive constant $c < 1$ such that for any given point $P \in S$ there is a unit vector $\mathbf{e}_P$ pointing in Ω and a neighborhood $A \subset S$, such that for any other point $Q \in A$ it is $|\mathbf{e}_{QP} \cdot \mathbf{e}_P| \leq c$ with $\mathbf{e}_{QP}$ the unit vector in the direction from Q to P. Looking at Fig. 13.1, one sees that if $c = \cos\vartheta$, with ϑ fixed for the whole surface S, the above means that $\alpha \geq \vartheta$ when $Q \in S$ belongs to a suitable neighborhood A of P.

As a matter of fact in the sequel of these notes we shall require a stronger regularity of the boundary S. For instance we shall assume that the exterior normal field $\mathbf{n}_P$ is everywhere defined on S and even that $\mathbf{n}_P$ is Lipschitz continuous, i.e.

$$P, Q \in S; \quad |\mathbf{n}_P - \mathbf{n}_Q| \leq c \cdot \overline{PQ};$$

this is basically the same as requiring that S has finite curvature at every point.

Under the above mentioned conditions we can apply an inverse radii transform, sometimes also called Kelvin or Rayleigh transform, which is as follows: put the origin O of $\mathbf{R}^3$ in B and take a spherical coordinate system (r, ϑ, λ), then define

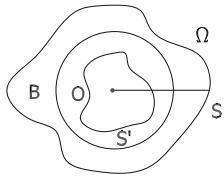


Fig. 13.2 The geometry of Rayleigh transform, with $B' = \mathcal{R}(\Omega)$, $S' = \mathcal{R}(S)$

$$s = \frac{R^2}{r},\ \vartheta' = \vartheta,\ \lambda' = \lambda \tag{13.1}$$

where R is any radius of a sphere totally inside B. Under (13.1), denoting

$$\xi = (r, \vartheta, \lambda),\ \xi' = (s, \vartheta, \lambda),\ \xi' = \mathcal{R}(\xi), \tag{13.2}$$

we obviously have that, putting

$$S' = \mathcal{R}(S),\ B' = \mathcal{R}(\Omega),\ \Omega' = \mathcal{R}(B), \tag{13.3}$$

then S' is totally inside the sphere $r = s = R$ (see Fig. 13.2), B' is inside S' and contains the origin, while Ω' is outside S'. In particular

$$\mathcal{R}(0) = \infty,\ \mathcal{R}(\infty) = 0. \tag{13.4}$$

Definition 1. A function u is harmonic in classical sense in B, denoted $u \in \mathcal{H}(B)$, if it is continuous with its second derivatives in B, and if $\Delta u = 0$ at any point $P \in B$ (recall that B is open). A function u is harmonic and regular in Ω, $u \in \mathcal{H}(\Omega)$, if it is continuous with its second derivatives in Ω and furthermore

$$\lim_{P \to \infty} u(P) = 0; \tag{13.5}$$

(13.5) means that $\forall \varepsilon > 0,\ \exists R_\varepsilon;\ |u(P)| < \varepsilon$ when $r_P > R_\varepsilon$.

Proposition 1. *Let B, S, Ω and B', S', Ω' be as in (13.3). We can show that if $u \in \mathcal{H}(B)$ then, defining*

$$v(s, \vartheta, \lambda) = \frac{1}{s} u\left(\frac{1}{s}, \vartheta, \lambda\right) = \mathcal{R}(u) \tag{13.6}$$

we have $v \in \mathcal{H}(\Omega')$. In other words if we put $v = \mathcal{R}(u)$, the Rayleigh transform of u, we have $\mathcal{R} : \mathcal{H}(B) \to \mathcal{H}(\Omega')$. Similarly $\mathcal{R} : \mathcal{H}(\Omega) \to \mathcal{H}(B')$. Moreover

$$\mathcal{R}^2(u) = \mathcal{R}(v) = u \tag{13.7}$$

for $\forall u \in \mathcal{H}(B)$ or $u \in \mathcal{H}(\Omega)$.

The above statement basically means that, when useful, we can study properties of spaces of harmonic functions on bounded domains, like B, and then derive the corresponding properties for spaces of regular harmonic for functions in Ω.

Among harmonic functions in B a special role play the polynomials which are also harmonic in B. Having to work with polynomials, it is convenient to adopt a multi-index notation already presented in Example 11 of Sect. 12.2.

Remember now that any polynomial $P_N(\xi)$ is defined everywhere in $\mathbf{R}^3$ and that two polynomials which coincide in a neighborhood of a point ξ_0 coincide everywhere, because then in ξ_0 they will have the same derivatives up to order N (all higher order derivatives are zero); this is the principle of identity of polynomials. The following conclusion can be drawn.

Proposition 2. *Any polynomial harmonic in an open set B is harmonic in the whole of $\mathbf{R}^3$, but of course not regular at ∞, unless it is identically zero.*

Proof. Let $P_N(\xi) = \sum_{n=0}^{N} \sum_{|\alpha|=n} c_\alpha \xi^\alpha$, $\xi = (x, y, z)$, be harmonic in an open set B. Then the polynomial of order $N-2$

$$P_{N-2}(\xi) = \sum_{n=0}^{N} \sum_{|\alpha|=n} c_\alpha (\Delta \xi^\alpha) \equiv 0, \quad \xi \in B$$

and therefore $P_{N-2}(\xi)$ is zero everywhere in $\mathbf{R}^3$. □

Accordingly, we can study the space of harmonic polynomials in R^3, without any specific reference to B. We call it $H\mathcal{P}_N^3$, when only polynomials up to degree N are taken into account. As it is obvious

$$H\mathcal{P}_N^3 \subset \mathcal{P}_N^3,$$

the space of all polynomials, already studied in Examples 1 and 11 in Chap. 12.

Since $\mathcal{P}_N^3$ is a HS of finite dimension, $H\mathcal{P}_N^3$ will also be a finite dimensional HS, under the same scalar product. In particular, to the orthogonal decomposition (see Example 11)

$$\mathcal{P}_N^3 = H_0^3 \oplus H_1^3 \oplus H_2^3 \ldots H_N^3, \tag{13.8}$$

where H_k^3 are spaces of polynomials homogeneous in ξ of degree k, there must correspond an analogous decomposition

$$H\mathcal{P}_N^3 = HH_0^3 \oplus HH_1^3 \oplus \ldots \oplus HH_N^3 \tag{13.9}$$

where each HH_k^3 contains all the harmonic polynomials, homogeneous of degree k. Note that the orthogonality of HH_ℓ^3 and $HH_k^3, \ell \neq k$, is already guaranteed by (13.8) and the obvious fact that

$$HH_k^3 \subset H_k^3. \tag{13.10}$$

The structure of the reasoning followed here is based on Krarup (2006) and has been used to develop explicit formulas for the traditional spherical harmonics.

Since this will be useful in our construction, we will simultaneously reason on $\mathcal{P}_N^2$, $H\mathcal{P}_N^2$ and HH_k^2.

Our first target will be to count the dimensions of HH_k^2 and HH_k^3.

Definition 2. In order to avoid confusion, let us agree on some notation. We put

$$\begin{aligned}
\boldsymbol{\chi} &= (x, y) \\
\rho &= |\boldsymbol{\chi}| \\
\boldsymbol{\xi} &= (x, y, z) = (\boldsymbol{\chi}, z) \\
r &= |\boldsymbol{\xi}| = \sqrt{\rho^2 + z^2}.
\end{aligned}$$

In particular $\rho = |\boldsymbol{\chi}|$ used in this context should be not be confused with the symbol ρ used sometimes in the text for mass density.

Proposition 3. *We have, with obvious notation,*

$$D_k^2 = \dim HH_k^2 = 2 - \delta_{k0}; \tag{13.11}$$

for each $k \neq 0$, the two homogeneous polynomials are given by the formulas

$$h_k(\boldsymbol{\chi}) = Re(x + iy)^k;\ h_{-k}(\boldsymbol{\chi}) = Im(x + iy)^k; \tag{13.12}$$

furthermore h_k and h_{-k} are orthogonal in H_k^2.

Proof. We note first of all that $HH_0^2 \equiv H_0^2$ with $h_0 \equiv 1$ being the unique linearly independent element, homogenous of degree zero. All the other elements of HH_0^2 are just constant everywhere. Similarly $HH_1^2 \equiv H_1^2$ and all homogenous (and harmonic) polynomials of degree 1 are obtained by combination of $h_1 \equiv x, h_{-1} \equiv y$; all that agrees with the statement of the proposition.

Now let $h(\boldsymbol{\chi}) \in HH_k^2$ with $k \geq 2$; then we can put $x = \rho\cos\lambda$, $y = \rho\sin\lambda$ and we have

$$h(\boldsymbol{\chi}) = \rho^k f_k(\lambda). \tag{13.13}$$

Let us impose to (13.13) to satisfy the Laplace equation in polar coordinates (ρ, λ) in R^2, i.e.

$$\Delta h(\boldsymbol{\chi}) = \left(\frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\lambda^2}\right) h(\boldsymbol{\chi}) = \rho^{k-2}\left[k^2 f_k(\lambda) + f_k''(\lambda)\right] \equiv 0.$$

This implies that

$$f_k(\lambda) = a_k \cos k\lambda + b_k \sin k\lambda, \tag{13.14}$$

i.e. we have two independent solutions, in polar coordinates, namely

$$k > 0,\ h_k(\rho,\lambda) = \rho^k \cos k\lambda\ ,\ h_{-k}(\rho,\lambda) = \rho^k \sin k\lambda \tag{13.15}$$

Now it is enough to observe that

$$\rho^k \cos k\lambda = Re(x+iy)^k,\ \rho^k \sin k\lambda = Im(x+iy)^k$$

to prove (13.12).

That h_k and h_{-k} are orthogonal to one another derives from the development of $(x+iy)^k$ in a binomial formula; separating the real from the imaginary part we see that they are linear combinations of monomials χ^α which can never be the same. Since such monomials are reciprocally orthogonal (cf. Example 11), we have proved what we wanted. □

Proposition 4. *Let us split H_k^3 into two orthogonal complements*

$$H_k^3 = HH_k^3 \oplus CH_k^3;\ CH_k^3 = (HH_k^3)^\perp, \tag{13.16}$$

then we have

$$CH_k^3 \equiv \{P_k(\xi) = r^2 P_{k-2}(\xi),\ P_{k-2} \in H_{k-2}^3\}; \tag{13.17}$$

furthermore, adopting a notation similar to (13.11),

$$D_k^3 = \dim HH_k^3 = 2k+1. \tag{13.18}$$

Proof. First we immediately see that CH_k^3 defined by (13.17) is orthogonal to HH_k^3; in fact let $h_k(\xi) \in HH_k^3$, then (cf. (12.80))

$$\forall P_{k-2} \in H_{k-2}^3,\ \ < r^2 P_{k-2}(\xi), h_k(\xi) > = P_{k-2}(\partial_\xi)\Delta h_k(\xi)\big|_{\xi=0} \equiv 0. \tag{13.19}$$

We note too, that CH_k^3 is a closed subspace of H_k^3. Now we have to show that CH_k^3 covers the whole complement of HH_k^3; it is enough to show that the orthogonal complement of CH_k^3, defined by (13.17), is in fact HH_k^3.

Since CH_k^3 is closed, $\forall P_k \in H_k^3$ we can make the orthogonal decomposition

$$P_k(\xi) = r^2 P_{k-2}(\xi) + R_k(\xi) \tag{13.20}$$

with $R_k(\xi) \perp CH_k^3$ i.e. R_k belongs to the orthogonal complement of CH_k^3. But this implies

$$< r^2 P_{k-2}, R_k >= P_{k-2}(\partial_\xi)\Delta R_k\big|_{\xi=0} =< P_{k-2}(\xi), \Delta R_k(\xi) >= 0, \quad (13.21)$$

$\forall P_{k-2} \in H^3_{k-2}$. Equation (13.21) then implies $\Delta R_k(\xi) = 0$ because this is a polynomial in H^3_{k-2}. Equation (13.20) is, as a matter of fact, (13.16).

Now, since CH^3_k is one to one with H^3_{k-2}, we have (cf. Exercise 1, Chap. 1)

$$\dim CH^3_k = \dim H^3_{k-2} = \frac{(k-1)k}{2};$$

(13.18) follows from

$$\begin{aligned} D^3_k &= \dim H^3_k - \dim CH^3_k \\ &= \frac{(k+1)(k+2)}{2} - \frac{(k-1)k}{2} = 2k+1. \end{aligned}$$ □

The Exercise 3 is preparatory for the next proposition; the reader is advised at least to read it before continuing.

Proposition 5. *The following decomposition formula holds*

$$P_N(\xi) = h_N(\xi) + r^2 h_{N-2}(\xi) + r^4 h_{N-4}(\xi) + \ldots \quad (13.22)$$

the summation of terms $r^{2k} h_{N-2k}(\xi)$ being extended up to $k = \left[\frac{N}{2}\right]$ (the smallest integer $\leq N/2$); as shown in Exercise 3 each $h_{N-2k}(\xi)$ is a harmonic polynomial in HH^3_{N-2k}; $h_N(\xi)$, which is the orthogonal projection of $P_N(\xi)$ onto HH^3_N, is given by the inverse formula

$$h_N(\xi) = P_N(\xi) + q_1 r^2 \Delta P_N(\xi) + q_2 r^4 \Delta^2 P_N(\xi) + \ldots \quad (13.23)$$

where q_k are suitable constants independent of the specific P_N once N is fixed.

Equation 13.23 is also known as Pizzetti's formula *in mathematical literature (Dunford and Schwarz 1958; Courant and Hilbert 1962).*

Proof. We just re-write (13.20) in the form

$$P_N(\xi) = h_N(\xi) + r^2 P_{N-2}(\xi), \quad (13.24)$$

where, as we have seen, h_N is harmonic, i.e. $h_N \in HH^3_N$. By iterating (13.24) we get (13.22). Notice that since the degree jumps 2 by 2 from N, we end up with $P_1(\xi)$ or $P_0(\xi)$, depending whether N is odd or even. But $P_1(\xi)$ or $P_0(\xi)$ are already harmonic by default.

Now we apply to (13.22) successively $r^2\Delta, r^4\Delta^2 \ldots$ and we get, taking Exercise 3 into account,

$$\begin{aligned} r^2\Delta P_N &= A_{11}r^2 h_{N-2} + A_{12}r^4 h_{N-4} + A_{13}r^6 h_{N-6} + \dots \\ r^4\Delta^2 P_N &= A_{22}r^4 h_{N-4} + A_{23}r^6 h_{N-6} + \dots \\ r^6\Delta^3 P_N &= A_{33}r^6 h_{N-6} + \dots \end{aligned} \tag{13.25}$$

As we see (13.25) can be considered as a triangular system with $(r^{2k}h_{N-2k})$ as unknowns and $(r^{2k}\Delta^k P_N)$ as known terms. Since (cf. Exercise 3)

$$A_{kk} = 2k \cdot (2k-2) \dots 2 \cdot (2N-2k+1) \cdot (2N-2k-1) \dots (2N-4k+3)$$

are always positive (remember that k goes from 1 to $\left[\frac{N}{2}\right]$) the system is invertible. So solving (13.25) and substituting back in (13.22) we get the expression (13.23). We note that q_k can be computed as we suggest in Exercise 4, however here it is only important to strengthen that q_k are independent of P_N, i.e. they are the same $\forall P_N \in H_N^3$. □

At this point we are ready to derive the first important result of this chapter. In fact we note that the elements of HH_N^3, being also elements of H_N^3, enjoy the reproducing property

$$h_N(\boldsymbol{\xi}) = < K_N(\boldsymbol{\xi}, \boldsymbol{\eta}), h_N(\boldsymbol{\eta}) >_{H_k^3} \tag{13.26}$$

with (cf. Example 13)

$$K_N(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{(\boldsymbol{\xi}^t \boldsymbol{\eta})^N}{N!}; \tag{13.27}$$

nevertheless $K_N(\boldsymbol{\xi}, \boldsymbol{\eta})$ is not the reproducing kernel of HH_N^3 because for any fixed $\overline{\eta}$, $K(\boldsymbol{\xi}; \overline{\boldsymbol{\eta}})$ does not belong to HH_N^3, namely it is not harmonic. The next Theorem will provide us with the correct RK of HH_N^3, which is nothing but the orthogonal projection of K_N onto HH_N^3. Hereafter we switch from N to n, to allow for the distinction of the maximum degree of the polynomial, N, from the homogeneous degree n.

Theorem 1. *Each subspace of homogeneous harmonic polynomials HH_n^3 is endowed with a RK, $H_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ given by*

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = A_n r_{\boldsymbol{\xi}}^n r_{\boldsymbol{\eta}}^n P_n(t) \tag{13.28}$$

where

$$\begin{aligned} & r_\xi = |\boldsymbol{\xi}|,\ r_\eta = |\boldsymbol{\eta}|,\ t = \frac{\xi^t \eta}{r_\xi r_\eta} = \cos\psi_{\xi\eta}, \\ & A_n = \frac{2^n n!}{(2n)!} \end{aligned} \tag{13.29}$$

and $P_n(t)$ are exactly the Legendre polynomials of degree n, already seen in Part I, Chap. 3.

Proof. Take $K_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ and apply to it, considered as a function of η, the formula (13.23), so as to define

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = K_n(\boldsymbol{\xi}, \boldsymbol{\eta}) + \sum_{k=1}^{[\frac{n}{2}]} q_k r_\eta^{2k} \Delta_\eta^k K_n(\boldsymbol{\xi}, \boldsymbol{\eta}). \tag{13.30}$$

By using the following, easy to prove, formula

$$\Delta_\eta^k (\boldsymbol{\xi}^t \boldsymbol{\eta})^n = n(n-1)\dots(n-2k+1) r_\xi^{2k} (\boldsymbol{\xi}^t \boldsymbol{\eta})^{n-2k}$$

in (13.30) we receive

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{1}{n!} \left\{ (\boldsymbol{\xi}^t \boldsymbol{\eta})^n + \sum_{k=1}^{[\frac{n}{2}]} p_k r_\eta^{2k} r_\xi^{2k} (\boldsymbol{\xi}^t \boldsymbol{\eta})^{n-2k} \right\} \tag{13.31}$$

where we have set

$$p_k = q_k n(n-1)\dots(n-2k+1). \tag{13.32}$$

If we put in evidence in (13.31) r_ξ^n, r_η^n, and we agree that $p_0 = 1$, we can write

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{r_\xi^n r_\eta^n}{n!} \sum_{k=0}^{[\frac{n}{2}]} p_k t^{n-2k} \equiv r_\xi^n r_\eta^n Q_n(t), \tag{13.33}$$

where $Q_n(t)$ is a polynomial in $t = \cos \psi_{\xi\eta}$ containing only even or odd powers, according to the parity of n. We note that $H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = H_n(\boldsymbol{\eta}, \boldsymbol{\xi})$, that $H_n(\overline{\boldsymbol{\xi}}, \boldsymbol{\eta})$ is harmonic in η by definition, and therefore $H(\boldsymbol{\xi}, \overline{\boldsymbol{\eta}})$ is harmonic in $\boldsymbol{\xi}$ too, and finally that $H_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ has the reproducing property in HH_n^3 because

$$\begin{aligned} < H_n(\boldsymbol{\xi}, \boldsymbol{\eta}), h_n(\eta) > &= < K_n(\boldsymbol{\xi}, \boldsymbol{\eta}), h_n(\boldsymbol{\eta}) > \\ &\quad + \sum_{k=1}^{[\frac{n}{2}]} p_k r_\xi^{2k} < r_\eta^{2k} (\boldsymbol{\xi}^t \boldsymbol{\eta})^{n-2k}, \ h_n(\boldsymbol{\xi}) > \\ &= < K_n(\boldsymbol{\xi}, \boldsymbol{\eta}), \ h_n(\boldsymbol{\eta}) >= h_n(\boldsymbol{\xi}). \end{aligned}$$

In fact all the terms multiplied by a power of r_η^{2k} are orthogonal to all harmonic polynomials.

Observe that $H_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ has to be harmonic in η, whatever is the vector $\boldsymbol{\xi}$, since neither the coefficients q_k nor p_k have to depend on the specific homogenous polynomial $(\boldsymbol{\xi}^t \boldsymbol{\eta})^n$, once n is fixed.

Then we can choose $\overline{\boldsymbol{\xi}} = (0, 0, 1)$, i.e. the unit vector along the z axis; then $t = \cos\vartheta$ and we must have that

$$H_n(\overline{\boldsymbol{\xi}}, \boldsymbol{\eta}) = r_\eta^n Q_n(t) \tag{13.34}$$

is harmonic. On the other hand we already know that $r^n P_n(t)$ is harmonic when $P_n(t)$ is a Legendre polynomial and this means that we must have

$$Q_n(t) = A_n P_n(t). \tag{13.35}$$

In fact, writing the Laplacian first in spherical coordinates (r, ϑ, λ) and then changing ϑ into $t = \cos\vartheta$, one gets

$$\begin{aligned} \Delta &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\left[\frac{\partial^2}{\partial \vartheta^2} + \operatorname{ctg}\vartheta\frac{\partial}{\partial\vartheta} + \frac{1}{\sin^2\vartheta}\frac{\partial^2}{\partial\lambda^2}\right] \\ &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\left[(1-t^2)\frac{\partial^2}{\partial t^2} - 2t\frac{\partial}{\partial t} + \frac{1}{1-t^2}\frac{\partial^2}{\partial\lambda^2}\right] \end{aligned} \tag{13.36}$$

We can use (13.36) on (13.34) to conclude that

$$r^{n-2}\left[n(n+1)Q_n(t) + (1-t^2)\frac{\partial^2}{\partial t^2}Q_n(t) - 2t\frac{\partial}{\partial t}Q_n(t)\right] \equiv 0. \tag{13.37}$$

If $Q_n(t)$ has to satisfy (13.37) and to be a polynomial with the same parity as n, then Q_n is fixed up to a constant (see the Exercise 5). Since $P_n(t)$ satisfies (13.37), the relation (13.35) must be true, so we have only to find A_n. We note that, (cf. (13.33) and (13.34)) the coefficient of t^n in $Q_n(t)$ is just $\frac{1}{n!}$. The coefficient of t^n in $P_n(t)$ is $\frac{2n!}{2^n(n!)^2}$ (see Exercise 6). So we must have $A_n = 2^n n!/(2n)!$, as it was to be proved. □

Definition 3. For reasons that will become soon clear, we define

$$L_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = r_\xi^n r_\eta^n (2n+1) P_n(\cos\psi_{\xi\eta}) \tag{13.38}$$

so that we have (see (13.35))

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{A_n}{2n+1} L_n(\boldsymbol{\xi}, \boldsymbol{\eta}). \tag{13.39}$$

Remark 1. Let us remember that $H_n(\boldsymbol{\xi}, \boldsymbol{\eta})$, considered as a family of functions of $\boldsymbol{\eta}$ indexed by $\boldsymbol{\xi}$, is total in HH_n^3 (see Proposition 15, Example 13); the same then must be true for $L_n(\boldsymbol{\xi}, \boldsymbol{\eta})$.

On the other hand we know that HH_n^3 has dimension $2n+1$ (cf. (13.17)), therefore there must be $(2n+1)$ points $\xi_i \neq 0$ such that $\{L_n(\boldsymbol{\xi}_i, \boldsymbol{\eta})\}$ is a basis

of HH_n^3. Furthermore, since $L_n(\lambda\boldsymbol{\xi}_i, \boldsymbol{\eta}) = \lambda^n L_n(\boldsymbol{\xi}_i, \boldsymbol{\eta})$, the point $\boldsymbol{\xi}_i$ can be chosen to belong to S_1, the sphere of radius 1.

This means that $\forall h_n \in HH_n^3$ we can put

$$h_n(\cdot) = \sum_{i=1}^{2n+1} \lambda_i L_n(\boldsymbol{\xi}_i, \cdot) \tag{13.40}$$

and that this correspondence is one to one, so that $h_n = 0 \Leftrightarrow \{\boldsymbol{\lambda}\} = 0$. Therefore, using (13.39) and the fact that $H_n(\cdot)$ is a RK,

$$\forall \boldsymbol{\lambda} \neq 0,\ \|h_n\|^2 = \frac{(2n+1)}{A_n} \Sigma_{i,j} \lambda_i \lambda_j L_n(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) > 0 \tag{13.41}$$

and we see that $\{L_n(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)\}$ is an invertible matrix.

On the other hand we know that (cf. Part I, (3.188)) $L_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ has also a nice reproducing property when $\boldsymbol{\xi}, \boldsymbol{\eta} \in S_1$ and we adopt an $L^2(S_1)$ scalar product; namely

$$\frac{1}{4\pi} \int_\sigma L_n(\boldsymbol{\xi}, \boldsymbol{\eta}) L_n(\boldsymbol{\xi}', \boldsymbol{\eta}) d\sigma_\eta = < L_n(\boldsymbol{\xi}, \cdot), L_n(\boldsymbol{\xi}', \cdot) >_{L^2(S_1)} = L_n(\boldsymbol{\xi}, \boldsymbol{\xi}').$$

Even more, we know that (cf. Part I, (3.182))

$$\begin{aligned} \frac{1}{4\pi} \int L_n(\boldsymbol{\xi}, \boldsymbol{\eta}) L_m(\boldsymbol{\xi}', \boldsymbol{\eta}) d\sigma_\eta &= < L_n(\boldsymbol{\xi}, \boldsymbol{\eta}), L_m(\boldsymbol{\xi}', \cdot) >_{L^2(S_1)} \\ &= \delta_{nm} L_n(\boldsymbol{\xi}, \boldsymbol{\xi}'). \end{aligned} \tag{13.42}$$

All that allows us to draw a number of conclusions that we state in the form of three Lemmas.

Lemma 1. *Let us introduce the trace operator* $\Gamma_{S_1} : \mathcal{P}_N^3 \to L^2(S_1)$

$$\forall P_N \in \mathcal{P}_N^3;\ \ \Gamma_{S_1} P_N(\boldsymbol{\xi}) = P_N(\boldsymbol{\xi})\big|_{r=1};$$

then the image of HH_n^3 *in* $L^2(S_1)$*, i.e.*

$$\Gamma_{S_1}(HH_n^3) \equiv Span_{\boldsymbol{\xi} \in S_1}\{L_n(\boldsymbol{\xi}, \boldsymbol{\eta})\} \equiv Span_{\boldsymbol{\xi} \in S_1}\{P_n(\cos \psi_{\boldsymbol{\xi}, \boldsymbol{\eta}})\}, \tag{13.43}$$

is isometric to HH_n^3*, up to a constant. Since by combining (13.40) and (13.41) we see that*

$$\forall h_n \in HH_n^3;\ \|h_n\|^2_{HH_n^3} = \frac{2n+1}{A_n} \|\Gamma_{S_1}(h_n)\|^2_{L^2(S_1)}; \tag{13.44}$$

in particular this implies that any two vectors orthogonal in HH_n^3 *with its original scalar product are orthogonal in* $L^2(S_1)$ *too and viceversa.*

Lemma 2. *If we consider the decomposition*

$$H\mathcal{P}_N^3 = HH_1^3 \oplus HH_2^3 \oplus, \ldots, \oplus HH_n^3, \tag{13.45}$$

which is orthogonal in the original topology in HH_n^3, we see that, thanks to (13.41), the same decomposition for the image $\Gamma_{S_1}(H\mathcal{P}_N^3)$ is orthogonal in $L^2(S_1)$ too, and

$$\forall P_N \in H\mathcal{P}_N^3;\ \|P_N\|^2_{H\mathcal{P}_N^3} = \sum_{n=0}^{N} \frac{(2n+1)}{A_n} \|\Gamma_{S_1}(P_N)\|^2_{L^2(S_1)}. \tag{13.46}$$

Equation 13.46 is a norm equivalence for any fixed N, but not when $N \to \infty$. Basically this means that the geometry of $H\mathcal{P}_N^3$ with the original scalar product and with the product of $L^2(S_1)$ is the same.

Lemma 3. *Any element $P_N \in H\mathcal{P}_N^3$ is uniquely determined by its trace on S_1.*

That this occurs for each component $h_n \in HH_n^3$ makes no surprise because $h_n(\lambda\boldsymbol{\xi}) = \lambda^n h_n(\boldsymbol{\xi})$, and then if we give h_n on S_1 we fix it in the whole of $\mathbf{R}^3$. But the Lemma claims that this is the same for all $P_N \in H\mathcal{P}_N^3$. The reason is that the following representation holds

$$\forall P_N \in H\mathcal{P}_N^3;\ P_N(\boldsymbol{\xi}) = \sum_{n=0}^{N} h_n(\boldsymbol{\xi}) = \sum_{n=0}^{N} \sum_{i=1}^{2n+1} \lambda_{ni} L_n(\boldsymbol{\xi}_{ni}, \boldsymbol{\xi}); \tag{13.47}$$

therefore

$$< P_N(\boldsymbol{\xi}), L_m(\boldsymbol{\xi}_{mj}, \boldsymbol{\xi}) >_{L^2(S_1)} = \sum_{i=1}^{2m+1} L_m(\boldsymbol{\xi}_{mi}, \boldsymbol{\xi}_{mj}) \lambda_{mi}, \tag{13.48}$$

$$m = 0, 1, \ldots, N,\ j = 1, 2, \ldots, 2m+1.$$

Equation 13.48 is a set of $N+1$ systems, one for each m, whose solutions exists as a consequence of (13.39). Since the known terms of (13.48) depend only on $P_N(\boldsymbol{\xi})$ on S_1, the Lemma is proved.

In a sense Lemma 3 is nothing but a theorem of existence of the solution of the Dirichlet problem for Laplace equation in polynomial spaces. In fact if we go back to (13.22) we see that $\forall P_N \in \mathcal{P}_N^3$, taking its trace on S_1, i.e. putting $r = \|\xi\| = 1$, we get the same function as the trace of the polynomial $h_N(\boldsymbol{\xi}) + r^2 h_{N-2}(\boldsymbol{\xi}) + \ldots$ and such a trace, as we saw in Lemma 3, is sufficient to know each individual component.

Then, as nicely stated in Krarup (2006): "given any polynomial $P_N(\boldsymbol{\xi})$ in B_1 there is one and only one harmonic polynomial agreeing with it on S_1."

The above reasoning and Theorem 1 lead us to one of the main results of this chapter, which we propose in the form of a Theorem.

Theorem 2 (Summation theorem). *Given in HH_n^3 any ON set of polynomials $\{\varphi_{nm}(\boldsymbol{\xi})\}$, that we shall call spherical harmonics of degree n and order m, we must have*

$$H_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \sum_{m=1}^{2n+1} \varphi_{nm}(\boldsymbol{\xi})\varphi_{nm}(\boldsymbol{\eta}); \tag{13.49}$$

because of (13.38) and (13.39), by simply changing the normalization of $\varphi_{nm}(\boldsymbol{\xi})$, i.e. putting

$$\varphi_{nm}(\boldsymbol{\xi}) = \sqrt{\frac{A_n}{2n+1}}\overline{\varphi}_{nm}(\boldsymbol{\xi}) \tag{13.50}$$

we get

$$L_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \sum_{m=1}^{2n+1} \overline{\varphi}_{nm}(\boldsymbol{\xi})\overline{\varphi}_{nm}(\boldsymbol{\eta}); \tag{13.51}$$

$\{\overline{\varphi}_{nm}\}$ are then normalized in $L^2(S_1)$, contrary to $\{\varphi_{nm}\}$ that are normalized in HH_n^3.

Proof. Simply apply Theorem 3 on RKHS. □

13.3 Spherical Harmonics

We can observe that (13.51) holds whatever is the CON system $\{\overline{\varphi}_{nm}(\boldsymbol{\xi})\}$; however there is a particular system of this kind, that we shall study in detail in the next proposition, characterized by the fact that if we express $\boldsymbol{\xi}$ in polar coordinates (r, ϑ, λ) we obtain spherical harmonics in which the three variables separate, in the sense that

$$\begin{aligned} &\overline{\varphi}_{nm}(\boldsymbol{\xi}) = r^n Y_{nm}(\vartheta, \lambda) = r^n f_m(\lambda)\overline{P}_{nm}(\vartheta) \\ &f_m(\lambda) = \cos m\lambda, \;\; f_{-m}(\lambda) = \sin m\lambda, \;\; m = 0, 1, 2 \ldots n \end{aligned} \tag{13.52}$$

Such functions are called, by antonomasia, *inner solid spherical harmonics*. The adjective "inner" refers to the fact that one can apply to (13.52) the Rayleigh transform (see Proposition 1) with respect to the unit sphere, $R = 1$, to obtain the "outer" *solid spherical harmonic*

$$\widetilde{\varphi}_{nm}(\boldsymbol{\xi}) = \frac{1}{r^{n+1}} Y_{nm}(\vartheta, \lambda) \tag{13.53}$$

which are regular harmonic functions in the whole $\mathbf{R}^3$, including the infinity but excluding the origin.

We note that, since $\overline{\varphi}_{nm}(\boldsymbol{\xi})$ are normalized in $L^2(S_1)$, we must have in fact

$$\frac{1}{4\pi}\int Y_{nm}(\vartheta,\lambda)Y_{jk}(\vartheta,\lambda)dSq = \delta_{nj}\delta_{mk}. \tag{13.54}$$

The function $Y_{nm}(\vartheta,\lambda)$ are called *surface spherical harmonics*: they are the trace on S_1 of solid spherical harmonics. The indexes n and m of Y_{nm} are called respectively the *degree* and the *order* of the spherical harmonics.

Now if we use (13.52) in (13.51) we get a very useful, and widely used, Corollary.

Corollary 1. *We have*

$$P_n(\cos\psi_{\xi\eta}) = \frac{1}{2n+1}\sum_{m=-n}^{n} Y_{nm}(\vartheta_\xi,\lambda_\xi)Y_{nm}(\vartheta_\eta,\lambda_\eta), \tag{13.55}$$

where $\boldsymbol{\xi} = (r_\xi,\vartheta_\xi,\lambda_\xi)$, $\boldsymbol{\eta} = (r_\eta,\vartheta_\eta,\lambda_\eta)$ *and* $\psi_{\xi\eta}$ *is the spherical angle between the directions of* $\boldsymbol{\xi}$ *and* $\boldsymbol{\eta}$.

Proposition 6. *For every degree* n *we find* $2n+1$ *homogeneous harmonic polynomials of the form (see Definition 2 and formula (13.52) for the notation)*

$$S_{nm}(\rho,\lambda,z) = \rho^{|m|} f_m(\lambda) Q_{n-|m|}(\rho,z) \tag{13.56}$$
$$m = -n,-n+1,\ldots,n-1,n.$$

Note that these $Q_{n-|m|}(\rho,z)$ *should not be confused with the Legendre functions of second kind, which by the way are functions of one variable only.*

In (13.56) $Q_{n-|m|}(\rho,z)$ *is a polynomial homogenous of degree* $n-|m|$ *in* (ρ,z), *with a form of the type*

$$Q_{n-|m|}(\rho,z) = z^{n-|m|} + q_1 z^{n-|m|-2}\rho^2 + \ldots \tag{13.57}$$
$$= \sum_{k=0}^{I} q_k z^{n-|m|-2k}\rho^{2k}$$

where we have put for the sake of simplicity

$$I = \left[\frac{n-|m|}{2}\right] \tag{13.58}$$

and

$$q_0 = 1. \tag{13.59}$$

The functions $S_{nm}(\rho,\lambda,z)$ *are called solid spherical harmonics and when we go to spherical coordinates* (r,ϑ,λ) *by putting* $\rho = r\sin\vartheta, z = r\cos\vartheta$ *they get the form*

$$S_{nm}(r,\vartheta,\lambda) = r^n f_m(\lambda)\widetilde{P}_{n-|m|}(\vartheta) \tag{13.60}$$

where

$$\widetilde{P}_{n-|m|}(\vartheta) = (\sin\vartheta)^{|m|} Q_{n-|m|}(\sin\vartheta, \cos\vartheta) \tag{13.61}$$
$$= (\sin\vartheta)^{|m|} \sum_{k=0}^{I} q_k (\cos\vartheta)^{n-|m|-2k} (1-\cos^2\vartheta)^k .$$

i.e. $\widetilde{P}_{nm}(\vartheta)$ *is the product of* $(\sin\vartheta)^{|m|}$ *by a polynomial of degree* $n-|m|$ *in* $\cos\vartheta$.

Furthermore the functions $S_{nm}(r,\vartheta,\lambda)$ *are* $L^2(S_1)$ *orthogonal, i.e.*

$$\frac{1}{4\pi}\int_{S_1} S_{n,m}(1,\vartheta,\lambda) S_{n,m'}(1,\vartheta,\lambda) dS_1 = 0 \quad m \neq m'. \tag{13.62}$$

Finally we note that the polynomials $Q_{n-|m|}(\rho,z)$ *in (13.56) and therefore the functions* $\widetilde{P}_{n,m}(\vartheta)$ *in (13.61) are defined up to a proportionality constant which here is fixed by the normalization condition (13.59); we shall see in the sequel other normalization conditions for such functions.*

Proof. We basically must prove that there exist constants $q_0 = 1, q_1, \ldots, q_I$, univocally fixed by the condition that $S_{n,m}(\rho,\lambda,z)$, given by (13.56) and (13.57), satisfies Laplace equation.

First of all we observe that if we put $m = \pm n$ (observe that we claimed $Q_{n-|m|}$ to depend on $|m|$ and not on m as it will be soon justified) in (13.56), we get $Q_0(\rho,z)$ which reduces to $Q_0 \equiv 1$ and

$$S_{n,\pm n} = \rho^n f_{\pm n}(\lambda) \tag{13.63}$$

which are harmonic in (x,y) (cf. (13.66)) and therefore also in (x,y,z).

Furthermore with $m = \pm(n-1)$, (13.56) yields

$$S_{n,\pm(n-1)} = q_1 z \rho^{n-1} f_{\pm(n-1)}(\lambda) \tag{13.64}$$

which is again straightforwardly harmonic in (x,y,z), because it is the first order in z and harmonic in (x,y).

Now take $S_{n,m}$ from (13.56) and impose to it to satisfy Laplace equation which we choose to write in cylindrical coordinates, namely putting

$$\Delta = \frac{\partial^2}{\partial z^2} + \frac{\partial^2}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\lambda^2}. \tag{13.65}$$

Considering that one has

$$\nabla f_m(\lambda)\cdot\nabla\rho^{|m|} Q_{n-|m|}(\rho,z) \equiv 0$$

because (ρ, λ, z) is an orthogonal coordinate system, one finds

$$\begin{aligned}
&\Delta\left[f_m(\lambda)\rho^{|m|}Q_{n-|m|}(\rho,z)\right] \\
&\quad = \Delta\left[f_m(\lambda)\right]\cdot\rho^{|m|}Q_{n-|m|}(\rho,z) + f_m(\lambda)\Delta\left[\rho^{|m|}Q_{n-|m|}(\rho,z)\right] \qquad (13.66)\\
&\quad = f_m(\lambda)\left\{-m^2\rho^{|m|-2}Q_{n-|m|} + \Delta[\rho^{|m|}Q_{n-|m|}]\right\} = 0
\end{aligned}$$

implying that the expression in parenthesis has to be zero. So, we are justified to assume $\rho^{|m|}Q_{n-|m|}$ to depend on $|m|$ as opposed to m. In other words, to both $f_m(\lambda)$ and $f_{-m}(\lambda)$ we can associate the same $\rho^m Q_{n-m}$ with $m \geq 0$. Considering the previous remark (see (13.63) and (13.64)) we can now assume that

$$m = 0, 1, \ldots n-1. \qquad (13.67)$$

Substituting (13.57) into (13.66) and considering that

$$\begin{aligned}
\Delta z^{n-m-2k}\rho^{m+2k} &= (n-m-2k)(n-m-2k-1)z^{n-m-2k-2}\rho^{m+2k} \\
&\quad + (m+2k)^2 z^{n-m-2k}\rho^{m+2k-2},
\end{aligned}$$

we find

$$\begin{aligned}
&\sum_{k=0}^{I}(n-m-2k)(n-m-2k-1)z^{n-m-2k-2}\rho^{m+2k}q_k \qquad (13.68)\\
&\quad + \sum_{k=0}^{I}4k(m+k)z^{n-m-2k}\rho^{m+2k-2}q_k = 0
\end{aligned}$$

We explicitly note that the last term of the first summation is always zero because we have (remember (13.58))

$$(n-m-2I)(n-m-2I-1) \equiv 0,$$

while the first term of the second summation is also zero because of the factor k. As a result the two sums in (13.68) contain exactly the same monomials and equating the corresponding coefficients we get the recursive relation

$$q_k = -\frac{(n-m-2k+2)(n-m-2k+1)}{4k(m+k)}q_{k-1}. \qquad (13.69)$$

The equation (13.69) fixes all q_k when the normalization (13.59) is assumed.

The second claim of proposition is elementary in nature because

$$S_{nm}(1,\vartheta,\lambda) = f_m(\lambda)\widetilde{P}_{n-|m|}(\vartheta)$$

and, for $m' \neq m$,

$$< S_{nm}(1,\vartheta,\lambda),\ S_{nm'}(1,\vartheta,\lambda) >_{L^2(S_1)}$$
$$= \frac{1}{4\pi}\int_0^{\pi} d\vartheta \sin\vartheta \widetilde{P}_{n-|m|}(\vartheta)\widetilde{P}_{n-|m'|}(\vartheta)\int_0^{2\pi} f_m(\lambda) f_{m'}(\lambda) d\lambda = 0,$$

since the Fourier functions $\{f_m(\lambda)\}$ are well-known to be orthogonal in $L^2([0, 2\pi])$. □

Remark 2. As we have already recalled, there are other functions of ϑ which are used in geodetic literature to form spherical harmonics, namely, with $t = \cos\vartheta$ and $m = 0, 1, \ldots n$,

$$P_{nm}(\vartheta) = (1-t^2)^{m/2} D_t^m P_n(t) \tag{13.70}$$
$$\overline{P}_{nm}(\vartheta) = k_{nm} P_{nm}(\vartheta) \tag{13.71}$$
$$k_{nm} = \sqrt{(2-\delta_{m0})(2n+1)\frac{(n-m)!}{(n+m)!}} \tag{13.72}$$

The $P_{nm}(\vartheta)$ are known as *Legendre associated functions* of the first kind. They were found by studying Laplace equation in spherical coordinates, imposing that

$$\Delta[r^n f_m(\lambda) P_{nm}(\vartheta)] = 0. \tag{13.73}$$

By using formula (13.36) we find for P_{nm} as functions of $t = \cos\vartheta$, the equation

$$(1-t^2)P''_{nm}(t) - 2tP'_{nm}(t) + \left[n(n+1) - \frac{m^2}{1-t^2}\right] P_{nm}(t) = 0 \tag{13.74}$$

which is also known as *Legendre equation*.

It has to be underlined that if we put $m = 0$ in (13.70) we get

$$P_{n0}(t) = P_n(t), \tag{13.75}$$

i.e. the associated Legendre functions of order zero are simply the Legendre polynomials. This agrees with the fact that if we put $m = 0$ into (13.74) we go back to (13.37), i.e. the equation that is satisfied by $P_n(t)$. One can prove that P_{nm}, as given by (13.70), do satisfy (13.74) and then (13.73) too.

On the other hand we see from (13.70) that $P_{nm}(\vartheta)$ is $(\sin\vartheta)^m$ multiplied by a polynomial of degree $(n-m)$ in $\cos\vartheta = t$, i.e. it has exactly the same form of $\widetilde{P}_{nm}(\vartheta)$ (cf. (13.61)). Since by Proposition (13.6) it has been proved that $\widetilde{P}_{nm}(\vartheta)$ are unique, up to a multiplicative constant, we conclude that $\widetilde{P}_{nm}$ are the same as the Legendre functions with a different normalization, i.e.

$$\widetilde{P}_{nm}(\vartheta) = A_{nm} P_{nm}(\vartheta). \tag{13.76}$$

The constant A_{nm} are easy to find by comparing the coefficients of maximum degree in t in $D^m P_n$ and in $Q_{n-m}(\sin\vartheta, \cos\vartheta)$, but they are really not needed here. What is important is that (13.76) holds. Finally, a different normalization is in fact used in all practical computations as well as in theoretical formulas, namely that of the functions $\overline{P}_{nm}(\vartheta)$, also called *normalized Legendre functions*.

The normalization condition of $\overline{P}_{nm}(\vartheta)$ is derived from the request that the surface spherical harmonics

$$Y_{nm}(\vartheta, \lambda) = f_m(\lambda)\overline{P}_{nm}(\vartheta) \tag{13.77}$$

have norm one in $L^2(S_1)$, namely

$$\frac{1}{4\pi}\int Y_{nm}^2(\vartheta, \lambda)d\sigma = 1. \tag{13.78}$$

If we use the relations (remember the definition (13.52) of $f_m(\lambda)$)

$$\int_0^{2\pi} f_m^2(\lambda)d\lambda = (1 + \delta_{m0})\pi$$

into (13.78), i.e.

$$\begin{aligned}\frac{1}{4\pi}\int_0^{2\pi} d\lambda f_m^2(\lambda) \cdot \int_0^{\pi} d\vartheta \sin\vartheta \overline{P}_{nm}(\vartheta)^2 \\ = \frac{1 + \delta_{m0}}{4}\int_{-1}^{1} \overline{P}_{nm}(t)^2 dt = 1,\end{aligned}$$

we see that the constants k_{nm} have to be computed from

$$\frac{1 + \delta_{m0}}{4} \cdot k_{nm}^2 \int_{-1}^{1} P_{nm}^2(t)dt = 1, \tag{13.79}$$

i.e.

$$m \neq 0 \quad k_{nm} = 2\left\{\int_{-1}^{1} P_{nm}^2(t)dt\right\}^{-1/2} \tag{13.80}$$

$$m = 0 \quad k_{n0} = \sqrt{2}\left\{\int_{-1}^{1} P_{n0}^2(t)dt\right\}^{-1/2}. \tag{13.81}$$

In particular we have

$$k_{n0} = \frac{1}{\sqrt{2n+1}}, \quad \overline{P}_{n0}(t) = \sqrt{2n+1}P_n(t). \tag{13.82}$$

We conclude that the spherical harmonics, $\{Y_{nm}(\vartheta, \lambda\}$, here precisely defined are the same functions anticipated by the definition (13.52) and therefore they satisfy the summation theorem (13.55).

Now that we have set up an explicit construction of a CON system in HH_n^3, namely $\{r^n Y_{nm}(\vartheta, \lambda)\}$, when we endow this space with the $L^2(S_1)$ product, we have to find suitable numerical methods for an efficient computation of the spherical harmonic functions $\{Y_{nm}(\vartheta, \lambda)\} = \overline{P}_{nm}(\vartheta) f_m(\lambda)$, i.e. of the associated Legendre functions $\overline{P}_{nm}(\vartheta)$. This is done by establishing recursive relations, among which two are relatively simple and widely used in practice.

Proposition 7. *The following recursive relation on the degree n for* $\overline{P}_{nm}(t), \overline{P}'_{nm}(t)$ *(as functions of* $t = \cos\vartheta$*) holds*

$$-\begin{vmatrix} \overline{P}_{n+1,m}(t) \\ \overline{P}'_{n+1,m}(t) \end{vmatrix} = A_{nm} \begin{vmatrix} t & 0 \\ 1 & t \end{vmatrix} \begin{vmatrix} \overline{P}_{nm}(t) \\ \overline{P}'_{nm}(t) \end{vmatrix} - B_{nm} \begin{vmatrix} \overline{P}_{n-1,m}(t) \\ \overline{P}'_{n-1,m} \end{vmatrix}, \tag{13.83}$$

where

$$A_{nm} = \left[\frac{(2n+1)(2n+3)}{(n+1-m)(n+1+m)}\right]^{1/2}$$
$$B_{nm} = \left[\frac{(2n+3)(n+m)(n-m)}{(2n-1)(n+1-m)(n+1+m)}\right]^{1/2};$$

for every $m \neq 0$ *such relations can start from*

$$\begin{cases} \overline{P}_{mm}(t) = k_{mm}(1-t^2)^{m/2} = k_{mm}(\sin\vartheta)^m \\ \overline{P}'_{mm}(t) = -k_{mm} m (1-t^2)^{m/2-1} t \end{cases} \tag{13.84}$$

and

$$\overline{P}_{m-1,m} \equiv 0, \ \overline{P}'_{m-1,m} \equiv 0; \tag{13.85}$$

for $m = 0$ *we can start from*

$$\begin{aligned} &\overline{P}_{00} = 1, \ \overline{P}_{10} = \sqrt{3}t \\ &\overline{P}'_{00} = 0, \ \overline{P}'_{10} = \sqrt{3}. \end{aligned} \tag{13.86}$$

Proof. Of the two relations (13.83) we need proving only the first one, as the second is just the derivative with respect to t of the first.

We take the recursive relation (cf. Part I, (3.24)) for $P_n(t)$

$$(n+1)P_{n+1}(t) = (2n+1)tP_n - nP_{n-1}$$

and apply D^m to obtain

$$(n+1)D^m P_{n+1} = (2n+1)tD^m P_n + (2n+1)mD^{m-1}P_n - nD^m P_{n-1}. \tag{13.87}$$

Then we remember that (see Part I, (3.37))

$$P_n = \frac{1}{2n+1}\left[P'_{n+1} - P'_{n-1}\right]$$

so that

$$D^{m-1}P_n = \frac{1}{2n+1}[D^m P_{n+1} - D^m P_{n-1}];$$

substituting back in (13.87) and re-ordering, we get

$$(n+1-m)D^m P_{n+1} = (2n+1)tD^m P_n - (n+m)D^m P_{n-1}. \tag{13.88}$$

Now we multiply (13.88) by $(1-t^2)^{m/2}$ arriving at

$$P_{n+1,m} = \frac{2n+1}{n+1-m}tP_{n,m} - \frac{n+m}{n+1-m}P_{n-1,m}. \tag{13.89}$$

We note here that during the step (13.87) whenever it happens that $n<m$ we can put

$$n<m,\ D^m P_n \equiv 0 \Rightarrow P_{n,m} \equiv 0, \tag{13.90}$$

because P_n is a polynomial of degree n in t; this already justifies (13.85).

Finally in (13.89) we can multiply and divide each P_{nm} by k_{nm} (cf. (13.71) and (13.72)), i.e. we can put

$$P_{\ell,m} = k^{-1}_{\ell,m}\overline{P}_{\ell,m},\ \ell = n+1, n, n-1$$

and simplify, to obtain the first of (13.83). The relation (13.84) is just the definition of $\overline{P}_{nm}$ and its derivative; (13.85) is already justified.

For $m=0$, we never have $n<m$, but we can initialize (13.84) with (13.86), which are again just definitions. □

Proposition 8. *The following recursive relations, on the order m, hold*

$$P_{n,m+1} = \frac{2t}{\sqrt{1-t^2}}mC_{nm}\overline{P}_{nm} - C_{nm}D_{nm}\overline{P}_{n,m-1} \tag{13.91}$$

$$C_{nm} = \left[\frac{1}{(n-m)(n-m+1)}\right]^{1/2},\quad m<n$$

$$D_{nm} = [(n+m)(n-m+1)]^{1/2} \cdot \sqrt{1+\delta_{m1}},$$
$$\overline{P}'_{nm} = \frac{t}{1-t^2} m \overline{P}_{nm} - \frac{D_{nm}}{\sqrt{1-t^2}} \overline{P}_{n,m-1}. \tag{13.92}$$

We note that, although we could indeed put (13.92) in a form where $\overline{P}'_{n,m+1}$ is given as a combination of $\overline{P}'_{nm}$ and $\overline{P}'_{n,m-1}$, such equation which can be computed in sequence after (13.91), has a simpler form which we prefer.

We note also that (13.91) and (13.92) can be triggered by a previous computation of $\overline{P}_{n0}$, for all the degrees needed, and then

$$\begin{cases} \overline{P}_{n1} = \sqrt{\dfrac{2}{n(n+1)}} (1-t^2)^{1/2-1} P_{n0} \\ \overline{P}'_{n1} = \dfrac{t}{1-t^2} \overline{P}_{n1} - \dfrac{\sqrt{2n(n+1)}}{\sqrt{1-t^2}} \overline{P}_{n0}. \end{cases} \tag{13.93}$$

Proof. We start from the notable relation, proved in Exercise 13.9,

$$(1-t^2) P_n^{(m+2)} = 2(m+1) t P_n^{(m+1)} - (n-m)(n+m+1) P_n^{(m)}, \tag{13.94}$$

where

$$P_n^{(m)}(t) = D^m P_n(t).$$

We substitute $(m-1)$ to (m) and multiply it by $(1-t^2)^{m/2}$, to get

$$P_{n,m+1} = \frac{2t}{\sqrt{1-t^2}} m P_{nm} - (n+m)(n-m+1) P_{n,m-1}. \tag{13.95}$$

Substituting

$$\overline{P}_{nj} = k_{nj} P_{nj}, \ j = m+1, m, m-1 \tag{13.96}$$

in (13.95) and simplifying, we get (13.91). Now we go back to the definition

$$P_{nm} = (1-t^2)^{m/2} D^m P_n$$

and differentiate, obtaining

$$P'_{nm} = -m \frac{t}{1-t^2} P_{nm} + \frac{1}{\sqrt{1-t^2}} P_{n,m+1}. \tag{13.97}$$

Using (13.95) in (13.97) and normalizing with (13.96) we find (13.92).

Finally the first of (13.93) is just the definition of $\overline{P}_{n1}$, while the second is (13.92) with $m = 1$. □

Remark 3. Whatever recursive relations are used to compute

$$\overline{P}_{nm}(t), \overline{P}'_{nm}(t) = -\frac{1}{\sin\vartheta}\frac{\partial}{\partial\vartheta}\overline{P}_{nm}(\vartheta),$$

we can then easily compute the second derivative $\overline{P''}_{nm}(t)$ by exploiting the equation (13.74) suitably normalized, i.e.

$$\overline{P}''_{nm} = \frac{2t}{1-t^2}\overline{P}'_{nm} - \frac{1}{1-t^2}\left[n(n+1) - \frac{m^2}{1-t^2}\right]\overline{P}_{nm}. \tag{13.98}$$

Remark 4. It is possible to see that for $m \neq 0$, $\overline{P}_{nm}(t) \to 0$, when $n \to \infty$. Therefore when $Y_{nm}(\vartheta, \lambda)$ are used with sums up to very high degree and order, for instance several thousands, the relative error in the calculus of such harmonics can increase significantly, and recursive relations on the degree n, starting from $\overline{P}_{mm}(t)$, are not any more providing reliable results, specially when $\overline{P}_{mm}$ itself is already very small.

So if one has to compute a function like

$$f(\vartheta, \lambda) = \sum_{n=0}^{N}\sum_{m=-n}^{n} f_{nm}Y_{nm}(\vartheta, \lambda) \tag{13.99}$$

for N equal to several thousands, one has to use (13.91) and (13.92), which for low orders give a good approximation and when they start giving bad results one can truncate the summation, because the functions $\overline{P}_{nm}(t)$ are in any case so small that they contribute little to sums like (13.99).

Among others, this is also the reason why we start (13.91) at $m = 0$ instead of $m = n$.

Alternatively one can still use the recursion on the degree n, however one starts from a $\widetilde{P}_{nm} = H_{nm}P_{nm}$ suitably re-normalized, and in the end one divides again the result by H_{nm}. This simple trick allows the accurate computation of P_{nm} for all degrees and orders up to some thousands.

A longer number of interesting relations like those presented above are known in literature, including different forms of the summation theorem (Martinec 1998).

13.4 Hilbert Spaces of Harmonic Functions and First Theorems of Potential Theory

It is now time to abandon the use of simple harmonic polynomials, which implies working only in finite dimensional spaces, and rather go to HS of harmonic functions, namely to transform sums into series and to study limit properties when the dimension of the space goes to infinity.

The material of this section is covered by several books in geodesy, like Moritz (1980), Krarup (2006), and Heiskanen and Moritz (1967). A recent simple mathematical book on the subject is Axler et al. (2001).

The first result to be presented is so important that we state it in the form of a Theorem.

Theorem 3. *The sequence of normalized spherical harmonics*

$$\{Y_{nm}(\vartheta,\lambda);\ |m|\le n,\ n=0,1,2\ldots\} \tag{13.100}$$

is a CON system in $L^2(S_1)$ *that is* $\forall f(\vartheta,\lambda)\in L^2(S_1)$ *we have*

$$\begin{cases} f(\vartheta,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=-n}^{n} f_{nm}Y_{nm}(\vartheta,\lambda) \\ f_{nm}=\frac{1}{4\pi}\int_{S_1} f(\vartheta,\lambda)Y_{nm}(\vartheta,\lambda)d\sigma \end{cases}, \tag{13.101}$$

the series being convergent in $L^2(S_1)$*; furthermore*

$$\|f\|^2_{L^2(S_1)}=\frac{1}{4\pi}\int_{S_1} f^2(\vartheta,\lambda)d\sigma=\sum_{n=0}^{+\infty}\sum_{m=-n}^{n} f_{nm}^2. \tag{13.102}$$

Proof. That $\{Y_{nm}\}$ is an orthonormal system in $L^2(S_1)$ we already know; we have to prove that it is complete.

We could just invoke Proposition 12 here, but to prepare further results (theorem 7) we prefer to prove directly that $\{Y_{nm}\}$ is total in $L^2(S_1)$ and then use Proposition 9. So we need to prove that

$$\forall f\in C^1(S_1);\quad <f,Y_{nm}>_{L^2(S_1)}=0,\forall n,m\Rightarrow f=0. \tag{13.103}$$

First note that (13.103) implies

$$\forall n,\ \forall P\in S_1\quad <f(Q),P_n(\psi_{PQ})>_{L^2(S_1)}\equiv 0, \tag{13.104}$$

because

$$<f(Q),P_n(\psi_{PQ})>_{L^2(S_1)}=\sum_{m=-n}^{n}(2n+1)^{-1}Y_{nm}(P)<f,Y_{nm}>_{L^2(S_1)}. \tag{13.105}$$

Now consider the single-layer potential

$$V(P)=\frac{1}{4\pi}\int\frac{f(Q)}{\ell_{PQ}}d\sigma_Q=\frac{1}{4\pi}\int_{S_1}\frac{f(Q)}{\sqrt{1+r_P^2-2r_P\cos\psi_{PQ}}}d\sigma_Q; \tag{13.106}$$

if we take $r_P < 1$ or $r_P > 1$ we have, respectively

$$r_P < 1, \quad \frac{1}{\ell_{PQ}} = \sum_{n=0}^{+\infty} r_P^n P_n(\cos\psi_{PQ}), \tag{13.107}$$

$$r_P > 1, \quad \frac{1}{\ell_{PQ}} = \sum_{n=0}^{+\infty} \frac{1}{r_P^{n+1}} P_n(\cos\psi_{PQ}). \tag{13.108}$$

The two series (13.107) and (13.108) converge uniformly in ψ_{PQ} because

$$|P_n(\cos\psi)| \leq 1, \ \forall\psi$$

so that we can substitute them in (13.106) and exchange summation and integral to find

$$r_P < 1, \quad V(P) = \sum_{n=0}^{+\infty} r_P^n \left\{ \frac{1}{4\pi} \int f(Q) P_n(\cos\psi_{PQ}) d\sigma_Q \right\} = 0,$$

$$r_P > 1, \quad V(P) = \sum_{n=0}^{+\infty} \frac{1}{r_P^n} \left\{ \frac{1}{4\pi} \int f(Q) P_n(\cos\psi_{PQ}) d\sigma_Q \right\} = 0.$$

In other words, since the potential of an L^2 single layer admits almost everywhere on S, radial limits, (cf. Miranda 1970), we find

$$V(P) \equiv 0 \tag{13.109}$$

everywhere in $\mathbf{R}^3$.

On the other hand remember that the following jump relations for the normal derivatives, taken across S_1, hold (cf. Part I, (1.54))

$$f(P) = -\frac{1}{4\pi} \left\{ \left(\frac{\partial V}{\partial \nu} \right)_+ - \left(\frac{\partial V}{\partial \nu} \right)_- \right\} \tag{13.110}$$

so that we find, because of (13.109), that it is also

$$f(P) \equiv 0, \tag{13.111}$$

as it was to be proved.

The relation (13.102) is just Parseval's identity for this specific case. □

Example 1. The following is the Dirichlet problem for a ball B_R of radius R. Let a function $f(\vartheta, \lambda)$ be given on S_R, the boundary of B_R, and for the sake of

definiteness we assume $f \in L^2(S_R)$; we want to find a $h(r, \vartheta, \lambda)$ which is harmonic in B_R and agrees, in a suitable sense, to be here defined, with f on S_R

$$u(R, \vartheta, \lambda) = f(\vartheta, \lambda). \tag{13.112}$$

Let us state the convention that, at the level of notation, when we represent a series of spherical harmonics without specifying the summation limits, we implicitly mean that we add over all degrees and orders, namely

$$\sum_{n,m} f_{nm} Y_{nm}(\vartheta, \lambda) \equiv \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} f_{nm} Y_{nm}(\vartheta, \lambda).$$

Because of Theorem 3 we know that we can put

$$P \in S_R, \ f(P) = \sum_{n,m} < f(Q), Y_{nm}(Q) > Y_{nm}(P) = \sum_{n,m} f_{nm} Y_{nm}(P);$$

since for each degree and order we know that

$$f_{nm} S_{nm}(r, \vartheta, \lambda) = f_{nm} \left(\frac{r}{R}\right)^n Y_{nm}(\vartheta, \lambda),$$

is indeed harmonic and agrees with

$$f_{nm} Y_{nm}(\vartheta, \lambda) \text{ on } S_R,$$

we guess that the sought solution is given by

$$u(r, \vartheta, \lambda) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} f_{nm} \left(\frac{r}{R}\right)^n Y_{nm}(\vartheta, \lambda). \tag{13.113}$$

The problem is whether the series is convergent and we are allowed to apply the Laplace operator term-wise, so that from $\Delta \left[\left(\frac{r}{R}\right)^n Y_{nm}(\vartheta, \lambda) \right] = 0$ we can deduce $\Delta u(r, \vartheta, \lambda) = 0$. Remember that if we put $\vartheta = \vartheta', \lambda = \lambda'$ in (13.55), i.e. $\xi = \eta$ and $\psi_{\xi\eta} = 0$, we have

$$\frac{1}{2n+1} \sum_{m=-n}^{n} Y_{nm}^2(\vartheta, \lambda) = P_n(1) = 1. \tag{13.114}$$

This implies that (see also Martinec 1998)

$$|Y_{nm}(\vartheta, \lambda)| \leq \sqrt{2n+1}. \tag{13.115}$$

Then, from (13.113), by using Schwarz inequality, we have

$$|u(r,\vartheta,\lambda)| \le \sum_{n=0}^{+\infty} \left(\frac{r}{R}\right)^n \left\{ \sum_m f_{nm}^2 \cdot \sum_m Y_{nm}^2 \right\}^{1/2}$$
$$\le \sum_{n=0}^{+\infty} \sqrt{2n+1} \left(\frac{r}{R}\right)^n \cdot \sqrt{\sum_m \cdot f_{nm}^2}$$
$$\le A \cdot \sum_{n=0}^{+\infty} \sqrt{2n+1} \left(\frac{r}{R}\right)^n ; \tag{13.116}$$

the last step in (13.116) is justified because from (13.102) we know that $\sum_{m=-n}^{n} f_{nm}^2 \to 0$, so that there must be a constant A such that

$$\forall n \ge 0, \quad \sqrt{\sum_{m=-n}^{n} f_{nm}^2} \le A.$$

The relation (13.116) shows that our series is absolutely and uniformly convergent in every ball strictly contained in B_R and concentric with that. Since the multiplication of $\left(\frac{r}{R}\right)^n$ by any (fixed) polynomial in n does not modify the convergence of (13.116), we deduce that we can apply termwise such operators as $\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}$ and $r^{-2}\Delta_\sigma$ and verify that in fact (13.113) is harmonic.

As for the Dirichlet boundary condition (13.112) contrary to intuition, it is not enough to put $r = R$ in (13.113) and then verify that we are left with the harmonic development of $f(P)$, because this series is not really convergent in a pointwise sense but only in $L^2(S_1)$, i.e. in a mean square sense over S_1.

The correct definition is as follows: we take the trace of $u(r,\vartheta,\lambda)$ at any sphere with $r = R - \delta$ and we take the difference

$$f(\vartheta,\lambda) - u(R-\delta,\vartheta,\lambda) = \sum_{n,m} f_{nm} \left[1 - \left(\frac{R-\delta}{R} \right)^n \right] Y_{nm}(\vartheta,\lambda);$$

we evaluate the $L^2(S_1)$ norm of such difference, namely

$$\| f(\vartheta,\lambda) - u(R-\delta,\vartheta,\lambda) \|_{L^2(S_1)}^2 = \sum_{n,m} f_{nm}^2 \left[1 - \left(1 - \frac{\delta}{R} \right)^n \right]^2 . \tag{13.117}$$

Since each term of the positive series (13.114) is bounded above by f_{nm}^2, because indeed

$$\left[1 - \left(1 - \frac{\delta}{R} \right)^n \right]^2 \le 1, \ n = 0, 1, \ldots$$

we can pass to the limit for $\delta \to 0$ under the series and find

$$\lim_{\delta \to 0} \| f(\vartheta, \lambda) - u(R - \delta, \vartheta, \lambda) \|^2_{L^2(S_1)} = 0. \tag{13.118}$$

Remark 5. This interpretation of the Dirichlet boundary condition has been introduced by Cimmino (1952) and it has been taken up in geodesy in a number of works on BVP (Cimmino 1955; Sansò and Venuti 1998). As a matter of fact, the theory holds in general, with suitable changes, for smooth surfaces, S, and for functions $f(P)$ square integrable over S, as we shall soon see.

Example 2 (Poisson integral). In this example we continue Example 1, giving to the solution of Dirichlet problem for the sphere (13.113) the form of a Poisson integral.

In fact, substituting

$$f_{nm} = \frac{1}{4\pi} \int f(\vartheta', \lambda') Y_{nm}(\vartheta', \lambda') d\sigma'$$

into (13.113) and recalling the convergence result of such series, we can claim that

$$\begin{aligned} \forall r < 1, \quad u(r, \vartheta, \lambda) = & \qquad (13.119) \\ &= \frac{1}{4\pi} \int \left\{ \sum_{n,m} Y_{nm}(\vartheta, \lambda) \left(\frac{r}{R} \right)^n Y_{nm}(\vartheta' \lambda') \right\} f(\vartheta', \lambda') d\sigma' \\ &= \frac{1}{4\pi} \int \left\{ \sum_{n=0}^{+\infty} \left(\frac{r}{R} \right)^n (2n+1) P_n(\cos \psi) \right\} f(\vartheta', \lambda') d\sigma', \end{aligned}$$

ψ being the spherical angle between the direction of $P(\vartheta, \lambda)$ and $P'(\vartheta', \lambda')$ on the unit sphere.

Now recall that

$$G(s,t) = \frac{1}{\{1 + s^2 - 2st\}^{1/2}} = \sum_{n=0}^{+\infty} s^n P_n(t)$$

and observe that

$$2s \frac{\partial}{\partial s} G(s,t) + G(s,t) = \sum_{n=0}^{+\infty} (2n+1) s^n P_n(t). \tag{13.120}$$

Using (13.120) in (13.119), with $s = \frac{r}{R}$ and $t = \cos \psi$, one receives, after re-arranging,

$$\begin{cases} u(r, \vartheta, \lambda) = \dfrac{1}{4\pi} \displaystyle\int \Pi_{Ri}(r, \psi) f(\vartheta', \lambda') d\sigma \\ \Pi_{Ri}(r, \psi) = \dfrac{R(R^2 - r^2)}{[R^2 + r^2 - 2rR\cos\psi]^{3/2}}. \end{cases} \tag{13.121}$$

The function $\Pi_{Ri}(r,\psi)$ is the so-called *internal Poisson kernel* for the ball of radius R. We note that Π_{Ri} is a function of the point $P \equiv (r,\vartheta,\lambda)$ and $P' \equiv (R,\vartheta',\lambda')$ in the sense that

$$\Pi_{Ri}(P,P') = \frac{R(R^2 - |\mathbf{r}_P|^2)}{|\mathbf{r}_{P'} - \mathbf{r}_P|^3}. \tag{13.122}$$

With the notation of (13.122) and observing that $R^2 d\sigma \equiv dS$, the area element of the sphere of radius R, we can re-write (13.121) as

$$u(P) = \frac{1}{4\pi R^2} \int_S \Pi_{Ri}(P,P') f(P') dS_{P'}. \tag{13.123}$$

With the aid of (13.123) it becomes quite evident that $u(P)$, inside any sphere of radius $R' < R$, is in fact continuous with all its derivatives, i.e. $u \in C^\infty(B_R)$, because the Poisson kernel enjoys the same property.

We conclude with some properties of the Poisson kernel. Since, with $f(P') \equiv 1$ we are to find $u(P) \equiv 1$ in (13.123), we see that

$$\frac{1}{4\pi R^2} \int \Pi_{Ri}(P,P') dS_{P'} \equiv 1;$$

moreover $\prod_{Ri}(P,P') > 0$ when $r < R$ so that

$$\int \left|\Pi_{Ri}(P,P')\right| dS_{P'} < \text{const},$$

in fact it is equal to 1; furthermore, taking once $\psi \neq 0$ and then $\psi = 0$ in (13.121) we find

$$\psi \neq 0 \quad \lim_{r_P \to R} \Pi_{Ri}(P,P') = 0$$

$$\psi = 0 \quad \lim_{r_P \to R} \Pi_{Ri}(P,P') = +\infty.$$

The above four properties are enough to guarantee that, $\forall \varphi(P') \in C(S_R)$ and $P_0 \in S_R$

$$\lim_{\substack{P \to P_0 \\ (r_P \to R)}} \frac{1}{4\pi R^2} \int \Pi_{Ri}(P,P') \varphi(P') dS_{P'} = \varphi(P_0). \tag{13.124}$$

Proposition 9 (Mean value property). *Let $u(P)$ be harmonic in a domain B and let $B_R(P_0)$ be a ball of centre P_0 and radius R such that $B_R(P_0) \subset B$; put*

$$M_{B_{P_O R}}\{u\} = \frac{1}{\frac{4}{3}\pi R^3}\int_{B_{P_0 R}} u(Q)dB \tag{13.125}$$

$$M_{S_{P_O R}}\{u\} = \frac{1}{4\pi R^2}\int_{S_{P_0 R}} u(Q)dS; \tag{13.126}$$

i.e. the mean value of $u(P)$ *over* $B_{P_0 R}$ *or over* $S_{P_0 R}$ *respectively; then we have*

$$M_{B_{P_0 R}}\{u\} = M_{S_{P_0 R}}\{u\} = u(P_0). \tag{13.127}$$

Proof. $u(P)$ has to be continuous on $S_{P_0 R}$ because this surface is contained in B; then $u(P)|_{S_{P_0 R}}$ is also in $L^2(S_{P_0 R})$ and we can apply (13.122) and (13.123) with $P = P_0, |\mathbf{r}_P| = |\mathbf{r}_{P_0}| = 0,\ |\mathbf{r}_{P'} - \mathbf{r}_{P_0}| = R$ so that $\prod_{Ri}(P_0, P') \equiv 1$; the result is the second equality of (13.127). Moreover if we multiply both members of (13.126) by $R^2 dR$ and integrate, we find

$$\int_0^{\overline{R}} R^2 dR \cdot u(P_0) = \frac{1}{4\pi}\int_0^{\overline{R}}\int_{S_{P_0 R}} u(Q)dS dR = \frac{1}{4\pi}\int_{B_{P_0 \overline{R}}} u(Q)dB; \tag{13.128}$$

noting that $\int_0^{\overline{R}} R^2 dR = \frac{1}{3}\overline{R}^3$ and dividing both members of (13.128) by such quantity, we get the first equality of (13.127). □

Proposition 10. *This is the inverse of Proposition 9; namely, let* $u(P)$ *be defined and continuous up to second derivatives in* B*; assume further that (13.127) holds for* $u(P)$, $\forall R$ *such that* $B_{P_0 R} \subset B$*, then* $u(P)$ *is harmonic in* B

$$\Delta u(P) = 0 \ \text{ in } B. \tag{13.129}$$

Proof. Fix P_0 and let $\boldsymbol{\xi}$ be a vector of constant length R, such that $P_0 + \boldsymbol{\xi} \subset B$; we can write

$$u(P_0 + \boldsymbol{\xi}) = u(P_0) + \boldsymbol{\xi} \cdot \nabla u(P_0) + \frac{1}{2}\boldsymbol{\xi}^t D^2 u(P_0)\boldsymbol{\xi} + o_2(\boldsymbol{\xi}), \tag{13.130}$$

where we have put

$$D^2 u(P_0) = \left\{ \frac{\partial^2}{\partial\xi_i \partial\xi_k} u(P_0 + \boldsymbol{\xi})\Big|_{\boldsymbol{\xi}=0} \right\}.$$

Note that, by direct computation, one has

$$M_{S_{P_0 R}}\{\boldsymbol{\xi}\} = 0,\ M_{S_{P_0 R}}\{\boldsymbol{\xi}\ \boldsymbol{\xi}^t\} = \frac{R^2}{3} I. \tag{13.131}$$

Take $M_{S_{P_0R}}$ of both members of (13.130) and use (13.127) to find

$$u(P_0) = u(P_0) + \frac{1}{2} TrD^2 u(P_0) M_{S_{P_0R}}\{\boldsymbol{\xi}\boldsymbol{\xi}^t\} + o_2(R),$$

i.e., using (13.131),

$$\frac{R^2}{6} TrD^2 u(P_0) + o_2(R) = 0.$$

Dividing by R^2 and letting $R \to 0$ we see then that

$$TrD^2 u(P_0) = \Delta u(P_0) = 0. \qquad \square$$

Theorem 4 (Maximum principle). *Let $\{u(P)\}$ be harmonic in B and continuous up to the boundary, then, unless $u(P)$ is constant everywhere in B,*

$$\forall P \in B, \quad \min_{Q \in S} u(Q) < u(P) < \max_{Q \in S} u(Q). \tag{13.132}$$

Proof. Since it is a continuous function on the bounded closed set $\overline{B} = B \cup S$, $u(P)$ attains a minimum and a maximum value in such set. Let $\overline{P}$ be a point of absolute maximum; we prove that either $\overline{P} \in S$ or $u(P)$ is constant. In fact if $\overline{P} \in B$, then it is also a relative maximum so that, taking a suitable ball $B_{\overline{P},R}$, one has to find

$$u(\overline{P}) = M_{B_{\overline{P}R}}\{u(P)\} \leq u(\overline{P}); \tag{13.133}$$

equality in (13.133) can be achieved only if $u(P) \equiv u(\overline{P})$, $\forall P \in B_{\overline{P},R}$, i.e. if $u(P)$ is constant in $B_{\overline{P},R}$. By the principle of identity of harmonic functions that will soon be proved, we should then have $u(P) = \text{const}$ everywhere in B. Otherwise (13.133) becomes impossible, i.e $\overline{P} \in S$. The same reasoning holds for the minimum. $\square$

Corollary 2. *Also the derivatives of $\{u(P)\}$ are controlled by their extreme values on the boundary, on condition that we stay away from S with P. More precisely, let K be any bounded closed set such that $K \subset B$ and let δ be the distance between K and S*

$$\delta = \min_{\substack{P \in S \\ Q \in K}} |\mathbf{r}_P - \mathbf{r}_Q|;$$

then there is a constant A such that

$$\max_{P \in K} \left| \frac{\partial u}{\partial x_i}(P) \right| \leq A \cdot \delta^{-1} \max_{P \in S} |u(P)|.$$

Proof. First note that (13.127) implies

$$\max_{P\in B}|u(P)| \le \max_{P\in S}|u(P)|.$$

Then, since $\frac{\partial u}{\partial x_i}$ is a harmonic function too, write the mean value property for any $P_0 \in K$. As, it is (denoting with $\mathbf{e}_i$ the unit vector in the direction of the i-th axis)

$$\frac{\partial u}{\partial x_i}(P_0) = \frac{1}{\frac{4}{3}\pi R^3}\int_{B_{P_0R}} \frac{\partial u}{\partial x_i} dB = \frac{1}{\frac{4}{3}\pi R^3}\left\{\int_{S_{P_0R}} \mathbf{e}_i \cdot \mathbf{n} u dS\right\},$$

then

$$P_0 \in K, \quad \left|\frac{\partial u}{\partial x_i}(P_0)\right| \le \frac{3}{R}\,\frac{1}{4\pi R^2}\int_{S_{P_0R}} |u(Q)|dS \le \frac{3}{R}\max_{P\in S}|u(P)|. \quad (13.134)$$

Now we note that once K is fixed R can always be extended to became $R = \delta$, so that (13.134) implies

$$\max_{P_0\in K}\left|\frac{\partial u}{\partial x_i}(P_0)\right| \le 3\delta^{-1}\max_{P\in S}|u(P)|,$$

as it was to be proved. □

Remark 6. We can observe that the argument on extremal values in B holds even if $u(P)$ is not continuous up to boundary. For instance if $u(P)$ is only bounded on S, we can establish (13.132) in the weaker form

$$\sup_{P\in B}|u(P)| \le \sup_{P\in S}|u(P)|. \quad (13.135)$$

We notice too that (13.135) guarantees that the Dirichlet problem for functions harmonic in B and bounded in $\overline{B}$ has a unique solution. Furthermore if $\{u_n(P)\}$ is harmonic in P and uniformly convergent to $f(P)$ on S then $\{u_n(P)\}$ converges uniformly to some function $u(P)$ in $\overline{B}$ such that $u(P)|_S \equiv f(P)$ and even more $u(P)$ is harmonic too. In fact, going to the limit in

$$\lim_{n\to\infty} u_n(P_0) = u(P_0) = \lim_{n\to\infty} M_{B_{P_0R}}\{u_n\} = M_{B_{P_0R}}\{u\},$$

we see that $u(P)$ has to satisfy the mean value property and then it is harmonic by dint of Proposition 10.

As a matter of fact, all that means that if we take a space of harmonic continuous functions in $\overline{B}$ with norm $\|u\|_{C(\overline{B})}$ and the corresponding space of continuous functions on S which are traces on S of the former, $f(P) \equiv u(P)|_S$, then the two spaces are in a one-to-one correspondence one to the other, and in addition this correspondence is isometric, i.e.

$$\|f\|_{C(S)} = \|u\|_{C(\overline{B})}.$$

That such correspondence is onto, i.e. that

$$\{f;\ f = u|_S, \Delta u = 0 \text{ in } B,\ u \in C(\overline{B})\} \equiv C(S),$$

when S is smooth, as we have assumed, is basically a theorem that is classical in potential theory (cf. Kellog 1953; Miranda 1970).

We state it here, without proof, in a form that will be used in the sequel.

Theorem 5. *When S is a closed smooth surface (for instance with a normal field $\mathbf{n}_P$ such that $|\mathbf{n}_P - \mathbf{n}_Q| \leq c \cdot \overline{PQ}$, i.e. it is Lipschitz continuous), the traces of harmonic functions $u(P) \in C(\overline{B})$ cover the whole $C(S)$, so that the problem of Dirichlet*

$$\begin{cases} \Delta u = 0 \text{ in } B \\ u|_S = f \text{ given on } S \end{cases}$$

has one and only one solution, for every $f \in C(S)$.

In addition if $f(P)$ has λ-Hölder continuous derivatives along the boundary, i.e., with ∇_t denoting the tangential component of the gradient,

$$|\nabla_t f(P) - \nabla_t f(Q)| \leq c \cdot (\overline{PQ})^\lambda,\ \forall PQ \in S$$

then also $u(P)$ has λ-Hölder continuous derivatives in $\overline{B}$, i.e.

$$|\nabla u(P)| \leq c', \quad |\nabla u(P) - \nabla u(Q)| \leq c''(\overline{PQ})^\lambda,\ \forall P, Q \in \overline{B},$$

for suitable constants c' and c''. In particular the normal derivative of $u(P)$ on S is continuous and therefore bounded.

Definition 4. Following (Krarup 2006), we define the radius of convergence of a series of spherical harmonics, $\sum_{n,m} u_{nm} r^n Y_{nm}(\vartheta, \lambda)$, as

$$R_c = \sup\left\{r;\ \Sigma u_{nm}^2 r^{2n} < +\infty\right\}. \tag{13.136}$$

We note that indeed if we take any R, such that,

$$R < R_c, \tag{13.137}$$

then the series results to be convergent inside the ball with boundary S_R, i.e. for $r < R$.

In fact

$$\left|\sum_{n,m} u_{nm} r^n Y_{nm}(\vartheta, \lambda)\right|^2 \leq \left(\sum_{n,m} u_{nm}^2 R^{2n}\right)\left(\sum_n \left(\frac{r}{R}\right)^{2n} (2n+1)\right)$$

and the first series is convergent in force of definition (13.136) and condition (13.137), while the second one is convergent $\forall r < R$.

Proposition 11. *Let $u(P)$ be harmonic in B, take any P_0 and let δ_{P_0} be its distance from the boundary S; then $u(P)$ can be developed into a series of spherical harmonics with convergence radius*

$$R_c \geq \delta_{P_0}. \tag{13.138}$$

Proof. Indeed let $R < \delta_{P_0}$; then we can write in $B_{P_0 R}$

$$u(P) = \frac{1}{4\pi R^2} \int_{S_{P_0 R}} \Pi_{Ri}(P, P') u(P') dS_{P'}$$

$$= \sum_{n,m} u_{nm}(P_0, R) \left(\frac{r_P}{R}\right)^n Y_{nm}(\vartheta_P, \lambda_P)$$

with

$$\sum_{n,m} u_{nm}(P_0, R)^2 = \frac{1}{4\pi R^2} \int_{S_{P_0 R}} u^2(P') dS$$

which is finite because on $S(P_0, R), u(P)$ is continuous.

Since this is true $\forall R < \delta_{P_0}$ we have at least $R_c = \delta_{P_0}$; i.e. (13.138) is proved. □

Note that it can very well happen that $R_c > \delta_{P_0}$ and this means that $u(P)$ can be extended as a harmonic function to a region larger then B.

Theorem 6 (Principle of identity of harmonic functions). *Remember that by hypothesis B is a bounded simply (arcwise) connected set with "smooth" boundary. Let $u(P), v(P)$ be two functions harmonic in B and P_0 a point in B; if, for some $R_0 > 0$,*

$$u(P) \equiv v(P), \ P \in B_{P_0 R_0}, \tag{13.139}$$

i.e. the two functions coincide in a neighborhood of P_0, then

$$u(P) \equiv v(P), \quad \forall P \in B. \tag{13.140}$$

Proof. Note that taking $u - v$ instead of u and 0 instead of v, we have to prove that

$$\{\exists R_0; u(P) = 0, \ \forall P \in B_{P_0 R_0}\} \Rightarrow \{u(P) \equiv 0, \ \forall P \in B\}. \tag{13.141}$$

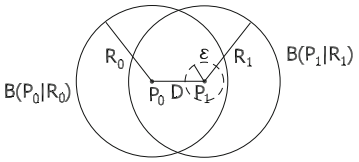


Fig. 13.3 Extension of $u(P)$ from $B_{P_0R_0}$ to $B_{P_1R_1}$

First we show that if $u(P) \equiv 0$ in $B_{P_0R_0}$ and if we take any P_1 with $\overline{P_0P_1} = D < R_0$ and $u(P)$ is harmonic in $B_{P_1R_1}$ with $R_1 > R_0 - D$ (see Fig. 13.3), then

$$u(P) \equiv 0, \quad P \in B_{P_1R_1}. \tag{13.142}$$

In fact under the above hypothesis we can put in $B_{P_1R_1}$, thanks to Proposition 11,

$$u(P) = \sum_{n,m} u_{n,m}(P_1, R_1) \left(\frac{r}{R_1}\right)^n Y_{nm}(\vartheta, \lambda); \tag{13.143}$$

On the other hand, since $D < R_0$, we have for a sufficiently small $\varepsilon > 0$, (cf. Fig. 13.3)

$$B_{P_1\varepsilon} \subset B_{P_0R_0}.$$

But then we can write

$$\forall n, m,\ u_{nm}(P_1, R_1) \left(\frac{\varepsilon}{R_1}\right)^n = \frac{1}{4\pi\varepsilon^2} \int_{S_\varepsilon} Y_{nm}(P')u(P')dS_{P'} = 0,$$

because $u(P')$ is identically zero in $B_{P_0R_0}$.

This implies $u_{nm}(P_1, R_1) = 0,\ \forall(n, m)$, and then (13.141).

Note that if we take any two spheres B_0, B_1 partially overlapping, one can always find a third sphere B' which is in the same position as discussed above, with respect to each of them (see Fig. 13.4). This implies that

$$u(P) \equiv 0 \text{ in } B_0 \Rightarrow u(P) \equiv 0 \text{ in } B' \Rightarrow u(P) \equiv 0 \text{ in } B_1.$$

Now take any $\overline{P} \in B$; according to our hypothesis we have an arc of finite length $L_{P_0\overline{P}}$, joining P_0 to $\overline{P}$ such that

$$L_{P_0\overline{P}} \in B;$$

let then

$$\begin{aligned} \delta &= \text{dist}\,(L_{P_0\overline{P}}, S) \\ &= \inf_{\substack{P \in L_{P_0\overline{P}} \\ Q \in S}} |\mathbf{r}_P - \mathbf{r}_Q| > 0. \end{aligned}$$

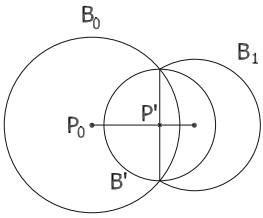


Fig. 13.4 Propagation of $u(P) \equiv 0$ from B_0 to B' to B_1. We may conclude then that if $u(P) \equiv 0$ in B_0, then the same happens in B' and also in B_1

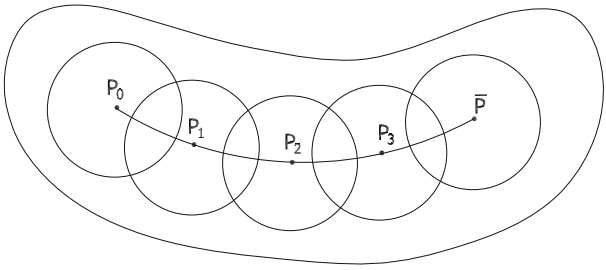


Fig. 13.5 Note that $B_{P_k R} \subset B$ by the choice of R

It is now obvious that we can join P_0 to $\overline{P}$ with a finite number of spheres $B_{P_k \delta}$ with

$$P_k \in L_{P_0 \overline{P}}, \ B_{P_k \delta} \subset B$$

and each $B_{P_k \delta}$ partially overlapping with $B_{P_{k-1} \delta}$ and $B_{P_{k+1} \delta}$ (cf. Fig. 13.5). By using the above argument then we see that $u(P)$ is necessarily zero in each $B_{P_k \delta}$ and then also in $\overline{P}$. □

We are now ready to prove one of the fundamental results of this chapter.

Theorem 7. *Let us consider the traces of solid spherical harmonics* $S_{nm}(P) = S_{nm}(r, \vartheta, \lambda) = r^n Y_{nm}(\vartheta, \lambda)$ *on the boundary* S, $S_{nm}(Q)|_S$. *This system of functions is complete in* $L^2(S)$, *i.e. if* $f \in L^2(S)$ *and*

$$\int_S f(Q) r_Q^n Y_{nm}(\vartheta_Q, \lambda_Q) dS_Q = 0, \ \forall n, m \tag{13.144}$$

then $f = 0$ *a.e. on* S.

Proof. We go for a direct proof, showing that $(S_{nm}|_S)$ is total in $L^2(S)$ and then recall Proposition 9.

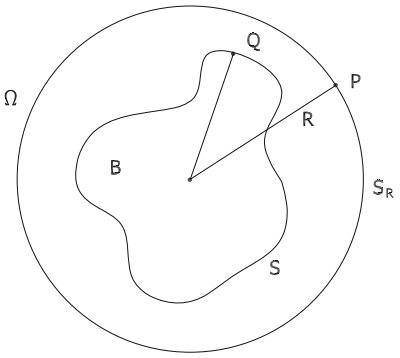


Fig. 13.6 Note that $Q \in S, r_Q < r_P = R$

Consider the single layer potential

$$V(P) = \int_S \frac{f(Q)}{\ell_{PQ}} dS_Q \tag{13.145}$$

and note that $V(P)$ is continuous everywhere in $R^3 \backslash S$ and harmonic in both B and Ω (see Fig. 13.6). Now take R so that $S_R \subset \Omega$ and take $Q \in S, P \in S_R$; we have then

$$\begin{aligned} \frac{1}{\ell_{PQ}} &= \sum_n \frac{r_Q^n}{R^{n+1}} P_n(\cos \psi_{PQ}) \\ &= \sum_{n,m} \frac{r_Q^n}{R^{n+1}} \frac{Y_{nm}(\vartheta_P, \lambda_P) Y_{nm}(\vartheta_Q, \lambda_Q)}{2n+1} \end{aligned}$$

and the series converges uniformly on S_R (Fig. 13.6). But then

$$P \in S_R,\ V(P) = \sum_{n,m} \frac{Y_{nm}(\vartheta_P, \lambda_P)}{(2n+1)R^{n+1}} \int_S f(Q) r_Q^n Y_{nm}(\vartheta_Q, \lambda_Q) dS \equiv 0.$$

The same reasoning indeed holds for any sphere outside S_R; in other words $V(P)$ is identically zero outside S_R, but then, by dint of Theorem 6, $V(P) \equiv 0$ in Ω. Since $V(P)$ has limits almost everywhere on S along the normal (see Miranda 1970, Chap. II Sect. 14), it has to be too

$$V(P) \equiv 0,\ P \in S. \tag{13.146}$$

On the other hand $V(P)$ is continuous and harmonic in B too and therefore (13.146), by the uniqueness of the solution of the Dirichlet problem (see Remark 6), implies that $V(P)$ is identically zero in B. Now it is enough to use the jump relations (see (13.110))

$$f(Q) = -\frac{1}{4\pi} \left\{ \left(\frac{\partial V}{\partial \nu} \right)_+ - \left(\frac{\partial V}{\partial \nu} \right)_- \right\}$$

to see that one has

$$f(Q) \equiv 0, \ Q \in S$$

as it was to be proved. □

Remark 7. It has to be clear that the systems $\{r^n Y_{nm}|_S\}\left\{\frac{1}{r^{n+1}} Y_{nm}|_S\right\}$ are complete in $L^2(S)$ but not orthogonal in this space, unless S is itself a sphere. This means that if one wants to approximate any $f(Q) \in L^2(S)$ by means of a finite combination of $S_{nm}(P)$ (internal or external), then one cannot use a simple projection argument by using orthogonality relations.

The orthogonal projection of $f(P)$ on

$$\text{Span}\{S_{nm}, n \leq N\} \equiv \left\{\sum_{n=0}^{N} \sum_{m=-n}^{n} \lambda_{nm} r_P^n Y_{nm}(\vartheta_P, \lambda_P)\right\}$$

has to be found by solving a Galerkin system (cf. Sect. 15.5), namely

$$\sum_{n=0}^{N} \sum_{m=-n}^{n} \lambda_{nm} \left\{\frac{1}{4\pi} \int_S r^{n+j} Y_{nm}(\vartheta_P, \lambda_P) Y_{jk}(\vartheta_P, \lambda_P) dS_P\right\}$$
$$= \frac{1}{4\pi} \int_S f(P) r_P^j Y_{jk}(\vartheta_P, \lambda_P) dS_P, \tag{13.147}$$

providing the function of $\text{Span}\{S_{nm}, n \leq N\}$ which is closest to $f(Q)$ in the $L^2(S)$ norm. The system (13.147) can indeed become very large, having as many as $(N+1)^2$ unknowns and its "normal" matrix is fully populated when S has not any particular symmetry. So its numerical solution can be sought more easily by iterative methods rather than by exact methods, like Cholesky.

It is worth noting the strict analogy of (13.147) with the standard least squares normal system, so widely used in Geodesy.

13.5 Green's Function and Krarup's Theorem

The reason why we are so interested in establishing the completeness of $\{S_{nm}\}$ in $L^2(S)$ is that we hope that while simple potentials like

$$u_N(P) = \sum_{n=0}^{N} \sum_{m=-n}^{n} u_{nm} S_{nm}(P)$$

do approach a given $L^2(S)$ function $f(P)$ on the boundary, on the same time inside B they do converge to some harmonic function $u(P)$ which then could be considered as solution of the Dirichlet problem with $f(P)$ as given boundary value, at least in some suitable sense.

In order to get a result of this kind we need to introduce the classical concept of Green's function and its use in potential theory.

Proposition 12 (Green's function). *Given B with a smooth boundary S, as above specified, there is a function $G(P, Q)$ (called Green's function of B) of two points $P, Q \in B$, such that, for fixed $P \in B$,*

$$\Delta_Q G(P, Q) = -4\pi\delta(P, Q) \tag{13.148}$$

$$G(P, Q)|_{Q\in S} = 0. \tag{13.149}$$

The Green function $G(P, Q)$ of B enjoys the following properties:

(a) Put

$$v(P) = \frac{1}{4\pi}\int_B G(P, Q)g(Q)dB_Q, \tag{13.150}$$

with g a measurable bounded function in B, then, at least in distribution sense,

$$\begin{cases} \Delta v(P) = -g(P) \ \text{ in } B \\ v(P)|_S = 0; \end{cases} \tag{13.151}$$

moreover $v(P)$ results to be continuous with its first derivatives in $\overline{B}$,

(b) Put

$$u(P) = -\frac{1}{4\pi}\int_S G_{n_Q}(P, Q)f(Q)dS_Q, \tag{13.152}$$

where $G_{n_Q}(P, Q)$ is the normal derivative of $G(P, Q)$ at $Q \in S$, $f \in C(S)$, then

$$\begin{cases} \Delta u(P) = 0 \qquad \text{in } B \\ u(P)|_S = f(P), \end{cases} \tag{13.153}$$

(c)

$$G(P, Q) \geq 0, \ P, Q \in B, \tag{13.154}$$

$$-G_{n_Q}(P, Q) \geq 0, \ Q \in S \tag{13.155}$$

(d) $G(P, Q)$ is symmetric, i.e.

$$G(Q, P) = G(P, Q)$$

Proof. Since

$$\Delta_Q \frac{1}{\ell_{PQ}} = -4\pi\delta(P, Q),$$

it is clear that if we define $h(P, Q)$ for every P fixed in B, such that

$$h(P, Q) : \begin{cases} \Delta_Q h(P, Q) = 0 \\ h(P, Q)|_{Q \in S} = \frac{1}{\ell_{PQ}} \end{cases} \tag{13.156}$$

and we put

$$G(P, Q) = \frac{1}{\ell_{PQ}} - h(P, Q), \tag{13.157}$$

we satisfy (13.148) and (13.149).

That $h(P, Q)$ exists $\forall P \in B$ (remember that B is open), is a consequence of Theorem 5. From the same theorem we derive that $h(P, Q)$, as well as $G(P, Q)$, has λ-Hölder continuous derivatives in $\overline{B}$; in particular it will be

$$Q \in S, \quad -G_{n_Q}(P, Q) = \left|G_{n_Q}(P, Q)\right| \leq C, \tag{13.158}$$

as far as P is fixed in B, which implies that also $\frac{1}{\ell_{PQ}}$ is continuous and with Lipschitz continuous derivatives on the boundary S.

Property (a) is proved by using the definition of Laplacian in distribution sense, namely by recalling that (remember that $\mathcal{D}(B)$ is the linear space of functions that are C^∞ in B and that are identically equal to zero outside a closed, bounded set $K \subset B$)

$$\forall \varphi \in \mathcal{D}(B), \int \varphi(P)\Delta v(P) dB_P = \int \Delta\varphi(P) v(P) dB_P$$

and by interchanging the integration order when we use a $v(P)$ as in (13.150). Then (13.148), together with the symmetry of $G(P, Q)$, (point d)), means exactly that

$$\int \Delta\varphi(P) \left(-\frac{1}{4\pi} G(P, Q)\right) dB_Q = \varphi(Q),$$

and we find, for every smooth $\varphi(P)$

$$\int \varphi(P)\Delta v(P)dB_P = -\int \varphi(Q)g(Q)dB_Q,$$

i.e. (13.151). That $v(P)|_S = 0$, comes from symmetry of $G(P, Q)$ and from (13.149).

We don't prove here that $v(P)$ is continuous with its first derivatives in $\overline{B}$.

To prove (b), start with the function u that is solution of (13.153) and assume it is continuous with its first derivatives in $\overline{B}$, i.e. $u \in C^1(\overline{B})$, so that we can write (cf. Part I, (1.61))

$$P \in B, \quad u(P) = \frac{1}{4\pi}\int_S \left\{u_n(Q)\frac{1}{\ell_{PQ}} - u(Q)\partial_{n_Q}\frac{1}{\ell_{PQ}}\right\} dS_Q. \quad (13.159)$$

On the other hand, when $Q \in S$, $\frac{1}{\ell_{PQ}} \equiv h(P, Q)$, so that using the identity

$$\int_S u_{n_Q}(Q)h(P, Q)dS_Q = \int_S u(Q)h_{n_Q}(P, Q)dS_Q,$$

we receive from (13.159)

$$\begin{aligned} u(P) &= \frac{1}{4\pi}\int_S u(Q)\left\{h_{n_Q}(P, Q) - \partial_{n_Q}\frac{1}{\ell_{PQ}}\right\} dS_Q \\ &= -\frac{1}{4\pi}\int_S G_{n_Q}(P, Q)u(Q)dS_Q; \end{aligned} \quad (13.160)$$

since $u(P)$ is continuous in $\overline{B}$ and $u(Q)|_S \equiv f(Q)$, and (13.152) and (13.153) are proved.

The restrictive condition $u(P) \in C^1(\overline{B})$ is eliminated by taking a sequence $f_n(P) \in C^1(S)$ such that $f_n(P) \to f(P)$ uniformly on S; then by the maximum principle $u_n(P) \to u(P)$ inside B, and (13.159) holds for $u(P)$ and $f(P)$ because $G_{n_Q}(P, Q)$ is a bounded function when $P \in B$.

Note that, since $u(P)$ is continuous up to the boundary, (13.160) tells us that (cf. Fig. 13.8)

$$P_0 \in S, \quad u(P_0) = \lim_{\delta\to 0} u(P_0 - \delta\mathbf{n}) = \lim_{\delta\to 0} -\frac{1}{4\pi}\int_S G_n(P_0 - \delta\mathbf{n}, Q)u(Q)dS_Q$$

meaning exactly that $\left\{-\frac{1}{4\pi}G_\nu(P_0 - \delta\mathbf{n}, Q)dS_Q\right\}$ tends to a measure of mass one concentrated in P_0 as a measure on S, when $\delta \to 0$.

Point c) is a consequence of the maximum principle; in fact fix P in B and a small sphere $B_\varepsilon(P)$ around P all contained in B (cf. Fig. 13.7),

then

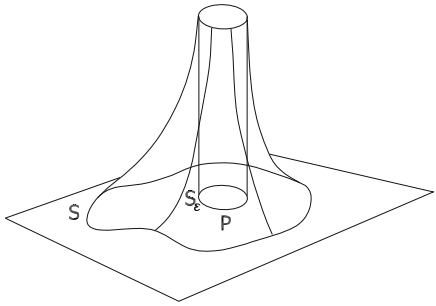


Fig. 13.7 An image of $G(P, Q)$

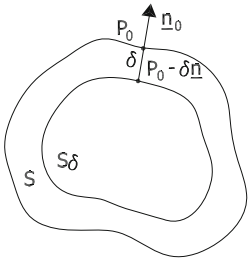


Fig. 13.8 The surface S and its internal translate at distance δ. This exists for sufficiently small δ

$$\varepsilon < \min_{Q \in S} \ell_{PQ}.$$

On the other hand $h(P, Q)$ for P fixed in B and Q variable, has extremes on S, so that

$$0 =< \min_{Q \in S} \frac{1}{\ell_{PQ}} \leq h(P, Q) \leq \max_{Q \in S} \frac{1}{\ell_{PQ}} < \frac{1}{\varepsilon}.$$

But then on S_ε

$$Q \in S_\varepsilon,\ G(P, Q) = \frac{1}{\varepsilon} - h(P, Q) > 0.$$

Since $G(P, Q)$ is harmonic between S_ε and S, where $G(P, Q)$ is zero, and ε is arbitrary, we find that (13.154) has to hold.

Moreover, if you put

$$V(\delta) = G(P, Q_0 - \delta \mathbf{n}), \quad Q_0 \in S,$$

you see that

$$-V(\delta) \leq 0,\ V(0) = 0$$

so that you find

$$\frac{\partial}{\partial\delta}V(\delta)\Big|_{\delta=0} = -G_{n_Q}(P, Q_0) \geq 0,$$

i.e. (13.155).

Finally, to prove point d) one needs to prove only that $h(P, Q)$ is symmetric. Write the identity (cf. (13.152), (13.157) and the second of (13.156))

$$h(P,Q) = \frac{1}{4\pi}\int_S h(P,Q')\left\{h_{n_{Q'}}(Q,Q') - \partial_{n_{Q'}}\frac{1}{\ell_{Q,Q'}}\right\}dS_{Q'} \tag{13.161}$$

$$= \frac{1}{4\pi}\int_S h(P,Q')h_{n_{Q'}}(Q,Q')dS_{Q'} - \frac{1}{4\pi}\int_S \frac{1}{\ell_{PQ'}}\partial_{n_{Q'}}\frac{1}{\ell_{QQ'}}dS_{Q'}.$$

The first integral is symmetric because one can move $\frac{\partial}{\partial n_{Q'}}$ from $h(Q, Q')$ and apply it to $h(P, Q')$, as a consequence of the second Green identity applied to two harmonic functions (cf. Part I, (1.57)). As for the second term note that, for $P \neq Q$,

$$\frac{1}{4\pi}\int_S\left\{\frac{1}{\ell_{PQ'}}\partial_{n_{Q'}}\frac{1}{\ell_{QQ'}} - \frac{1}{\ell_{QQ'}}\partial_{n_{Q'}}\frac{1}{\ell_{PQ'}}\right\}dS_{Q'}$$

$$= \frac{1}{4\pi}\int_B\left\{\frac{1}{\ell_{PQ'}}\Delta\frac{1}{\ell_{QQ'}} - \frac{1}{\ell_{QQ'}}\Delta\frac{1}{\ell_{PQ'}}\right\}dB_{Q'}$$

$$= -\int_B\left\{\frac{1}{\ell_{PQ'}}\delta(Q,Q') - \frac{1}{\ell_{QQ'}}\delta(P,Q')\right\}dB_{Q'} = -\left(\frac{1}{\ell_{PQ}} - \frac{1}{\ell_{PQ}}\right) = 0;$$

therefore also the second integral in (13.161) is symmetric, as it was to be proved. □

We are now ready to prove a theorem which extends (13.152) and (13.153) to any $f \in L^2(S)$.

Theorem 8. *Given any $f \in L^2(S)$ we find a unique solution of the Dirichlet problem, i.e. we can extend the Green operator (13.152) to $L^2(S)$ by continuity, in the sense that we find a sequence of functions harmonic in B, $\{u^{(N)}\}$ such that $u^{(N)}\big|_S = f^{(N)}$, with $f^{(N)} \in C(S)$ and $u^N(P) \to u(P)$, uniformly in any closed subset of B, and simultaneously $f^N \to f$ in $L^2(S)$; $u(P)$ is related to f by (13.152), for any P in B open, and therefore it is harmonic in B.*

Proof. Assume we have proven a majorization of the type

$$\int_B u^2(P)dB_P \leq C\int_S u^2(Q)dS_Q, \tag{13.162}$$

at least $\forall u \in C(\overline{B})$.

Then we can take any $f^{(N)} \in C(S)$ and such that $\|f^{(N)} - f\|_{L^2(S)} = \int_S (f^{(N)} - f)^2 dS$ tends to zero and define the corresponding $u^{(N)}$ through (13.152); obviously $u^{(N)}$ satisfties (13.153). On the same time, due to (13.162), $u^{(N)}$ has an $L^2(B)$ limit in B, i.e. there is $u(P) \in L^2(B)$ such that

$$\|u - u^{(N)}\|_{L^2(B)} \to 0. \tag{13.163}$$

Now, take any closed set $K \subset B$; then there is a $\delta > 0$ such that

$$\underset{\substack{P \in K \\ Q \in S}}{\text{Dist}}(P, Q) \geq \delta > 0. \tag{13.164}$$

Since $u^{(N)}(P)$ are harmonic they satisfy the mean value property, i.e. $\forall P \in K$, taken the sphere $B_\delta(P)$ one has $B_\delta(P) < B$ and

$$u^{(N)}(P) = \frac{1}{4/3\pi\delta^3} \int_{B_\delta} u^{(N)}(Q) dB_Q;$$

so, $\forall P \in K$,

$$\begin{aligned} &\left|u^{(N+k)}(P) - u^{(N)}(P)\right| \\ &\quad \leq \frac{1}{4/3\pi\delta^3} \int_{B_\delta(P)} \left|u^{(N+k)}(Q) - u^{(N)}(Q)\right| dB_Q \\ &\quad \leq \sqrt{\frac{1}{4/3\pi\delta^3}} \|u^{(N+k)} - u^{(N)}\|_{L^2(B_\delta(P))} \\ &\quad \leq \sqrt{\frac{1}{4/3\pi\delta^3}} \|u^{(N+k)} - u^{(N)}\|_{L^2(B)}. \end{aligned} \tag{13.165}$$

Since (13.165) holds uniformly in k and δ is fixed we have that $\{u^{(N)}(P)\}$ converges uniformly in K and $u(P)$ is then continuous in every $K \subset B$, i.e. in the whole B.

Furthermore

$$u(P) = \lim_{N \to \infty} -\int_S G_n(P, Q) f^{(N)}(Q) dS_Q = -\int_S G_n(P, Q) f(Q) dS_Q \tag{13.166}$$

the limit being justified by the fact that the distance of P from S is positive and then $G_n(P, Q)$ is continuous and bounded for $Q \in S$. The relation (13.166) proves that $u(P)$ is harmonic in B. The same conclusion can be derived from the fact that $u(P)$ has to satisfy the mean value property too and then, on account of Proposition 10, $u(P)$ has to be harmonic.

The inequality (13.162) is proved as follows: first assume $u(P)$ to be continuous with its gradient in $\overline{B}$ and apply the Green identity to $u^2(Q)$ and $G(P,Q)$, for any fixed P in B. Recalling that $G(P,Q)=0$, when $Q\in S$, and that $\Delta u^2 = 2|\nabla u|^2$

$$\begin{aligned}
&\int_B 2|\nabla u|^2 G(P,Q)dB_Q + 4\pi u^2(P)\\
&\quad = \int_B \left\{\left[\Delta u^2(Q)\right]G(P,Q) - u^2(Q)\Delta G(P,Q)\right\}dB_Q\\
&\quad = \int_S \left\{\left[\partial_n u^2(Q)\right]G(P,Q) - u^2(Q)G_{n_Q}(P,Q)\right\}dS_Q \qquad (13.167)\\
&\quad = -\int u^2(Q)G_{n_Q}(P,Q)dS_Q.
\end{aligned}$$

Since $G(P,Q)>0$, when $P,Q\in B$, (13.167) implies

$$u^2(P) \le -\frac{1}{4\pi}\int_S G_{n_Q}(P,Q)u^2(Q)dS_Q\ . \qquad (13.168)$$

Integrating over B one gets

$$\int_B u^2(P)dB_P \le \int_S \left[-\frac{1}{4\pi}\int_B G_{n_Q}(P,Q)dB_P\right]u^2(Q)dS_Q. \qquad (13.169)$$

On the other hand

$$\begin{aligned}
V(Q) &= -\frac{1}{4\pi}\int_B G(P,Q)dB_P\\
&= -\frac{1}{4\pi}\int_B G(Q,P)dB_P
\end{aligned}$$

is a function of the type (13.150), which is then continuous up to the boundary with its first derivatives. But then

$$\left|\ \partial_{n_Q}V(Q)\right|\big|_S \le C,$$

which inserted into (13.169) gives (13.162). □

Corollary 3. *The harmonic function*

$$u(P) = -\frac{1}{4\pi}\int_S G_{n_Q}(P,Q)f(Q)dS_Q \qquad (13.170)$$

with $f(Q)$ in $L^2(S)$ admits in fact $f(Q)$ as trace on the boundary S in the sense that, taken a sufficiently small δ as in Fig. 13.8, one has

$$\lim_{\delta \to 0} \int_S [u(P_0 - \delta \mathbf{n}) - f(P_0)]^2 dS_{P_0} = 0. \tag{13.171}$$

Proof. For the proof of (13.171) see Cimmino (1952, 1955) and Sansò and Venuti (1998). Here we make only a small reasoning which satisfies our intuition that $f(P)$ has to be given by the values attained by $u(P)$ on the boundary S. In fact consider that the harmonic polynomials $\{r^n P_{nm}(\vartheta, \lambda)|_S\}$ do form a total system in $L^2(S)$; therefore they can always be orthonormalized in $L^2(S)$ (see Remark 7) providing so a basis of polynomials, that we shall call $h_N(P)$, such that $\{h_N(P)|_S\}$ is a CON system in $L^2(S)$. We shall have

$$h_N(P) = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm} r^n Y_{nm}(\vartheta, \lambda),$$

so that $h_N(P)$ are harmonic and certainly continuous in $\overline{B}$. So if we put

$$f^{(N)}(P) = \sum_{k=0}^{N} f_k h_k(P)$$

we get a sequence such that, for suitable fixed coefficients $\{f_k\}$,

$$\lim_{N \to \infty} \| f(P) - f^{(N)}(P) \|^2_{L^2(S)} = 0,$$

i.e.

$$f(P) = \sum_{k=0}^{+\infty} f_k h_k(P), \quad P \in S. \tag{13.172}$$

On the other hand the functions $f^{(N)}(P)$ are well-defined and harmonic throughout all B so that we can take

$$u^{(N)}(P) = f^{(N)}(P), \ P \in B;$$

furthermore the sums $u^{(N)}(P)$ do converge uniformly to $u(P)$ in every closed set $K \subset B$

$$u(P) = \sum_{k=0}^{+\infty} f_k h_k(P). \tag{13.173}$$

Now if we simply take $P \in S$ in (13.173), we find that

$$u(P) = f(P), \ P \in S,$$

such equality meaning that the series (13.173) is $L^2(S)$ convergent to $f(P)$.

This rather heuristic proof can be made more rigorous, but cannot substitute (13.171), which has to be proved by a further specific analysis. □

Proposition 13. *The set of functions $\{u(P)\}$ which are harmonic in B and such that $\| u(P)|_S \|_{L^2(S)}$ is finite, is a Hilbert space with scalar product*

$$< u, v >_{L^2(S)} = \int_S u(P)v(P)dS_P; \tag{13.174}$$

both, scalar products and norms, have to be understood as limits of similar expressions from inside; for instance (13.174) means

$$< u, v >_{L^2(S)} = \lim_{\delta \to 0} \int u(P_0 - \delta \mathbf{n})v(P_0 - \delta \mathbf{n})dS_{P_0}; \tag{13.175}$$

we call this Hilbert space $HL^2(S)$.

Proof. That limits like (13.175) do exists is in fact consequence of the Corollary of Theorem 8.

That $HL^2(S)$ is a Hilbert space descends from the fact that the correspondence

$$u \in HL^2(S) \Leftrightarrow f = u\big|_S \in L^2(S)$$

is one-to-one thanks to Theorem 8, and isometric in the sense that

$$\|u\|_{HL^2(S)} = \|f\|_{L^2(S)}.$$

So any convergent sequence in one space corresponds to a convergent sequence in the other; moreover $L^2(S)$ *is* a Hilbert space, i.e. it is complete and so the same is true for $HL^2(S)$. □

We are able now to prove a very important theorem which is known in geodetic literature with the name of Runge-Krarup theorem, sometimes also associated to the name of Keldysh-Laurentiev (cf. Krarup 2006; Moritz 1980).

As a matter of fact this piece of theory, specially in the formulation of Krarup, is very general, however we will provide here a version which is adapted to that part of potential theory that is explored in these notes and, in particular, to the case of potentials which are in $HL^2(S)$.

Theorem 9. *Let B be an open domain as specified at the beginning of Sect. 2.2 and B_0 another open domain, satisfying similar hypotheses and such that*

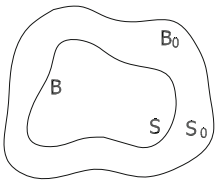


Fig. 13.9 The two nested domains B and B_0. Note that, due to condition (13.175), S and S_0 can never touch each other

$$B_0 \supset \overline{B} \tag{13.176}$$

(see Fig. 13.9).

Denote with S_0, S the boundaries of B_0, B and with $HL^2(S_0), HL^2(S)$ two Hilbert spaces of functions harmonic in B_0 and B respectively. Define the restriction operator $\mathcal{R}_B$ so that to any $u_0(P) \in HL^2(S_0)$ we associate the same function but restricted to the domain B; it is clear that such a function will be harmonic in B and even more it will be in $HL^2(S)$ because S is completely included in B_0 and $u_0(P)$ is then continuous on S; formally

$$\mathcal{R}_B : HL^2(S_0) \to HL^2(S);\ \mathcal{R}_B(u_0) = u_0(P)\big|_B\,; \tag{13.177}$$

then the set

$$\mathcal{R}_B[HL^2(S_o)] \equiv \{u \in HL^2(S); u = \mathcal{R}_B u_0, u_0 \in HL^2(S_0)\}$$

is dense in $HL^2(S)$. This means that

$$\forall u \in HL^2(S),\ \exists\{u_N\} \in HL^2(S_0)$$
$$\Rightarrow \|u - u_N\|_{HL^2(S)} = \left\{\int_S [u(P) - u_N(P)]^2 dS_P\right\}^{1/2} \to 0.$$

Proof. The proof is straightforward. We just note that $\{S_{nm}(r, \vartheta, \lambda)\} \in HL^2(B_0)$ and on the other hand this sequence is total in $HL^2(S)$.

So we have simultaneously

$$\text{Span}\{S_{nm}(r, \vartheta, \lambda)\} \subset HL^2(S_0);\ \mathcal{R}_B\text{Span}\{S_{nm}(r, \vartheta, \lambda)\} \subset HL^2(S).$$

Then, by taking the closure of the second relation in $HL^2(S)$, one has

$$HL^2(S) = \overline{\mathcal{R}_B\text{Span}\{S_{nm}\}}; \tag{13.178}$$

at the same time, by the first relation

$$\mathcal{R}_B\text{Span}\{S_{nm}\} \subseteq \mathcal{R}_B HL^2(S_0) \subseteq HL^2(S)$$

which, closed in $HL^2(S)$, yields

$$HL^2(S) = \overline{\mathcal{R}_B \mathrm{Span}\{S_{nm}\}} \subseteq \overline{\mathcal{R}_B HL^2(S_0)} \subseteq HL^2(S). \tag{13.179}$$

(13.178) and (13.179) together prove the theorem. □

Remark 8. Since B_0 in the previous theorem is arbitrary, one can use as B_0 a ball and indeed instead of $HL^2(B_0)$ one can use any Hilbert space of functions harmonic in B_0, such that all the $S_{nm}(r, \vartheta, \lambda)$ do belong to it.

For instance, take B_0 to be a ball of radius R, such that $B_0 \supset \overline{B}$, and take the Hilbert space HK with reproducing kernel (cf. Theorem 3)

$$\begin{aligned} K(P, Q) &= \sum_{n,m=0}^{+\infty} k_n \left(\frac{r_P}{R}\right)^n \left(\frac{r_Q}{R}\right)^n Y_{nm}(\vartheta_Q, \lambda_P) Y_{nm}(\vartheta_Q, \lambda_Q) \\ &= \sum_{n,m=0}^{+\infty} k_n S_{nm}(P) S_{nm}(Q) \qquad (k_n > 0, \ \forall n) \end{aligned} \tag{13.180}$$

That HK is a Hilbert space is easy to verify, that it contains all the solid spherical harmonics is a consequence of (13.180) and in particular of the condition $k_n > 0$; in fact recalling Theorem 3, formula (13.180) tells us that $\{\sqrt{k_n} S_{nm}(P)\}$ is a CON system in HK.

That the functions in HK are harmonic in B_0 is also clear from the shape of $K(P, Q)$ and the fact that by definition

$$f(P) = < K(P, Q), f(Q) >_{HK} . \tag{13.181}$$

Finally, in order that (13.180) be not a pure formal expression, one needs to impose some convergence conditions to the coefficients $\{k_n\}$. Observing that (13.180) can be written as

$$K(P, Q) = \sum_{n=0}^{+\infty} (2n + 1) k_n \left(\frac{r_P r_Q}{R^2}\right)^n P_n(\cos \psi_{PQ}) \tag{13.182}$$

and recalling that $|P_n(t)| \leq 1$, one immediately sees that under the condition

$$\sum_{n=0}^{+\infty} k_n (2n + 1) < +\infty, \tag{13.183}$$

the series (13.180) is uniformly convergent up to the boundary, i.e. up to the sphere of radius R.

The Theorem 9 is so relevant to the understanding of physical geodesy, that we restate it, in the form of a Corollary, in its outer version, which holds automatically true by virtue of the inverse radii transformation (see Proposition 1).

Corollary 4. *Let B, S be as in Theorem 9 and let $\Omega = (\overline{B})^c$, be the space exterior to S; let now B_0, with boundary S_0, be such that*

$$\overline{B}_0 \subset B \tag{13.184}$$

and $\Omega_0 = (\overline{B}_0)^c$, so that

$$\Omega_0 \supset \overline{\Omega}; \tag{13.185}$$

let $H_e L^2(S)$ be the Hilbert space of functions harmonic in Ω, regular at infinity endowed with the norm

$$\|u\|^2_{H_e L^2(S)} = \int_S u^2(P) dS_P, \tag{13.186}$$

and let $H_e L^2(S_0)$ be the similar space for Ω_0.

Note that we have added an index e to signify that here we are dealing with functions harmonic in the external domains (Ω, Ω_0) as opposed to the case discussed in Theorem 9. Let us define $\mathcal{R}_\Omega$ as the operator of restriction to Ω, applied to functions in $H_e L^2(S_0)$; then we have

$$\overline{\mathcal{R}_\Omega [H_e L^2(S_0)]} = H_e L^2(S); \tag{13.187}$$

i.e. $\forall u \in H_e L^2(S)$ there is a sequence $\{u_N(P)\} \in H_e L^2(S_0)$, harmonic in Ω_0 such that

$$\lim_{N \to \infty} \int_S [u(P) - u_N(P)]^2 dS = 0 \tag{13.188}$$

and that consequently $u_N(P) \to u(P)$ pointwise in Ω and even uniformly in every closed bounded set contained in Ω.

Moreover if, in analogy with Remark 8, we consider the case that B_0 is a ball of radius R and S_0 a so-called Bjerhammar sphere*, and the Hilbert space of functions harmonic in Ω_0, HK_e, endowed with the reproducing kernel*

$$K_e(P, Q) = \sum_{n,m} k_n \left(\frac{R}{r_P}\right)^{n+1} \left(\frac{R}{r_Q}\right)^{n+1} Y_{nm}(\vartheta_P, \lambda_P) Y_{nm}(\vartheta_Q, \lambda_Q)$$

$$= \sum_{n=0}^{+\infty} (2n+1) k_n \left(\frac{R^2}{r_P r_Q}\right)^{n+1} P_n(\cos \psi_{PQ}) \qquad (k_n > 0, \ \forall n), \tag{13.189}$$

we still have

$$\overline{R_\Omega[HK_e]} = H_e L^2(S) \tag{13.190}$$

and (13.188) holds with $\{u_N(P)\} \in HK_e$, *i.e. harmonic in* Ω_0, *down to the Bjerhammar sphere* S_0.

13.6 Exercises

Exercise 1. Prove Proposition 1 by showing, with the use of spherical coordinates and assuming $R = 1$, that

$$\begin{aligned}\Delta_s v(s,\vartheta,\lambda) &= \frac{\partial^2 v}{\partial s^2} + \frac{2}{s}\frac{\partial v}{\partial s} + \frac{1}{s^2}\Delta_\sigma v \\ &= \frac{1}{s^5}\left[u''\left(\frac{1}{s},\vartheta,\lambda\right) + 2su'\left(\frac{1}{s},\vartheta,\lambda\right) + s^2\Delta_\sigma u\left(\frac{1}{s},\vartheta,\lambda\right)\right] \\ &= \frac{1}{s^5}\Delta_r u(r,\vartheta,\lambda) = 0,\end{aligned}$$

where $u'(r,\vartheta,\lambda) = \frac{\partial}{\partial r}u(r,\vartheta,\lambda)$.

Exercise 2. Compute $h_m(x,y), h_{-m}(x,y)$ directly for $m = 2, m = 3$ and prove that they give the same result as those computed from (13.12), namely

$$\begin{aligned} h_2 &= x^2 - y^2,\ h_{-2} = xy \\ h_3 &= x^3 - 3xy^2,\ h_{-3} = y^3 - 3x^2y.\end{aligned}$$

Exercise 3. Let h_{N-2k} be a harmonic polynomial in HH^3_{N-2k}; prove the formula

$$\begin{cases} \Delta^m r^{2k} h_{N-2k} = A_{mk} r^{2k-2m} h_{N-2k},\ m \le k, \\ A_{mk} = 2k(2k-2)\dots(2k-2m+2) \\ \cdot(2N-2k+1)(2N-2k-1)\dots(2N-2k-2m+3) \end{cases}. \tag{13.191}$$

For this purpose first prove that

$$\Delta r^{2\ell} h_{N-2k} = 2\ell(2N+2\ell-4k+1) r^{2\ell-2} h_{N-2k} \tag{13.192}$$

and then apply the Laplace operator m times to $r^{2k}h_{N-2k}$, using iteratively such a relation.

(**Hint**: to prove the second of the above relations, use $\Delta(fg) = (\Delta f)g + 2\nabla f \cdot \nabla g + f\Delta g$; note that $\Delta r^{2\ell} = 2\ell(2\ell+1)r^{2\ell-2}, \nabla r^{2\ell} = 2\ell r^{2\ell-2}\xi$ (with $\boldsymbol{\xi} = \frac{\mathbf{r}}{r}$) and for any function f homogeneous of degree α we have

$$\boldsymbol{\xi} \cdot \nabla f(\boldsymbol{\xi}) = \alpha f(\boldsymbol{\xi}) \,).$$

Exercise 4. Since we already know that formula (13.23) holds true, one can compute q_k just by imposing that it has to be $\Delta h_N = 0$. Prove that

$$q_1 = -\frac{1}{2(2N-1)}, \; q_2 = \frac{1}{2 \cdot 4(2N-1)(2N-3)} \dots \tag{13.193}$$

(**Hint**: prove, by using the same argument as in Exercise 3, that

$$\Delta(r^{2k}\Delta^k P_N) = 2k(2N-2k+1)r^{2k-2}\Delta^k P_N + r^{2k}\Delta^{k+1}P_N,$$

then impose $\Delta h_N = 0$, considering $r^{2k}\Delta^{k+1}P_N$ as independent variables).

Exercise 5. Prove that if

$$\Delta[r^n(t^n + a_1 t^{n-2} + a_2 t^{n-4} + \dots)] = 0$$

then a_1, a_2 are univocally determined.

(**Hint**: by using (13.36) prove that

$$\Delta\left[r^n t^{n-2\ell}\right] = r^{n-2}\left\{2\ell(2n-2\ell+1)t^{n-2\ell} + (n-2\ell)(n-2\ell-1)t^{n-2\ell-2}\right\}\Big).$$

Exercise 6. Prove that the coefficient c_n of t^n in $P_n(t)$ is

$$c_n = \frac{(2n!)}{2^n(n!)^2} \tag{13.194}$$

(**Hint**: recall that

$$(n+1)P_{n+1}(t) = (2n+1)tP_n(t) - nP_{n-1}(t)$$

and derive the recursive relation

$$(n+1)c_{n+1} = (2n+1)c_n.$$

Observe that from the expression of c_n for $n = 0$, $n = 1$, we correctly obtain $c_0 = c_1 = 1$ and then prove that c_n satisfies the above recursive relation).

Exercise 7. Compute the spherical harmonics $r^n Y_{nm}(\vartheta, \lambda)$ for all orders and degrees 2 and 3 and transform them back to polynomials in (x, y, z), verifying their harmonicity.

Exercise 8. Verify the summation rule (13.55) for degree 2 and 4. (Warning: note that for (13.55) to hold it is necessary to use fully-normalized spherical harmonics.)

Exercise 9. Prove that (13.74) is satisfied by the functions (13.70).
(**Hint**: first call L the Legendre operator

$$L\cdot = D_t(1-t^2)D_t\cdot$$

and remember that (cf. (13.37))

$$LP_n(t) = -n(n+1)P_n(t),$$

or

$$(1-t^2)D^2P_n = 2tDP_n - n(n+1)P_n \tag{13.195}$$

Recalling also that $P_{nm} = (1-t^2)^{m/2}\, P_n^{(m)}$, $P_n^{(m)} = D_n^m$, prove that

$$\begin{aligned} LP_{nm}(t) = L\left[(1-t^2)^{m/2}P_n^{(m)}\right] = m[mt^2-(1-t^2)](1-t^2)^{\frac{m}{2}-1}P_n^{(m)} + \\ -2t(m+1)(1-t^2)^{m/2}P_n^{(m+1)} + (1-t^2)^{\frac{m}{2}+1}P_n^{(m+2)}; \end{aligned} \tag{13.196}$$

then from (13.195), by applying D^m to both members and recalling that

$$D^m(fg) = \sum_{k=0}^{m}\binom{m}{k} D^k(f)\cdot D^{m-k}(g),$$

derive

$$\begin{aligned} &D^m[(1-t^2)D^2P_n] \\ &\quad = (1-t^2)P_n^{(m+2)} - 2mtP_n^{(m+1)} - m(m-1)P_n^{(m)} \\ &\quad = 2tP_n^{(m+1)} + 2mP_n^{(m)} - n(n+1)P_n^{(m)}; \end{aligned}$$

rearranging the last equality you get

$$(1-t^2)P_n^{(m+2)} = 2t(m+1)P_n^{(m+1)} + [m(m+1) - n(n+1)]P_n^{(m)};$$

then substitute back in LP_{nm}).

Exercise 10. Derive the normalization constant

$$k_{n0} = \sqrt{2n+1}$$

from the reproducing relation (cf. Part I, (3.188)).

$$\frac{1}{4\pi}\int_{S_1} P_m(\cos\psi_{PQ})P_n(\cos\psi_{P'Q})d\sigma = (2n+1)^{-1}P_n(\cos\psi_{PP'})$$

(**Hint**: put $P = P'$ at the North Pole).

Exercise 11. Compute $\overline{P}_{nm}(t)$, up to degree and order 4, using (13.83) and (13.91).

Exercise 12. Repeat the reasoning of Example 2 for the exterior Dirichlet problem proving that, $\forall f(P) \in L^2(S_R)$,

$$\begin{cases} u(P) = \sum_{n,m} f_{nm}\left(\frac{R}{r_P}\right)^{n+1} Y_{nm}(\vartheta_P, \lambda_P) \\ \\ f_{nm} = \frac{1}{4\pi}\int f(\vartheta', \lambda')Y_{nm}(\vartheta', \lambda')d\sigma' \end{cases}$$

and that, accordingly

$$r_P > R, \quad u(P) = \frac{1}{4\pi}\int \Pi_{Re}(P, P')f(P')d\sigma_{P'}$$

with $\prod_{Re}(P, P')$, the external Poisson kernel,

$$r > R, \quad \Pi_{Re}(P, P') = \frac{R(r^2 - R^2)}{[r^2 + R^2 - 2rR\cos\psi]^{3/2}}.$$

Exercise 13. Using a complementary argument to that of Theorem 7 and a small sphere inside B, prove that the sequence of outer spherical harmonics $\left\{\frac{1}{r^{n+1}}Y_{nm}(\vartheta, \lambda\right\}$ restricted to S forms again a complete system in $L^2(S)$.

Exercise 14. Prove that the Green function of the sphere with radius R is given by

$$G(P, Q) = \frac{1}{\sqrt{r_P^2 + r_Q^2 - 2r_P r_Q\cos\psi_{PQ}}} - \frac{1}{\sqrt{R^2 + \frac{r_P^2 r_Q^2}{R^2} - 2r_P r_Q\cos\psi_{PQ}}};$$

moreover verify that

$$-\frac{\partial}{\partial r_Q}G(P, Q)\Big|_{r_Q=R} = -\frac{1}{R^2}\Pi_{Ri}(P, Q)$$

as it has to be in view of (13.123) and (13.152)

(**Hint**: that $G(P,Q)|_{r_Q=R}=0$ is obvious. You need only to prove that $\left(R^2+\frac{r_P^2 r_Q^2}{R^2}-2r_P r_Q\cos\psi_{PQ}\right)^{-1/2} = h(P,Q)$ is harmonic in $Q\in B$; this is clear if one observes that, with $\mathbf{r}_P^*=\frac{R^2}{r_P^2}\mathbf{r}_P$, implying $|r_P^*|>R$, one can write

$$h(P,Q)=\frac{r_P}{R}\left(r_P^{*2}+r_P^2-2r_Q^* r_Q\cos\psi\right)^{-1/2}=\frac{r_P}{R}|\mathbf{r}_P^*-\mathbf{r}_Q|^{-1}\big)$$

Exercise 15. Prove that, when B is a ball of radius R, then the inequality (13.162) can be put in the rather expressive form (with $|B|=\frac{4}{3}\pi R^3,\ |S|=4\pi R^2$)

$$\frac{1}{|B|}\int_B u^2 dB\le\frac{1}{|S|}\int_S u^2 dS.$$

Similarly when we use a regular potential u which is harmonic in Ω, the space outside a sphere of radius R, one can write the inequality

$$\int_\Omega u^2\frac{1}{r^2}d\Omega\le R\int_\sigma u^2 d\sigma$$

(**Hint**: use just the two representations

$$u(P)=\sum_{n=0}^{+\infty}\sum_{m=-n}^{n}u_{nm}\left(\frac{r}{R}\right)^n Y_{nm}(\vartheta,\lambda),\qquad r\le R$$

$$u(P)=\sum_{n=0}^{+\infty}\sum_{m=-n}^{n}u_{nm}\left(\frac{R}{r}\right)^{n+1} Y_{nm}(\vartheta,\lambda),\qquad r\ge R$$

and compute directly $\int_S u^2 dS$, $\int_B u^2 dB$ in the first case and $\int_\Omega u^2\frac{1}{r^2}d\Omega$ in the second case. Remember that $\{Y_{nm}(\vartheta,\lambda)\}$ is orthonormal with

$$\frac{1}{4\pi}\int Y_{nm}(\vartheta,\lambda)^2 d\sigma=1\big).$$

Chapter 14
A Quick Look to Classical Boundary Value Problems (BVP) Solutions

14.1 Outline of the Chapter

We shall present into the next chapter, together with a link to proper numerical methods, the problem of approximating the gravity field potential in a modern mathematical form. Yet the same item has been treated in the past by different authors leading to numerical solutions, transforming the problem into integral equations, which are still applied in some cases. This matter is summarized in Sect. 14.2 from the historical point of view.

As it has been pointed out long ago, the classical solution by Molodensky is basically equivalent to a downward continuation followed by classical Stokes solution. This point is analyzed and explained in Sect. 14.3. Molodensky's formulas can be applied for the purpose of a local approximation of T too. This application, though, requires that the implied error could be at least roughly estimated. The problem is discussed and solved in Sect. 14.4.

In Sect. 14.5 a short review is presented, following a different approach dating back to Helmert ideas. Although this approach is somehow outside the main line followed in the book, the subject is presented for the sake of completeness.

Therefore this short chapter can be considered as an intermezzo in which the older material is re-organized and the definition of geodetic boundary value problems in modern form is prepared.

14.2 The Classical Molodensky Approach: A Historical Excursus

Since in what follows we shall reason on the residual part only of the anomalous potential and of the free air gravity anomaly field, we will be entitled to use a spherical approximation approach (cf. Part I, Sect. 2.6). So we could formulate our

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_14,

problem in the so-called form of a simple Molodensky problem (see Sect. 15.3), namely given a free air anomaly $\Delta g(P)$ on the telluroid surface S, to find a potential T, harmonic outside S, regular at infinity and such as to satisfy the boundary condition (cf. (15.35))

$$-\frac{\partial T}{\partial r} - \frac{2}{r}T\bigg|_S = \Delta g(P). \tag{14.1}$$

For the moment we shall assume that $\Delta g(P)$ is given all over the telluroid S, and we will come later on to prove that the approximate solution we are going to find can be reasonably applied as well, when $\Delta g(P)$ is given only on a certain portion A of S.

The problem with this and other similar formulations has been studied by a number of authors starting from J.J. Levallois, to J. de Graaf Hunter, to arrive at the milestone paper by M.S. Molodensky, V.F. Jeremyev and M.I. Yourkina and the subsequent studies by V.V. Brovar, L.P. Pellinen, and H. Moritz.

Yet the first systematic and complete solution of the simple Molodensky problem is due to T. Krarup in one of his famous *Letters on Molodensky's problem* that have been published only posthumous (Krarup 2006), though they have been so influential in the history of geodesy.

Funny enough, none of these authors have clearly stated that in reality there were two distinct problems that have been analyzed with a certain confusion between the two, due to the fact that both, linearized and put in spherical approximation, were leading to a boundary value problem of the form (14.1).

Yet they were in fact quite distinct from one another, because in one case, the so-called *vector Molodensky problem*, not only $\Delta g(P)$ had to be given on S, but also the deflections of the vertical (ξ, η), contrary to the so-called *scalar Molodensky problem*, where it is sufficient to assume that Δg only is known. This has been clarified by the author in Sacerdote and Sansò (1986).

The full modern anlaysis of the simple Molodensky problem as well as that of the just linearized Molodensky problem can be found into the next chapter and it stems from quite recent researches (Sansò and Venuti 2008), representing a benchmark of a number of previous works performed specially by Holota, (see Holota 1983 and Holota, Nesvadbe 2007).

The key point in all these analyses is that the function

$$v(P) = -r\frac{\partial}{\partial r}T - 2T = -\mathbf{r}\cdot\nabla T - 2T \tag{14.2}$$

is in fact harmonic whenever T is harmonic, as it is easily verified (cf. Proposition 4).

However the operator $r\frac{\partial}{\partial r} + 2$ is annihilating the first degree harmonics, so that $v(P)$ has to be void of such harmonics, when $r_P \to \infty$, if we want to be able to invert (14.2). As a matter of fact, since we organize things in such a way that both the harmonic developments of T and Δg, and hence $v(P)$, when $r_P \to \infty$ be lacking the zero and first degree harmonics, (14.2) can be explicitely inverted by the formula

(cf. Heiskanen and Moritz 1967, and Chap. 4 below)

$$T(r,\vartheta,\lambda) = \frac{1}{r^2}\int_r^{+\infty} s\cdot v(s,\vartheta,\lambda)ds, \tag{14.3}$$

as the reader is invited to verify directly.

All that understood, the solution of the simple Molodensky problem becomes quite easy: first one solves a Dirichlet problem searching for a $v(P)$ harmonic outside S, satisfying

$$v|_S = r_P \Delta g(P), \tag{14.4}$$

which is well-known in potential theory; then one computes the wanted potential T by (14.3). This is the core of the so-called *Prague method*. Its solution can be reduced to the solution of (14.4) which in turn is obtained by writing a suitable integral equation after representing v as the potential of a double layer. This is in fact one of the oldest tools applied in potential theory and to related boundary value problems. This solution though, is not very comfortable because it requires first to solve an integral equation and then to transform its solution by another surface integral in order to get T (Krarup 2006).

In this respect Brovar theory, or its equivalent formulation by Pellinen, is more manageable (see Heiskanen and Moritz 1967, Chap. 8, or Moritz 1980, Part D, Sect. 44 and 45); nevertheless the justification of the corresponding theory is based on a series development which has never been formally proved to converge (see Moritz 1980, Part D, Sect. 43, 45). A key issue in all these approaches is that $\Delta g(P)$ on S has to be considered as derived form a potential T, which is harmonic down to the mean earth sphere (rememeber that we are here reasoning in spherical approximation). Since in any way we work ultimately with discrete data and we have available the Runge–Krarup Theorem 9, we shall accept this hypothesis and we shall develop into the next section a quite simple interpretation of the theory based only on formulas approximate to the first order in $\left(\frac{h_P}{R}\right)$, h_P denoting the ellipsoidal height of S; this is very similar to what one finds in the classical text book by Heiskanen and Moritz (1967).

The resulting equation is as a matter of fact almost the only one applied in practice. In any case interesting references on more recent work on the same items are (Heck 2003a; Martinec 1998).

14.3 The Approximate Solution of Molodensky's Problem by Downward Continuation

We recall that we are working in spherical approximation, so that the earth ellipsoid $\mathcal{E}$ is mapped to the mean earth sphere $S_{\overline{R}}$, of radius $\overline{R}$, and points P with ellipsoidal heights h_P are mapped to points with radial distance

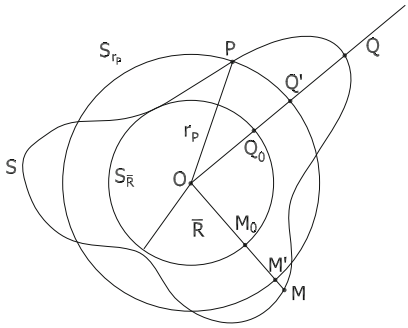


Fig. 14.1 Geometry of the downward continuation from S to S_{r_P}; note that $P, Q' \in S_{r_P}$, $Q \in S$; note too that Q, Q', Q_0 are in a one-to-one correspondence

$$r_P = \overline{R} + h_P. \tag{14.5}$$

Let us assume now that, according to the discussion of Sect. 14.2, Δg_P is a free air anomaly of a potential $T(P)$ harmonic down to $S_{\overline{R}}$ so that at any point P of the telluroid it is a very smooth function. To be precise, this is not exact for points P having a negative ellipsoidal heights. Nevertheless form the theory developed in this section, it will be apparent that even changing $\overline{R}$ by 0,1% of its value we reach conclusions with the same order of approximation. Therefore, since negative ellipsoidal heights are never so negative on the telluroid, the remark above becomes irrelevant.

Then we shall solve the simple Molodensky problem in two steps:

(a) Analytical continuation of $\Delta g(Q)$ from the telluroid surface S to a sphere S_{r_P}, with radius r_P,
(b) Application of the simple Stokes formula (cf. Part I, Example 3 of Chap. 3) to go from Δg to T.

In fact we have:

(a) Given that

$$r_P - r_Q = h_P - h_Q, \tag{14.6}$$

we can write (see Fig. 14.1)

$$\Delta g(Q') \cong \Delta g(Q) - \frac{\partial \Delta g}{\partial r}(h_Q - h_P) = \Delta g(Q) + G_1 \tag{14.7}$$

because $h_{Q'} = h_P$ by construction.

Note that with the above definition the correction term G_1 is a function of both P and Q,

(b) Once $\Delta g(Q')$ is approximately known all over S_{r_P}, we are in condition of applying the simple Stokes formula (cf. Example 3), namely

$$T(P) = \frac{r_P}{4\pi}\int S(\psi_{PQ'})\Delta g(Q')d\sigma_{Q'} \tag{14.8}$$
$$= \frac{r_P}{4\pi}\int S(\psi_{PQ'})[\Delta g(Q) + G_1(P,Q)]d\sigma_{Q'}.$$

Of course the problem here is to have a sensibile formula to express G_1, namely the $\frac{\partial \Delta g}{\partial r}$ at the point Q as a function of the known values of Δg. We have been working on this problem in another context, namely Part I, Sect. 2.4, however here we want to find a better approximation by expressing directly $\frac{\partial \Delta g}{\partial r}$ as an integral transform of $\Delta g(Q)$, without any coarse geometric reasoning on the mean curvature C of the equipotential surface.

To this aim we first rewrite (14.7) for the full downward continuation of $\Delta g(Q)$ to the level of sea, namely to Q_0,

$$\Delta g(Q_0) \cong \Delta g(Q) - \frac{\partial \Delta g(Q)}{\partial r} h_Q \tag{14.9}$$

and then we try to invert the Poisson integral equation, that relates $v(Q) = r_Q \Delta g(Q)$ to $v(Q_0) = r_{Q_0}\Delta g(Q_0)$. In fact, recalling that $v(Q)$, in spherical approximation, is harmonic, we must have (cf. Exercise 2)

$$v(Q) = \frac{\overline{R}(r_Q^2 - \overline{R}^2)}{4\pi}\int \frac{v(M_0)}{\ell_{QM_0}^3} d\sigma_{M_0}, \tag{14.10}$$

where M_0 is a point running on $S_{\overline{R}}$.

Since (14.10) has to hold for any harmonic function, we can write the identity (note that $\left.\frac{\overline{R}}{r_Q}\right|_{Q\in S_{\overline{R}}} \equiv 1$)

$$\frac{\overline{R}}{r_Q} = \frac{\overline{R}(r_Q^2 - \overline{R}^2)}{4\pi}\int \frac{1}{\ell_{QM_0}^3} d\sigma_{M_0}, \tag{14.11}$$

which by the way one can directly verify. If we multiply (14.11) by $v(Q_0)$ and subtract from (14.10) we get

$$v(Q) - \frac{\overline{R}}{r_Q} v(Q_0) = \frac{\overline{R}(r_Q^2 - \overline{R}^2)}{4\pi}\int \frac{v(M_0) - v(Q_0)}{\ell_{QM_0}^3} d\sigma_{M_0}. \tag{14.12}$$

Returning to Δg and rearranging we find

$$\Delta g_{Q_0} = \left(\frac{r_Q}{\overline{R}}\right)^2 \Delta g_Q + \frac{r_Q(r_Q^2 - \overline{R}^2)}{4\pi}\int \frac{\Delta g_{Q_0} - \Delta g_{M_0}}{\ell_{QM_0}^3} d\sigma_{M_0}. \tag{14.13}$$

Now we observe that $r_Q = \overline{R} + h_Q$ so that, retaining only first order quantities in $\frac{h_Q}{\overline{R}}$ which is of the maximum order of 10^{-3}, we have

$$\left(\frac{r_Q}{\overline{R}}\right)^2 \cong 1 + 2\frac{h_Q}{\overline{R}}, r_Q(r_Q^2 - \overline{R}^2) \cong 2\overline{R}^3\frac{h_Q}{\overline{R}} \tag{14.14}$$

Using (14.14) in (14.13) we receive

$$\Delta g_{Q0} \cong \left(1 + 2\frac{h_Q}{\overline{R}}\right)\Delta g_Q + \frac{h_Q}{\overline{R}}\frac{1}{2\pi}\int \frac{\overline{R}^3}{\ell_{QM_0}^3}[\Delta g_{Q_0} - \Delta g_{M_0}]d\sigma_{M_0}. \tag{14.15}$$

As intuition suggests, one can write

$$\frac{\overline{R}^3}{\ell_{QM_0}^3} \cong \frac{\overline{R}^3}{\ell_{Q_0M_0}^3}\left[1 + O\left(\frac{h_Q}{\overline{R}}\right)\right]. \tag{14.16}$$

Formally this is justified by the following reasoning: call ψ the spherical angle between Q and M_0, then

$$\ell_{QM_0} = [\overline{R}^2 + r_Q^2 - 2\overline{R}r_Q\cos\psi]^{(1/2)}$$

Now substitute $r_Q = \overline{R} + h_Q$, collect $\overline{R}^2$ and retain only first order terms in $\frac{h_Q}{\overline{R}}$ to arrive at

$$\begin{aligned}\ell_{QM_0} &\cong \sqrt{2}\,\overline{R}\left[\left(1 + \frac{h_Q}{\overline{R}}\right)(1 - \cos\psi)\right]^{(1/2)} \\ &\cong 2\overline{R}\sin\frac{\psi}{2}\left(1 + \frac{1}{2}\frac{h_Q}{\overline{R}}\right) = \ell_{Q_0M_0}\left(1 + \frac{1}{2}\frac{h_Q}{\overline{R}}\right);\end{aligned}$$

this proves (14.16).

On the other hand, when Q does not belong to $S_{\overline{R}}$, ℓ_{QM_0} can never become zero, while $\ell_{Q_0M_0}$ goes to zero when M_0 runs over the whole $S_{\overline{R}}$. Therefore the integral transform

$$I(\Delta g) = \int \frac{\overline{R}^3}{\ell_{Q_0M_0}^3}[\Delta g_{Q_0} - \Delta g_{M_0}]d\sigma_{M_0} \tag{14.17}$$

has a strong singularity at Q_0 (Mikhlin 1964).

Nevertheless if Δg_{Q_0} is smooth enough, for instance has bounded second derivatives, it is not hard to see that (14.17) has a precise meaning in terms of a so-called integral in Cauchy principal part; this means that (14.17) is computed first excluding an ε-cap around Q_0 and then letting $\varepsilon \to 0$ (Miranda 1970; Mikhlin 1964).

When the above hypotesis is true, we see that terms like $\frac{\overline{R}^3}{\ell^3_{Q_0M_0}} \cdot \frac{h_Q^2}{\overline{R}^2}$ which arise form substituting (14.16) into (14.15) can be considered as negligible.

So we finally can write instead of (14.15), with no significant loss of accuracy,

$$\Delta g_{Q_0} = \left(1 + 2\frac{h_Q}{\overline{R}}\right)\Delta g_Q + \frac{h_Q}{\overline{R}}\frac{1}{2\pi}\int \frac{\overline{R}^3}{\ell^3_{Q_0M_0}}[\Delta g_{Q_0} - \Delta g_{M_0}]d\sigma_{M_0}. \quad (14.18)$$

Already (14.18) shows that $\Delta g_Q = \Delta g_{Q_0} + O\left(\frac{h_Q}{\overline{R}}\right)$, so that if we substitute $\Delta g_Q - \Delta g_M$ (see Fig. 14.1) in the integral in (14.18), we commit a negligible error of the second order in $\frac{h_Q}{\overline{R}}$; therefore we can write

$$\Delta g_{Q_0} = \left(1 + 2\frac{h_Q}{\overline{R}}\right)\Delta g_Q + \frac{h_Q}{\overline{R}}\frac{1}{2\pi}\int \frac{\overline{R}^3}{\ell^3_{Q_0M_0}}[\Delta g_Q - \Delta g_M]d\sigma_{M_0}. \quad (14.19)$$

Comparing this with (14.9) we see that, at the present level of accuracy,

$$\frac{\partial \Delta g(Q)}{\partial r} \cong -\frac{2}{\overline{R}}\Delta g(Q) - \frac{\overline{R}^2}{2\pi}\int \frac{\Delta g(Q) - \Delta g(M)}{\ell^3_{Q_0M_0}}d\sigma_{M_0}. \quad (14.20)$$

This formula has now the great advantage that, from our data given on the telluroid S, namely $\Delta g[r_Q(\vartheta,\lambda),\vartheta,\lambda]$, we can easily compute $\frac{\partial \Delta g(Q)}{\partial r}$ as a convolution on the sphere $S_{\overline{R}}$, which, as it has been shown in Part II, can be quite favourably reckoned with the help of Fourier transforms.

From (14.20) the G_1 term can be calculated and from (14.8) we get the sought solution, that we summarize in the coupled equations

$$G_1(P,Q) = -\frac{\partial \Delta g(Q)}{\partial h}(h_Q - h_P) \quad (14.21)$$

$$= 2\left(\frac{h_Q - h_P}{\overline{R}}\right)\Delta g_Q + 2\left(\frac{h_Q - h_p}{\overline{R}}\right)\frac{\overline{R}^3}{4\pi}\int \frac{\Delta g_Q - \Delta g_M}{\ell^3_{Q_0M_0}}d\sigma_{M_0}$$

$$T(P) = \frac{r_P}{4\pi}\int S(\psi_{PQ'})[\Delta g(Q) + G_1(P,Q)]d\sigma_{Q'}. \quad (14.22)$$

To be precise and specific we notice that $S(\psi_{PQ})$ to be used in (14.22) is the purely angular version of the Stokes's function, namely, considering that $r_P = r_{Q'}$ by hypotheisis, the function (3.100) of Part I, i.e.

$$S(\psi_{PQ'}) = 1 + \left(\sin\frac{1}{2}\psi_{PQ'}\right)^{-1} - 6\sin\frac{1}{2}\psi_{PQ'} - 5\cos\psi_{PQ'} + \quad (14.23)$$

$$-3\cos\psi_{PQ'}\log(\sin\frac{1}{2}\psi_{PQ'} + \sin^2\frac{1}{2}\psi_{PQ'}).$$

14.4 On the Local Use of Molodensky's Formula

Formulas (14.21) and (14.22) put together constitute what we can call Molodensky's formula to the first order in $\frac{h_P - h_Q}{R}$. As a matter of fact this captures the most relevant terms of Molodensky's theory, since the higher order terms, even assuming that can provide a better approximation to T, are generally smaller than the errors implied by the assumption of the spherical approximation which we are applying here.

As we see in principle the application of (14.22) and (14.23) implies the knowledge of Δg all over the earth telluroid in both integrals, the one needed to compute the $G_1(P, Q)$ term and the Stokes integral.

On the other hand we have assumed here to be in the context of a *local* calculation, as we have already described in Part I, Sect. 5.10.

This means that:

- We have available data in a set A and we accept to derive T in a smaller set A_Δ (see Part I, Fig. 5.10)
- The data Δg that we shall use in the computation are residual gravity anomalies, where a global model and a residual terrain correction have already been subtracted.

Since when we are at the border of A_Δ we have data at most at an angular distance Δ, we propose to substitute the integral on the full sphere, corresponding to the range $S \equiv \{0 \leq \psi \leq \pi,\ 0 \leq \alpha < 2\pi\}$, with the truncated integral over the cap $C_\Delta \equiv \{0 \leq \psi \leq \Delta\ ;\ 0 \leq \alpha \leq 2\pi\}$.

So if we have to compute the convolution (cf. Part I, A.4)

$$u(P) = F * v = \frac{1}{4\pi} \int_{S_1} F(\psi_{PQ}) v(Q) d\sigma_Q \tag{14.24}$$

we rather compute a truncated version

$$u_\Delta(P) = F_\Delta * v = \frac{1}{4\pi} \int_{C_\Delta} F(\psi_{PQ}) v(Q) d\sigma_Q. \tag{14.25}$$

We immediately note that in (14.25) we can express the fact that the computation over the moving cap C_Δ is equivalent to a convolution over the whole sphere S_1 by using the truncated kernel

$$F_\Delta(\psi) = \begin{cases} F(\psi), & 0 \leq \psi \leq \Delta \\ 0 & \Delta < \psi \leq \pi. \end{cases} \tag{14.26}$$

So, subtracting (14.25) from (14.24) we find the pointwise truncation error $t(P)$,

$$t(P) = u(P) - u_\Delta(P) = (F - F_\Delta) * v \tag{14.27}$$
$$= F^\Delta * v = \frac{1}{4\pi} \int_{S_1 \setminus C_\Delta} F(\psi_{PQ}) v(Q) d\sigma_Q$$

where we have put

$$F^\Delta(\psi) = \begin{cases} 0, & 0 \le \psi \le \Delta \\ F(\psi), & \Delta < \psi \le \pi. \end{cases} \tag{14.28}$$

According to what we learnt in Part I, A.4, (14.28) can be put into a spectral form too. Namely, if we define the coefficients $t_{nm}, v_{nm}, F_n^\Delta$ in such a way that

$$t(P) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} t_{nm} Y_{nm}(\vartheta_P, \lambda_P)$$
$$v(P) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} v_{nm} Y_{nm}(\vartheta_P, \lambda_P)$$
$$F^\Delta(\psi) = \frac{1}{2} \sum_{n=0}^{+\infty} (2n+1) F_n^\Delta P_n(\cos\psi),$$

i.e.

$$F_n^\Delta = \int_0^\pi F^\Delta(\psi) P_n(\cos\psi) \sin\psi d\psi \tag{14.29}$$
$$= \int_\Delta^\pi F(\psi) P_n(\cos\psi) \sin\psi d\psi,$$

then we have

$$t_{nm} = \frac{1}{2} F_n^\Delta v_{nm}, \tag{14.30}$$

Now we can easily use the concepts of Part I, Chap. 5 to find a mean square truncation error, defined as

$$\mathcal{TE}^2 = E\{t(P)^2\}, \tag{14.31}$$

where the average $E\{\ \}$ is computed according to Part I, Sect. 5.4.

The result is important in itself and we state it in the form of a lemma.

Lemma 1 (Mean square truncation error). *Let $v(P)$ be a field on the unit sphere with degree variances $\sigma_n^2(v)$, such that*

$$C_{vv}(P,Q)=\sum_{n=0}^{+\infty}\sum_{m=-n}^{n}\sigma_n^2(v)Y_{nm}(\vartheta_P,\lambda_P)Y_{nm}(\vartheta_Q,\lambda_Q). \tag{14.32}$$

Then the mean square truncation error for the convolution (14.24), is given by

$$\mathcal{TE}^2(\Delta)=\sum_{n=0}^{+\infty}(2n+1)\left(\frac{1}{2}F_n^{\Delta}\right)^2\sigma_n^2(v). \tag{14.33}$$

Proof. The proof is indeed trivial if we use the results of Part I, Sect. 5.6 because (14.30) implies

$$\sigma_n^2(t)=\left(\frac{1}{2}F_n^{\Delta}\right)^2\sigma_n^2(v) \tag{14.34}$$

so that

$$\begin{aligned}\mathcal{TE}^2(\Delta)=C_{tt}(P,P)&=\sum_{n=0}^{+\infty}(2n+1)\sigma_n^2(t)\\&=\sum_{n=0}^{+\infty}(2n+1)\left(\frac{1}{2}F_n^{\Delta}\right)^2\sigma_n^2(v).\end{aligned}$$

□

It is obvious, though worthwhile, to underline that

$$\lim_{\Delta\to\pi}\mathcal{TE}^2(\Delta)=0. \tag{14.35}$$

In fact if we assume that v has finite variance on S, i.e.

$$C_{vv}(P,P)=\sum_{n=0}^{+\infty}(2n+1)\sigma_n^2(v)<+\infty,$$

and that $|F_n^{\Delta}|$ are uniformly bounded because $F(\psi)$ is integrable over $[0,\pi]$, we see that we can pass to the limit under the series in (14.33), and since $F_n^{\Delta}\to 0$ when $\Delta\to\pi$, (14.35) follows.

We are able now to apply our results to the integrals contained in (14.21) and (14.22). Let us start with

$$T(P)=\frac{R}{4\pi}\int S(\psi_{PQ})[\Delta g(Q)+G_1(P,Q)]d\sigma_Q. \tag{14.36}$$

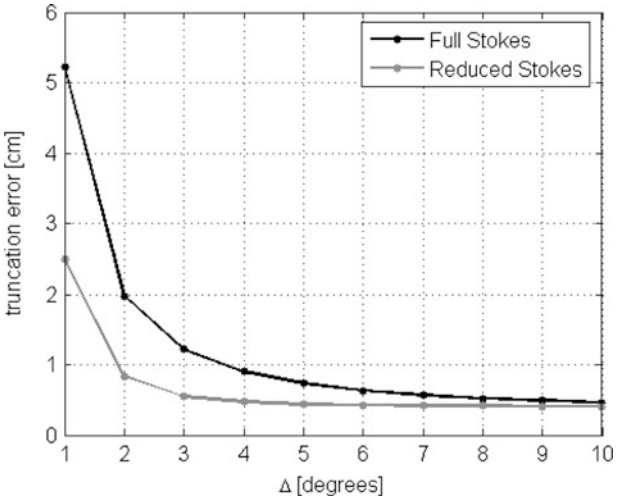


Fig. 14.2 The mean square truncation error $\mathcal{TE}(\Delta)$ for the height anomaly, according to the hypothesis of Example 1

We notice first of all that $G_1(P, Q)$ is typically a perturbation term, depending linearly on Δg but smaller than it, due to the presence of factors like $\frac{h_P - h_{Q_0}}{R}$. Accordingly, we can compute the mean square truncation error for T, referring only to the variability of Δg in (14.36). Then we need only to apply (14.33) to get

$$\mathcal{TE}^2(T;\Delta) = R^2 \sum_{n=0}^{+\infty} (2n+1) \left(\frac{1}{2} S_n^{\Delta} \right)^2 \sigma_n^2(\Delta g), \tag{14.37}$$

where

$$S_n^{\Delta} = \int_{\Delta}^{\pi} S(\psi) P_n(\cos\psi) \sin\psi \, d\psi. \tag{14.38}$$

The only warning in the use of (14.37) is that indeed the degrees 0 and 1 as always have zero degree variances; furthermore, since Δg is a residual anomaly, $\sigma_n^2(\Delta g)$ up to the degree M of the model used to reduce the original Δg, is just an error degree variance, in the sense explained in Part I, Sect. 5.7.

Example 1. In this example we assume that $\sigma_n^2(\Delta g) = 0$ up to degree $n = 180$, because we have subtracted a *true* global field up to this degree, and $\sigma_n^2(\Delta g)$ are those foreseen by the Tscherning-Rapp model, (see Part I, Sect. 5.9) when $n > 180$. In this way we can compute numerically (14.37) for different values of Δ and, expressing the result in meters, by considering $\zeta = \frac{T}{\gamma}$, we find the plot of Fig. 14.2.

We note that the level of 1 cm mean square truncation error in height anomaly is reached for $\Delta = 4°$.

An even better result can be obtained if we simultaneously subtract not only the Δg the first 180°, but also to the Stokes function its spectral development up to degree 180, i.e we compute (14.38) for a reduced Stokes function. In this case the 1 cm level of the truncation error is reached already at $\Delta = 2°$, as one can see again from Fig. 14.2.

As for the truncation error arising from the computation of the $G_1(P, Q)$ term, i.e.

$$G_1(P,Q) = \frac{1}{2}\frac{h_P - h_Q}{\overline{R}}\Delta g_Q + \frac{1}{2}\frac{h_P - h_Q}{\overline{R}}\frac{1}{4\pi}\int\left(\frac{\overline{R}}{\ell_{QQ'}}\right)^3 [\Delta g_Q - \Delta g_{Q'}]d\sigma_{Q'} \tag{14.39}$$

some numerical simulations prove that this is very small, at least if we choose $\Delta \geq 3°$. In fact, even a rough computation of order of magnitudes can show that in this case it goes below the 0.1 mGal level. This is due to the rather peaky shape of the kernel $\left(\frac{\overline{R}}{\ell_{QQ'}}\right)^3$. With $\Delta = 3°$ and with the rough estimates

$$O\left(\frac{h_P - h_Q}{\overline{R}}\right) = 10^{-3},\ O(\Delta g) = 10\,\text{mGal}$$

one finds from (14.39)

$$\begin{aligned}\mathcal{TE}(G_1) &\sim \frac{1}{2}10^{-3}\cdot 10\,\text{mGal} \\ &\cdot\frac{1}{4\pi}\int_0^{2\pi} d\alpha \int_\Delta^\pi \frac{2\sin\frac{\psi}{2}\cos\frac{\psi}{2}}{\left(2\sin\frac{\psi}{2}\right)^3}d\psi \sim \\ &\sim 10^{-2}\cdot\frac{1}{8}\left(\frac{1}{\sin\frac{\Delta}{2}} - 1\right)\text{mGal} \\ &= 0.047\,\text{mGal}\,.\end{aligned} \tag{14.40}$$

So we shall assume that G_1 can be computed form data up to an angular distance of 3°, without any further discussion.

Concluding, we can say that we have proved the applicability of Molodensky's formula, to the first order, to a local data set and we have learnt how to control the truncation error.

14.5 The Helmert Approach: A Short Review

During the last 20 years an old idea of Helmert (Martinec 1998), dating back to 1884, has been revitalized, revised and presented as a different approach to the determination of the *geoid*: the so-called *Helmert–Stokes approach*.

There is quite an extensive literature on this matter (Martinec 1998) and, although such an approach is out of the main line of reasoning of this book, we shortly illustrate it for the sake of completeness. In particular we shall follow Heck (Heck 2003a), because that presentation is in our opinion the most lucid and convincing.

The method attempts to perform a direct estimation of the anomalous potential outside and within the masses, down to the geoid level by subtracting first the influence of all the masses between geoid itself and the topographic surface, what we called in Part I, Chap. 4 the topographic correction. However, as we already know from Part I, Sect. 4.3, the full terrain correction has such an amplitude, that the magnitude of Δg is not reduced by applying it. For this reason, as well as to avoid to change the global mass of the earth depending on the correction applied, the idea of Helmert was to create an equivalent layer of condensed mass, disposed on a a sphere at some *compensation depth* D, i.e. with a radius $R_C = R - D$. The difference between the two effects is in fact the Helmert's correction, which by definition keeps the masses balanced.

In this sense, as observed in Heck (2003a), the method bears a similarity with the model of the isostatic correction, shortly illustrated in Part I, Remark 2; the main difference is in that Helmert's method substitutes the gravity of the *roots* (δH_r in Part I, Fig. 4.4) with that of an equivalent surface element of a mass which, using the notation of Part I, Sect. 4.3, is given by $(\delta_m - \delta_0)\delta H_r dS$. Since, due to the Airy-Heiskanen hypothesis, as matter of fact the root depth δH_r is related to the topographic height by

$$\delta_0 \cdot H = (\delta_m - \delta_0)\delta H_r,$$

we see that actually, at the compensation depth D, Helmert's method places a surface mass element corresponding to the mass of the topographic column squeezed on the compensating surface S_{R_c} at Q_c (see Fig. 14.3).

The idea then runs as follows: let Δg_{FA} be the full free air anomaly at P and $\Delta g_{TC}(P)$ the full terrain correction, computed integrating on the masses between the telluroid S and the reference surface S_R (taking the place of the ellipsoid in spherical approximation, that we are applying here); let further Δg_C be the *compensation* attraction created by the surface layer, with mass density $\delta_0 \cdot \delta H$, spread over the sphere S_{R_C}, at compensation depth $D = R - R_C$; then we compute the Helmert reduced anomalies

$$\Delta g_H = \Delta g_{FA} - (\Delta g_{TC} - \Delta g_C). \tag{14.41}$$

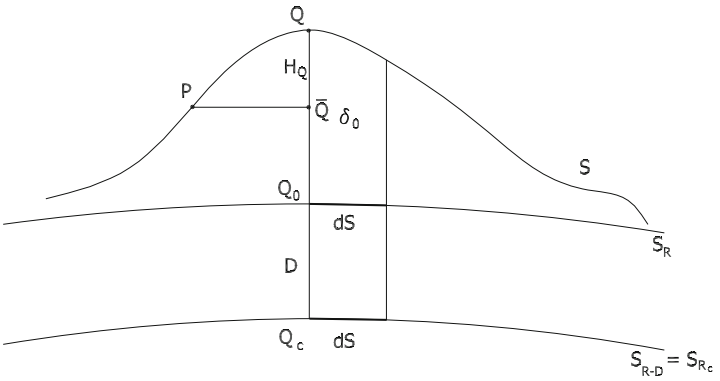


Fig. 14.3 The geometry of Helmert's correction with compensation depth D; note that $R_{\overline{Q}} = R + H_{\overline{Q}} = r_P = R + H_P$

With a good choice of the compensation depth D (typically between 20 and 30 km), it is known that Δg_H is smaller and smoother then Δg, and so better suited to downward continuation. Once Δg_H is reduced to the sphere S_R, the Stokes's formula can be legitimately applied to compute T_H on S_R and outside. Then the Helmert potential $T_{TC} - T_C$ has to be added, to retrieve the final solution

$$T = T_H + (T_{TC} - T_C). \tag{14.42}$$

The implementation of the method then relies on the computation of the following four quantities ($\psi = \psi_{PQ}$)

$$\begin{aligned} T_{TC}(P) = {} & \frac{4}{3}\pi G\delta_0 \frac{r_P^3 - R^3}{r_P} + \frac{1}{2} G\delta_0 \int d\sigma_{\overline{Q}} [r_Q \ell_{PQ} - r_P \ell_{P\overline{Q}} \\ & + 3 r_P \cos\psi (\ell_{PQ} - \ell_{P\overline{Q}}) + r_P^2 (3\cos^2\psi - 1) \\ & \cdot \log \frac{(r_Q - r_P \cos\psi) + \ell_{PQ}}{r_P(1 - \cos\psi) + \ell_{P\overline{Q}}}], \end{aligned} \tag{14.43}$$

$$T_C(P) = G \int \cdot \frac{\delta_0 H_Q}{\ell_{PQ_C}} R_C^2 \, d\sigma_{Q_C}, \tag{14.44}$$

$$\begin{aligned} \Delta g_{TC}(P) &= -\frac{\partial}{\partial r_P} T_{TC}(P) \\ &= -\frac{2}{r_P} T_{TC}(P) + \frac{G\delta_0}{r_P} \int d\sigma_Q \left[\frac{r_Q^3}{\ell_{PQ}} - \frac{R^3}{\ell_{PQ_0}} \right], \end{aligned} \tag{14.45}$$

$$\Delta g_C(P) = -\frac{\partial}{\partial r_P} T_C(P) = G\delta_0 \int \frac{H_Q [r_P - R_C \cos\psi]}{\ell_{PQ_C}} R_C^2 \, d\sigma_{Q_C}. \tag{14.46}$$

The interested reader can derive on its own such formulas, guided by the exercises at the end of the chapter.

Remark 1. The reasons why we shall not dwell anymore in the book on such an approach are several. First of all, despite its interpretation as one of the possible remove-restore techniques applicable to the computation of the geoid, it is often claimed in literature that Helmert's method is better founded on a physical ground. Yet for this to be true one has to believe that all the masses between S and the geoid have a constant density equal to the one we have chosen, for instance $\delta_0 = 2.67\,\mathrm{g\,cm^{-3}}$.

We have already solved in our Part I, Chaps. 3 and 4 all the main problems to reduce Δg_{FA} to a smaller and smoother signal on the sphere. In particular, as discussed in Part I, Sect. 4.3, the subtraction of a global model already copes with the long-wavelength (averaged) topographic effects and their isostatic compensation which, coming from below the surface, is always smooth in itself. As for the high-frequency part, we have accounted for it through the residual terrain correction (RTC) which has many advantages; first of all the masses in RTC always balance because the actual S is winding up and down around the average $\widetilde{S}$; secondly, possible errors due to variations in mass density are much less influential on the final result because they are affecting a part of the corrections of smaller size.

So, the only reason to look at a method like this could be related to the need of computing the geoid in areas where only little gravimetric material is available. In such areas in fact, for instance at present in most of the African continent, the performance of global models is weak, and having the possibility of smoothing Δg already by considering the topography only, can be of some advantage. Nevertheless one has to consider first of all that with poor gravimetric material, there will be in any case no possibility of getting a high resolution geoid with errors in the range of centimeters.

Furthermore already now and even more in the next future, global models up to degree 200–300 will be based on spatial observations. so they will help to bridge the holes in data at least with a spatial resolution of $\sim$80 km.

More comments in this point can be found in Sect. 15.7.

14.6 Exercises

Exercise 1. This exercise is just a preparation of the next in view of the proof of formula (14.43).

Prove by direct differentiation that, calling

$$I(P,Q)=\frac{1}{2}\Big[r_Q\ell_{PQ}-r_P\ell_{P\overline{Q}}+3r_P\cos\psi(\ell_{PQ}-\ell_{P\overline{Q}}) \tag{14.47}$$
$$+r_P^2(3\cos^2\psi-1)\log\frac{(r_Q-r_P\cos\psi)+\ell_{PQ}}{r_P(1-\cos\psi)+\ell_{P\overline{Q}}}\Big]$$

one has

$$\frac{\partial}{\partial r_Q}I(P,Q)=\frac{r_Q^2}{\ell_{PQ}}. \tag{14.48}$$

Furthermore, recalling that $\psi_{PQ}=\psi_{P\overline{Q}}, r_{\overline{Q}}=r_P$ (see Fig. 14.3), verify that

$$I(P,\overline{Q})\equiv 0,$$

so that the relation holds

$$I(P,Q)=I(P,\overline{Q})+\int_{R+H_{\overline{Q}}}^{R+H_Q}\frac{\partial}{\partial r_{Q'}}I(P,Q')dr_{Q'} \tag{14.49}$$
$$=\int_{R+H_P}^{R+H_Q}\frac{r_{Q'}^2}{\ell_{PQ'}}dr_{Q'},$$

Q' being a point running along the radius through Q.

Exercise 2. Observe that

$$T_{TC}(P)=G\delta_0\int d\sigma_{Q_0}\int_{r_{Q_0}}^{r_Q}\frac{r_{Q'}^2dr_{Q'}}{\ell_{PQ'}} \tag{14.50}$$
$$=G\delta_0\int d\sigma_{Q_0}\int_R^{r_{\overline{Q}}}\frac{r_{Q'}^2}{\ell_{PQ'}}dr_{Q'}+G\delta_0\int d\sigma_{Q_0}\int_{r_{\overline{Q}}}^{r_Q}\frac{r_{Q'}^2}{\ell_{PQ'}}dr_{Q'}$$

and, noting that $r_{\overline{Q}}=r_P=R+H_P$, prove (14.43), i.e., using (14.47),

$$T_{TC}(P)=\frac{4}{3}\pi G\delta_0\frac{r_P^3-R^3}{r_P}+G\delta_0\int d\sigma_{Q_0}I(P,Q). \tag{14.51}$$

(**Hint:** the first integral in (14.50) is given by (see Part I, Example 2 in Chap. 1, Sect. 3)

$$G\delta_0\int d\sigma_{Q_0}\int_R^{r_{\overline{Q}}}\frac{r_{Q'}^2}{\ell_{PQ'}}dr_{Q'}=\frac{G\delta_0\cdot\frac{4}{3}\pi(r_{\overline{Q}}^3-R^3)}{r_P}$$

for $r_P\geq r_{\overline{Q}}$. Then put $r_P=r_{\overline{Q}}$).

Exercise 3. Verify by direct differentiation the identity (Heck 2003a)

$$\frac{\partial}{\partial r_P}\frac{1}{\ell_{PQ}} = \frac{2}{r_P \ell_{PQ}} - \frac{1}{r_P r_Q^2}\frac{\partial}{\partial r_Q}\frac{r_Q^3}{\ell_{PQ}}. \tag{14.52}$$

Apply (14.52) to

$$\frac{\partial}{\partial r_P} T_{TC}(P) = \frac{\partial}{\partial r_P}\left(G\delta_0 \int d\sigma_{Q_0} \int_R^{r_Q} \frac{r_{Q'}^2}{\ell_{PQ'}} dr_{Q'}\right)$$

to prove (14.45).

Note that (14.44) and (14.46), don't need any particular development since they are respectively Newton's formula and its radial derivative, computed at a point outside the condensed masses because

$$r_P \geq R > R_C.$$

Chapter 15
The Analysis of Geodetic Boundary Value Problems in Linear form

15.1 Outline of the Chapter

Assume that S is a smooth surface, in the sense explained in Sect. 13.2, and that there is a function $u(P)$ harmonic in Ω (the exterior of S), regular at infinity, and we have performed a very large number of measurements that can be expressed as functionals of $u(P)$ at every point $P \in S$

$$F[u(P), P] = f(P), \tag{15.1}$$

then one can attempt to determine $u(P)$ by solving the BVP

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ F[u] = f & \text{on } S \\ u \to 0 & \text{at } \infty. \end{cases} \tag{15.2}$$

In our context $u(P)$ is in fact the anomalous gravity potential $T(P)$, which is actually related to the full gravity potential $W(P)$ through

$$T(P) = W(P) - U(P)\,; \tag{15.3}$$

$U(P)$, the normal potential, is a known function of the point P in space, where P is identified in terms of coordinates by means of a Cartesian frame centered to the reference ellipsoid. It is important to remember that (see Part I, (1.147))

$$O(T) \cong 10^{-5} O(U), \tag{15.4}$$

so that T can legitimately be considered as a perturbation of U.

In this chapter, S is either the earth topographic surface, or the telluroid (see Part I, Sect. 2.3), suitably smoothed by taking into account that our data $f(P)$

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0_15,

are not really given everywhere on the surface and that in any case what we are aiming at is only an approximation of the solution of (15.2) by means of a finite sum of spherical harmonics, namely a global model. In this context clearly S can be averaged over squares of some kilometers without increasing the approximation error, up to a maximum degree of a few thousands. This is particularly true when the influence of the uppermost thin layer of topographic masses is (approximately) accounted for by the residual terrain correction (cf. Part I, Chap. 4). Another warning is that, as we know, true global models are built by using other data than those referring to the boundary, in fact they are rather derived by space geodetic techniques, like ground satellite tracking, satellite-to-satellite tracking, satellite gradiometry etc. Here these data will be considered as known, since we concentrate on the BVP part only of this approximation procedure and we would like to know whether the procedure is stable, i.e. whether, if we use a certain norm for the data $\{f(P)\}$ and another norm for the solution $\{T(P)\}$, in order to be able to understand what is "small" and what is "large", to a small perturbation of data corresponds a small perturbation of the solution.

In this sense the theory that we shall outline in this chapter is a basis for the construction of the so-called *high resolution* earth gravity models, represented by a set of harmonic coefficients up to a maximum degree of some thousands. This can be done with or without the help of the knowledge of lower degree harmonics depending on what data we consider as boundary values.

Typical in this sense would be either the free air gravity anomaly or the gravity disturbance. The first, in a linearized version, writes

$$\Delta g(P) = \mathbf{e}_\gamma \cdot \nabla T(P) + \frac{\gamma'}{\gamma} T(P) \tag{15.5}$$

$$\left(\mathbf{e}_\gamma = \frac{\boldsymbol{\gamma}}{|\gamma|},\ \gamma' = \frac{\partial \gamma}{\partial h} \right).$$

The linearized equation (15.5) holds according to Molodensky's theory in the scalar version, where we know for each point of the boundary (ϑ, λ), W and g, (cf. Part I, Sect. 2.3, point 3). The second can be written

$$\delta g = \mathbf{e}_\gamma \cdot \nabla T\ ; \tag{15.6}$$

(15.6) applies when we assume to know beyond (ϑ, λ) also the ellipsoidal height h of the point P and the gravity modulus $g(P)$ (cf. Part I, Sect. 2.3, point 2).

In the first case we derive (15.5) by linearizing a free boundary BVP, or Molodensky's problem, where the ellipsoidal height $h(\vartheta, \lambda)$ of the boundary is unknown and in fact related to the known normal height $h^*(\vartheta, \lambda)$ by Bruns's relation (cf. Part I, (2.36))

$$h(\vartheta, \lambda) = h^*(\vartheta, \lambda) + \zeta\ ;\ \ \zeta = \frac{T(\vartheta, \lambda)}{\gamma(\vartheta, \lambda)}. \tag{15.7}$$

In the second case, (15.6) is derived by linearizing the expression of $|\mathbf{g}|$ on a known fixed boundary.

So, the surface where data (15.5) are given is a "smoothed" version of the telluroid, while for (15.6) we can think of a smoothed version of the actual topographic surface. In both cases we shall make on S the hypothesis that it is star-shaped, i.e. that it can be expressed in spherical coordinates by an equation of the form $r = R(\vartheta, \lambda)$; furthermore we shall assume that $R(\vartheta, \lambda)$ has bounded first and second derivatives, i.e. that S has a bounded inclination with respect to $\mathbf{e}_r$ and a bounded curvature. In order to speed up the notation of the section we shall use, through this chapter, the following symbols

$$\begin{aligned}
&\sigma \equiv (\vartheta, \lambda) = \text{ corresponding to a direction in space} \\
&\qquad\qquad \text{or a point on the unit sphere} \qquad (15.8)\\
&R_\sigma = R(\sigma) = R(\vartheta, \lambda) \\
&\nabla_\sigma = \mathbf{e}_\vartheta \frac{\partial}{\partial \vartheta} + \mathbf{e}_\lambda \frac{1}{\sin \vartheta} \frac{\partial}{\partial \lambda} \\
&\mathbf{n}_P = \text{ unit vector normal to } S \text{ at } P \\
&I = \text{ inclination of } S \text{ with respect to } \mathbf{e}_r \\
&\cos I = \mathbf{n}_P \cdot \mathbf{e}_r \\
&J = (\cos I)^{-1} \\
&d\sigma = \text{ unit sphere area element} \\
&dS = J R_\sigma^2 d\sigma = \text{ surface } S \text{ - area element}
\end{aligned}$$

Furthermore we shall use an index $+$ or $-$ to represent minimum and maximum of a certain quantity with respect to σ; so

$$\begin{aligned}
&R_+ = \max_\sigma R_\sigma, \; R_- = \min_\sigma R_\sigma, \; \delta R = R_+ - R_- \qquad (15.9)\\
&J_+ = \max_\sigma J = (\cos I_+)^{-1}
\end{aligned}$$

and so forth.

Finally, let us remark that we have used the notation

$$\mathbf{e}_\gamma \cdot \nabla u \cong \frac{\partial u}{\partial h} = u' \, ; \qquad (15.10)$$

in the sequel the same notation will be used as well for $\frac{\partial u}{\partial r} = u'$, when the context implies an unambiguous identification between the two alternatives.

All that given, in Sect. 15.2 we prepare the precise definitions of the BVP problems we are going to analyze, and of the spaces where the solution will be sought.

In Sect. 15.3 and in Sect. 15.4 we analyze the two main BVP's, namely Molodensky's problem and the fixed boundary BVP, proving theorems of existence and uniqueness under suitable conditions on the geometry of S (cf. Sansò and Venuti 2008). In Sect. 15.5 we start the discussion of the numerical implementation of an approximate solution of our BVP's; in particular we start from the traditional least squares method and we show how it compares with the classical Galerkin approach. Despite some simplifications, even Galerkin's equations are too complicated to allow for a direct numerical solution, although some numerical work has been done to study direct solutions on the surface S. So, following the geodetic tradition, some simplified iterative solutions have to be devised. These are illustrated in Sect. 15.6 (Sacerdote and Sansò 2010). Finally in Sect. 15.7 we briefly introduce new datasets relative to the gravity field, that space technology is recently providing. The use of such data sets can be done along the line of a solution of a BVP and for this reason they are shortly presented within this chapter.

15.2 A Precise Definition of the Two Main BVP's and of Their Solution Spaces

What is peculiar of this chapter is that in the rather large literature concerning geodetic BVP we shall choose for the data the $L^2(S)$ topology, because this is what is implicitly assumed in many approximation procedures, specially when we discretize S so that the $L^2(S)$ norm resembles a quite familiar sum of squares.

Correspondingly we expect $T' = \frac{\partial T}{\partial r}$ to be in $L^2(S)$ too so that for the solution a suitable norm could be that of $H^{1,2}(S)$, namely the one that guarantees that $|\nabla T|$ is in $L^2(S)$ too. This is essential if we want to build an approximation valid for gravity anomalies and deflections of the vertical, up to the boundary.

For technical reasons however we shall not use exactly the classical $L^2(S)$ and $H^{1,2}(S)$ norms but rather an equivalent form, namely (remember that we shall put an H in front of the symbols of our Hilbert spaces, to underline that we are dealing with spaces of harmonic functions) we shall put

$$v \in HL^2(S), \quad \|v\|_0^2 = \int v^2(R_\sigma, \sigma) R_\sigma d\sigma \tag{15.11}$$

$$u \in HH^{1,2}(S), \quad \|u\|_1^2 = \|r|\nabla u|\|_0^2 = \int |\nabla u(R_\sigma, \sigma)|^2 R_\sigma^3 d\sigma. \tag{15.12}$$

Let us notice that, based on the above remark, the norm $\| \ \|_0$ defined by (15.11) can be used for both harmonic functions defined through Ω, that admit a trace on S according to Cimmino (Cimmino 1952; Miranda 1970), or functions which are just defined on S, like the data f. The same is not true for the $\| \ \|_1$, (15.12), because ∇u implies also the knowledge of u'.

The target of this chapter is precisely to show that, assuming that data are bounded in $L^2(S)$, the solutions T of our geodetic BVP is bounded in $HH^{1,2}(S)$, i.e. that, under suitable conditions on R_σ, there is constant C such that

$$\|T\|_1 \leq C\,\|f\|_0 \tag{15.13}$$

where f is $-R_\sigma \Delta g$ or $-R_\sigma \delta g$, depending whether we treat the problem with data (15.5) or (15.6).

The *linearized Molodensky problem* with boundary values (15.5) is in fact the one we are still using for the determination of high resolution global models, yet it has more unfavourable mathematical properties due to the fact that in spherical approximation its solution is non-unique.

On the contrary the BVP with boundary values (15.6) is much easier to analyze and has superior stability properties. yet one could argue that the data for such a problem are not available. This is certainly true at the present time, however the possibility of a direct survey of the topographic surface from space by SAR and the nowadays common use of GPS together with gravimeters, providing the ellipsoidal height at every new point of gravity measurement, make this form of the geodetic BVP more and more important for the future.

A warning on the notation used in the chapter is that many times we need to define a not-better specified constant: for that we shall always use the symbol C, without necessarily implying that it is a specific constant assuming the same value in all cases.

To start to give the appropriate formulation of our problems we need here some preliminary propositions.

Proposition 1. *There are functions $\{Z_{\ell m}\}$ in $HL^2(S)$ such that, fixing a radius $\overline{R} > R_+$ and a sphere $\overline{S}$, with radius $\overline{R}$, encompassing S, $\forall u \in HL^2(S)$*

$$\begin{aligned} \langle Z_{\ell m}, u\rangle_0 &= \int Z_{\ell m}(R_\sigma,\sigma)u(R_\sigma,\sigma)R_\sigma\, d\sigma \\ &= \int_{\overline{S}} S_{\ell m}(\overline{R},\sigma)u(\overline{R},\sigma)d\sigma, \qquad \left(d\sigma = \frac{d\overline{S}}{\overline{R}^2}\right), \end{aligned} \tag{15.14}$$

with $S_{\ell m}$ the outer solid spherical harmonics of degree ℓ and order m.

Proof. Since $\operatorname{dist}(\overline{S}, S) = \overline{R} - R_+ > 0$, we have for the *Green function* of S

$$P \in \overline{S}, Q \in S\,; \quad |G_{n_Q}(P,Q)| \leq C. \tag{15.15}$$

Therefore, using (13.168) and the fact that $dS = JR_\sigma^2 d\sigma$, we have $\forall u \in HL^2(S)$

$$\int_{\overline{S}} u^2(P)d\sigma \leq C \int_S u^2(Q)dS \leq CJ_+R_+\|u\|_0^2. \tag{15.16}$$

On the other hand the linear functionals

$$L_{\ell m}(u) = \int_{\overline{S}} S_{\ell m} u d\sigma \tag{15.17}$$

are indeed bounded, since

$$|L_{\ell m}(u)|^2 \leq \int_{\overline{S}} S_{\ell m}^2 d\sigma \int_{\overline{S}} u^2 d\sigma \leq \frac{4\pi}{\overline{R}^{2\ell+2}} C J_+ R_+ \|u\|_0^2. \tag{15.18}$$

Therefore (15.14) is just Riesz theorem (Theorem 2) applied to $L_{\ell m}(u)$.

Let us underline that the functions $Z_{\ell m}$ so defined are in $HL^2(S)$, namely they are harmonic in the whole Ω and it is their trace on S that is used to verify the identity (15.14). □

Proposition 2. *Let us define*

$$V_L = Span\{Z_{\ell m}\,; |m| \leq \ell, \ell \leq L\} \tag{15.19}$$

and call $V_L^\perp$ the orthogonal complement of V_L in $HL^2(S)$; then

$$u \in V_L^\perp \Leftrightarrow u = O\left(\frac{1}{r^{L+2}}\right). \tag{15.20}$$

Furthermore $\{Z_{\ell m}\}$ is a system of linearly independent functions.

Proof. In fact if $u \in V_L^\perp$, i.e. $u \perp V_L$, we have (recall that $S_{\ell m}(\overline{R}, \sigma) = Y_{\ell m}(\sigma)\backslash\overline{R}^{\ell+1}$)

$$\overline{u}_{\ell m} = \frac{1}{4\pi} \int_{\overline{S}} Y_{\ell m}(\sigma) u(\overline{R}, \sigma) d\sigma = 0 \quad \forall m, \forall \ell \leq L \tag{15.21}$$

and viceversa. Hence, for $r \geq \overline{R}$,

$$u(r, \sigma) = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \overline{u}_{\ell m} \left(\frac{\overline{R}}{r}\right)^{\ell+1} Y_{\ell m}(\sigma), \tag{15.22}$$

confirming the statement (15.20). To prove that $\{Z_{\ell m}\}$ is a system of linearly independent functions, we note that

$$\sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m} = 0$$

implies

$$\begin{aligned} 0 &= \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m} \left\langle Z_{\ell m}, S_{jk} \right\rangle_0 \\ &= \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m} \int_{\overline{S}} \frac{Y_{\ell m}(\sigma)}{\overline{R}^{\ell+1}} \frac{Y_{jk}(\sigma)}{\overline{R}^{j+1}} d\sigma \\ &= \frac{4\pi}{\overline{R}^{2j+2}} a_{jk}, \qquad (|k| \le j,\ j \le L). \end{aligned}$$

□

Let us define for the moment the linearized Molodensky problem, modified to exploit the knowledge of low-degree harmonics up to order L, as

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ \mathbf{e}_\gamma \cdot \nabla T + \frac{\gamma'}{\gamma} T = -\Delta g - \sum_{\ell,m=0}^{L} a_{\ell m} \frac{1}{r} Z_{\ell m} & \text{on } S \\ T = O\left(\frac{1}{r^{L+2}}\right) & \text{for } r \to \infty. \end{cases} \tag{15.23}$$

The unknowns in (15.23) are both T and the coefficients $\{a_{\ell m}\}$. The boundary condition in (15.23) can be conveniently put into the perturbative form

$$\mathbf{r} \cdot \nabla T + 2T = f + \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} + \tag{15.24}$$

$$+ r(\mathbf{e}_r + \mathbf{e}_\gamma) \cdot \nabla T + \left(2 + r\frac{\gamma'}{\gamma}\right) T.$$

Proposition 3. *The equation (15.24) is perturbative in the sense that, calling as usual* $\boldsymbol{\nu}$ *the ellipsoidal normal,*

$$\begin{cases} \boldsymbol{\varepsilon} = \mathbf{e}_r + \mathbf{e}_\gamma \cong \mathbf{e}_r - \boldsymbol{\nu} \\ |\boldsymbol{\varepsilon}| \le \frac{1}{2} e^2 \end{cases} \tag{15.25}$$

$$(e^2 = \textit{ ellipsoid eccentricity})$$

and

$$\begin{cases} \eta = 2 + r\frac{\gamma'}{\gamma} \\ |\eta| \le 2e^2, \end{cases} \tag{15.26}$$

i.e. $\boldsymbol{\varepsilon}$ *and* η *can be taken as perturbation parameters, small to the first order in* e^2.

Proof. The estimate (15.25) is easily derived from the explicit expressions of $\boldsymbol{\nu}(\sigma)$ and $\mathbf{e}_r(\sigma)$, as functions of ellipsoidal colatitude $\overline{\vartheta}$ and longitude λ, namely

$$\boldsymbol{\nu}(\sigma) = \begin{vmatrix} \sin\overline{\vartheta}\cos\lambda \\ \sin\overline{\vartheta}\sin\lambda \\ \cos\overline{\vartheta} \end{vmatrix};$$

$$\mathbf{e}_r(\sigma) = \frac{1}{\sqrt{\sin^2\overline{\vartheta} + (1-e^2)^2\cos^2\overline{\vartheta}}} \begin{vmatrix} \sin\overline{\vartheta}\cos\lambda \\ \sin\overline{\vartheta}\sin\lambda \\ (1-e^2)\cos\overline{\vartheta} \end{vmatrix}.$$

In fact, by using an approximation to the order of e^2, we have

$$\mathbf{e}_r \sim (1 + e^2\cos^2\overline{\vartheta})\boldsymbol{\nu} - e^2 \begin{vmatrix} 0 \\ 0 \\ \cos\overline{\vartheta} \end{vmatrix}$$

i.e.

$$|\mathbf{e}_r - \boldsymbol{\nu}| \sim e^2|\cos\overline{\vartheta}| \cdot \left| \begin{vmatrix} \cos\overline{\vartheta}\sin\overline{\vartheta}\cos\lambda \\ \cos\overline{\vartheta}\sin\overline{\vartheta}\sin\lambda \\ \cos^2\overline{\vartheta} - 1 \end{vmatrix} \right|$$

$$= e^2|\cos\overline{\vartheta}|\sqrt{\cos^2\overline{\vartheta}\sin^2\overline{\vartheta} + \sin^4\overline{\vartheta}} = e^2|\cos\overline{\vartheta}\sin\overline{\vartheta}| \le \frac{1}{2}e^2$$

The estimate (15.26) is calculated form the approximate expression (see Part I, (2.122))

$$r\frac{\gamma'}{\gamma} \cong -\left(\frac{r}{\mathcal{M}} + \frac{r}{\mathcal{N}}\right) - 2\frac{\omega^2 r}{\gamma},$$

making the computation up to $O(e^2)$.

Remember that $\mathcal{M}$ and $\mathcal{N}$ are respectively the radius of curvature of the meridian and the grand normal already met in Part I, (1.206) and (1.137).

In fact, disregarding the height of the point on the surface S which gives a smaller contribution, one can write

$$r = \mathcal{N}[\sin^2\overline{\vartheta} + (1-e^2)^2\cos^2\overline{\vartheta}]^{(1/2)} \sim \mathcal{N}(1 - e^2\cos^2\overline{\vartheta})$$

$$\frac{\mathcal{N}}{\mathcal{M}} = \frac{1 - e^2\cos^2\overline{\vartheta}}{1-e^2} \cong 1 + e^2\sin^2\overline{\vartheta}$$

$$\frac{\omega^2 r}{\gamma} \cong \frac{1}{2}e^2;$$

the last estimate is done on a pure numerical basis with r equal to the mean radius of the earth. Accordingly one finds

$$\left| r\frac{\gamma'}{\gamma} + 2 \right| \sim e^2 |\sin^2 \overline{\vartheta} - 2\cos^2 \overline{\vartheta} + 1| = e^2 |2\sin^2 \overline{\vartheta} - \cos^2 \overline{\vartheta}| \leq 2e^2.$$

□

We are now able to give the definiton of the linearized Molodensky problem in perturbative form.

Definition 1. We say that the linearized Molodensky problem is to find the potential T and numbers $\{a_{\ell m}; 0 \leq \ell \leq L,\ |m| \leq \ell\}$ such that, denoting $rT' = \mathbf{r}\cdot\nabla T$,

$$\begin{cases} \Delta T = 0 \qquad \text{in } \Omega \\ rT' + 2T = f + \displaystyle\sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m} + r\boldsymbol{\varepsilon}\cdot\nabla T + \eta T \text{ on } S \\ T = O\left(\frac{1}{r^{L+2}}\right),\ r \to \infty. \end{cases} \tag{15.27}$$

As it is easy to verify, comparing with (15.23) and (15.24), in (15.27) we have put

$$f = -r\Delta g. \tag{15.28}$$

For future reference we note that, denoting

$$A_1 = r\left(-\frac{\partial}{\partial h} + \frac{\gamma'}{\gamma}\right) \tag{15.29}$$

the Molodensky boundary operator, we have used in (15.27) the perturbative form

$$A_1 = A_{1S} + D_1 \equiv \left(-\frac{\partial}{\partial r} - 2\right) + (r\boldsymbol{\varepsilon}\cdot\nabla + \eta) \tag{15.30}$$

$$\boldsymbol{\varepsilon} = \mathbf{e}_r - \boldsymbol{\nu},\ \eta = \frac{2}{r} + \frac{\gamma'}{\gamma}. \tag{15.31}$$

In a similar way the fixed boundary BVP in a linearized form can be written as

$$\begin{cases} \Delta T = 0 \qquad \text{in } \Omega \\ \mathbf{e}_\gamma \cdot \nabla T = \delta g \quad \text{on } S \\ T = O\left(\frac{1}{r}\right) \quad r \to \infty. \end{cases} \tag{15.32}$$

Note that we haven't introduced into (15.32) the knowledge of a certain number of harmonics of low degree and the corresponding unknowns $\{a_{\ell m}\}$; the reason is simply that (15.32) can be very easily analyzed without introducing such an artifice, what is not possible for problem (15.27).

Paralleling the Definition 1 we can put (15.32) too into a perturbative form, by exploiting (15.25).

Definition 2. The linearized fixed boundary BVP in perturbative form is to find a potential T in Ω satisfying

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ rT' = f + r\boldsymbol{\varepsilon} \cdot \nabla T & \text{on } S \\ T = O\left(\frac{1}{r}\right) & \to \infty. \end{cases} \tag{15.33}$$

Also here one finds that in (15.20) we have put

$$f = -r\delta g. \tag{15.34}$$

15.3 The Analysis of the Linearized Molodensky Problem

The results of this paragraph and of the next are based on the work (Sansò and Venuti 2008). The analysis of this problem can be performed basically in two steps. First of all we define a simpified problem, without perturbative terms, and we completely analyze it. Then we go back to the original form (15.27) and we get the desired result.

Definition 3 (simple Molodensky's problem). The *simple Molodensky problem* or *Molodensky's problem in spherical approximation* is to find $\{T, a_{\ell m}\}$ such that

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ rT' + 2T = f + \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m} & \text{on } S \\ T = O\left(\frac{1}{r^{L+2}}\right) & r \to \infty. \end{cases} \tag{15.35}$$

To proceed with the analysis of (15.35) we need a result which is adapted from Hörmander (cf. Hörmander 1976) to the specific star-shaped geometry used here.

Theorem 1 (energy integral). *Let T be harmonic in Ω and satisfy the equation*

$$rT' + \alpha T = \mathbf{r} \cdot \nabla T + \alpha T = v\,; \tag{15.36}$$

then v is harmonic in Ω too. Furthermore assume that $v \in HL^2(S)$, then T satisfies the identity

$$\|T\|_1^2 + (1 - 2\alpha) \int d\Omega |\nabla T|^2 = 2 \int_S v T_n dS, \tag{15.37}$$

with $T_n = \frac{\partial T}{\partial n}$ and $\mathbf{n}$ the exterior normal of S.

Proof. From (15.36) we derive by differentiation

$$\Delta v = \mathbf{r} \cdot \nabla(\Delta T) + (\alpha + 1)\Delta T = 0$$

proving that v is harmonic in Ω too. Now let T be harmonic in Ω; note that the following identity holds

$$\begin{aligned} &\nabla \cdot [(\mathbf{r} \cdot \nabla T + \alpha T)\nabla T] = [(\mathbf{r} \cdot \nabla)\nabla T] \cdot \nabla T + (\alpha + 1)|\nabla T|^2 \quad (15.38) \\ &\equiv \frac{1}{2} r \frac{\partial}{\partial r}(|\nabla T|^2) + (\alpha + 1)|\nabla T|^2. \end{aligned}$$

Remember that, to apply Gauss' theorem, we must consider that the normal $\mathbf{n}$ to S is pointing in Ω. So by integrating (15.38) over Ω, with $d\Omega = r^2 dr d\sigma$, we find

$$\begin{aligned} &-\int_S (\mathbf{r} \cdot \nabla T + \alpha T)\mathbf{n} \cdot \nabla T dS = -\int_S v T_n dS \quad (15.39) \\ &= \frac{1}{2} \int d\sigma \int_{R_\sigma}^{+\infty} dr r^3 \frac{\partial}{\partial r} |\nabla T|^2 + (\alpha + 1) \int |\nabla T|^2 d\Omega \\ &= \frac{1}{2} \int d\sigma R_\sigma^3 |\nabla T|^2 + \left(\alpha - \frac{1}{2}\right) \int |\nabla T|^2 d\Omega. \end{aligned}$$

Rearranging we get (15.37). □

Corollary 1. *Assume that* $\alpha \leq \frac{1}{2}$ *and* $v \in HL^2(S)$*, then* $T \in HH^{1,2}(S)$ *and*

$$\|T\|_1 \leq 2J_+ \|v\|_0; \quad (15.40)$$

in particular (15.40) holds for $\alpha = 0$*, i.e. for* $v = rT'$.

Assume viceversa that $\alpha > \frac{1}{2}$*, then*

$$\|T\|_1 \leq (2\alpha - 1)J_+ \|T\|_0 + 2J_+ \|v\|_0, \quad (15.41)$$

meaning that if one can prove that $T \in HL^2(S)$ *then we have* $T \in HH^{1,2}(S)$ *too.*

Proof. Note that, by the Schwarz inequality, whatever is v, the following inequality holds

$$\begin{aligned} |2\int_S v T_n dS| &= 2|\int_S v T_n R_\sigma^2 J d\sigma| \quad (15.42) \\ &\leq 2J_+ \left\{\int_S v^2 R_\sigma d\sigma\right\}^{(1/2)} \left\{\int_S T_n^2 R_\sigma^3 d\sigma\right\}^{(1/2)} \\ &\leq 2J_+ \|v\|_0 \|T\|_1. \end{aligned}$$

So if $(1-2\alpha) \geq 0$, from (15.37) and (15.42) we find

$$\|T\|_1^2 \leq |2\int_S vT_n dS| \leq 2J_+ \|v\|_0 \|T\|_1,$$

proving (15.40).

If, on the contrary, $1-2\alpha < 0$, we have from (15.37)

$$\begin{aligned} \|T\|_1^2 &= -(2\alpha-1)\int_S T\,T_n dS + 2\int_S vT_n dS \\ &= \int_S [-(2\alpha-1)T + 2v]T_n R_\sigma^2 J d\sigma; \end{aligned} \tag{15.43}$$

if we apply the Schwarz inequality to (15.43) we get (15.41). □

We are able now to proceed in the analysis of (15.35).

Proposition 4. *The simple Molodensky problem (15.35) is equivalent to the* modified Dirichlet problem

$$\begin{cases} \Delta v = 0 & \text{in } \Omega \\ v|_S = f + \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m} & \text{on } S \\ v = 0\left(\frac{1}{r^{L+2}}\right), \end{cases} \tag{15.44}$$

with

$$v = rT' + 2T \tag{15.45}$$

on condition that

$$L \geq 1 \tag{15.46}$$

Proof. If T is harmonic, v is harmonic too in force of Theorem 1. That the boundary condition in (15.44) is satisfied is tautological, given the definition (15.45). Furthermore, if $T = O\left(\frac{1}{r^{L+2}}\right)$, recalling Proposition 2, we must have for $r \geq \overline{R}$

$$T = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \overline{T}_{\ell m} \left(\frac{\overline{R}}{r}\right)^{\ell+1} Y_{\ell m}(\sigma), \tag{15.47}$$

so that

$$v = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} -(\ell-1)\overline{T}_{\ell m} \left(\frac{\overline{R}}{r}\right)^{\ell+1} Y_{\ell m}(\sigma), \tag{15.48}$$

and the third of (15.44) is satisfied.

Viceversa if v satisfies (15.44) with $L \geq 1$, we can reverse (15.47) and (15.48) in the sense that from the known development ($r \geq \overline{R}$)

$$v = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \overline{v}_{\ell m} \left(\frac{\overline{R}}{r}\right)^{\ell+1} Y_{\ell m}(\sigma) \tag{15.49}$$

we derive in $r \geq \overline{R}$, for T, the expression

$$T = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \frac{\overline{v}_{\ell m}}{\ell-1} \left(\frac{\overline{R}}{r}\right)^{\ell+1} Y_{\ell m}(\sigma) \tag{15.50}$$

which shows that in that region T is harmonic too and furthermore it satisfies the third of (15.44). Now from the identity

$$r\frac{\partial}{\partial r}(\Delta T) + 4\Delta T = \Delta v = 0 \text{ in } \Omega$$

multiplied by r^3, we can write

$$\frac{\partial}{\partial r}(r^4 \Delta T) = 0, \tag{15.51}$$

which integrated between r and $\overline{R}$, considering that $\Delta T|_{r=\overline{R}} = 0$, gives

$$r^4 \Delta T = 0, \ R_\sigma \leq r \leq \overline{R}. \tag{15.52}$$

Therefore $\Delta T = 0$ in the whole of Ω.

Note that critical in our reasoning is the fact we can never have $\ell = 1$ in (15.50), because for the smallest value of ℓ we have $L + 1 \geq 2$.

In fact, it is easy to see that there can be no function T which simultaneously satisfies

$$\Delta T = 0, \ rT' + 2T = \frac{Y_{1m}}{r^2}.$$

In addition, whatever is the first degree spherical harmonic $\frac{Y_{1m}}{r^2}$, one has

$$\left(r\frac{\partial}{\partial r} + 2\right)\frac{Y_{1m}}{r^2} \equiv 0,$$

i.e. there is no one-to-one correspondence between T and v, when first degree spherical harmonics are still present. This is avoided by condition (15.46).

Finally, we note that, when $L \geq 1$, we can write from (15.45)

$$\frac{\partial}{\partial r}(r^2 T) = rv \tag{15.53}$$

which integrates to

$$T(r,\sigma) = -\frac{1}{r^2}\int_r^{+\infty} sv(s,\sigma)ds. \tag{15.54}$$

Again the fact that $v = O\left(\frac{1}{r^3}\right)$, at least, guarantees the convergence of the integral in (15.54). □

Proposition 5. *Let us call w the solution of the* Dirichlet problem

$$\begin{cases} \Delta w = 0 \text{ in } \Omega \\ w = f \text{ on } S\,; \end{cases} \tag{15.55}$$

that $w \in HL^2(S)$ exists and is unique, when $f \in L^2(S)$, is a theorem by Cimmino that we don't prove here (cf. Cimmino 1952).

Then the solution of (15.44) is given by

$$v = -P_{V_L^\perp} w \tag{15.56}$$

with

$$P_{V_L^\perp} = I - P_{V_L} \tag{15.57}$$

the orthogonal projector on $V_L^\perp$ in $HL^2(S)$, and the solution of the simple Molodensky problem T is given by (15.54).

Proof. Equation 15.44 can be written as

$$\begin{cases} \Delta\left(v - \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m}\right) = 0 \quad \text{in } \Omega \\ v - \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} = f \qquad \text{on } S, \end{cases} \tag{15.58}$$

showing that we must have

$$v - \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} = w. \tag{15.59}$$

Since the third of (15.44) is equivalent to $v \in V_L^{\perp}$ (see Proposition 2), it is enough to apply $P_{V_L^{\perp}}$ to both members of (15.59) to get (15.56). □

We note also that (15.59) determines $\{a_{\ell m}\}$ too, since

$$\sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} = v - w = -P_{V_L} w \tag{15.60}$$

and $\{Z_{\ell m}\}$ are linearly independent.

We note as well that (15.56) implies the important relation

$$\|v\|_0 \leq \|w\|_0 = \|f\|_0 \,; \tag{15.61}$$

because the orthogonal projection of an element of a Hilbert space has always norm not larger than the projected vector.

Before we close the analysis of the simple Molodensky problem, we still need another technical result which we formulate as a proposition.

Proposition 6. *Let u be any function in $V_L^{\perp}$, i.e. $u = O\left(\frac{1}{r^{L+2}}\right)$ when $r \to \infty$; assume further that $u \in HH^{1,2}(S)$, i.e. that $\|u\|_1 < +\infty$; then the following inequality holds*

$$\|u\|_0 \leq C_{0L} \|R_\sigma u'\|_0 \leq C_{0L} \|u\|_1, \tag{15.62}$$

with

$$C_{0L} = J_+ \left(\frac{\delta R}{R_+} + \frac{2}{L+2} \right) \equiv J_+ C_L \,; \tag{15.63}$$

see (15.9) for the meaning of symbols.

Proof. Put

$$u_+ = u(R_+, \sigma) \tag{15.64}$$

and note that

$$u|_S = u(R_\sigma, \sigma) = u|_S - u_+ + u_+$$

so that one can write

$$\|u\|_0 \leq \|u_+\|_0 + \|u - u_+\|_0. \tag{15.65}$$

Note that for $r = R_+$ one can put

$$u_+ = \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} u_{\ell m}^+ Y_{\ell m}(\sigma) \tag{15.66}$$

and that for $r > R_+$ one has

$$ru' = -\sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} u_{\ell,m}^+ (\ell+1) \left(\frac{R_+}{r}\right)^{\ell+1} Y_{\ell m}(\sigma). \tag{15.67}$$

A direct computation shows that

$$\|u_+\|_0^2 = \int d\sigma R_\sigma u_+^2 \leq R_+ 4\pi \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} u_{\ell m}^{+^2} \tag{15.68}$$

$$\leq R_+ 4\pi \frac{2L+3}{(L+2)^2} \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \frac{(\ell+1)^2}{2\ell+1} u_{\ell m}^{+^2};$$

similarly

$$\int d\sigma \int_{R_+}^{+\infty} (ru')^2 dr = \int_{\Omega_+} (u')^2 d\Omega$$

$$= (4\pi R_+) \sum_{\ell=L+1}^{+\infty} \sum_{m=-\ell}^{\ell} \frac{(\ell+1)^2}{2\ell+1} u_{\ell m}^{+^2}. \tag{15.69}$$

Comparing (15.68) and (15.69), noticing that $\frac{(2L+3)}{(L+2)^2} < \frac{2}{L+2}$, we derive

$$\|u_+\|_0^2 \leq \frac{2}{L+2} \int_{\Omega_+} (u')^2 d\Omega. \tag{15.70}$$

On the other hand

$$|u_+ - u|_S|^2 = |u(R_+, \sigma) - u(R_\sigma, \sigma)|^2 = \left| \int_{R_\sigma}^{R_+} u' dr \right|^2 \tag{15.71}$$

$$\leq \int_{R_\sigma}^{R_+} r^2 (u')^2 dr \int_{R_\sigma}^{R_+} \frac{1}{r^2} dr = \frac{R_+ - R_\sigma}{R_\sigma R_+} \int_{R_\sigma}^{R_+} r^2 u'^2 dr.$$

Multiplying (15.71) by R_σ and integrating on $d\sigma$ we obtain

$$\|u_+ - u\|_0^2 \leq \frac{\delta R}{R_+} \int_{\Omega \setminus \Omega_+} (u')^2 d\Omega. \tag{15.72}$$

So, going back to (15.65) and applying the *Cauchy inequality*, we get

$$\begin{aligned}\|u\|_0^2 &\leq \left\{\sqrt{\frac{2}{L+2}}\left[\int_{\Omega_+}(u')^2 d\Omega\right]^{(1/2)} + \sqrt{\frac{\delta R}{R_+}}\left[\int_{\Omega\setminus\Omega_+}(u')^2 d\Omega\right]^{(1/2)}\right\}^2 \\ &\leq C_L\left[\int_{\Omega_+}(u')^2 d\Omega + \int_{\Omega\setminus\Omega_+}(u')^2 d\Omega\right] \\ &= C_L\|u'\|^2_{L^2(\Omega)} \leq C_L\|\nabla u\|^2_{L^2(\Omega)}. \end{aligned} \tag{15.73}$$

On the other hand

$$\begin{aligned}\int_\Omega |\nabla u|^2 d\Omega &= -\int_S u u_n dS = -\int u u_n J R_\sigma^2 d\sigma \\ &\leq J_+\left(\int u^2 R_\sigma d\sigma\right)^{(1/2)}\left(\int u_n^2 R_\sigma^3 d\sigma\right)^{(1/2)} \leq J_+\|u\|_0 \cdot \|u\|_1. \end{aligned} \tag{15.74}$$

By using (15.74) into (15.73) and simpifying $\|u\|_0$, we get (15.62). □

We are ready now to derive the sought result for the simple Molodensky problem.

Theorem 2 (simple Molodensky's problem). *The solution of the problem (15.35), explicitly provided by formula (15.52) with v defined in (15.45) and satisfying the inequality (15.61), is such that, if*

$$4J_+C_{0L} < 1, \tag{15.75}$$

then

$$\|T\|_1 \leq C_{1L}\|f\|_0 \tag{15.76}$$

with

$$C_{1L} = \frac{2J_+}{1-4J_+C_{0L}}. \tag{15.77}$$

Proof. Write (15.45) as

$$rT' = v - 2T \tag{15.78}$$

to the effect that (use (15.61) and (15.62) with $u = T$)

$$\begin{aligned}\|R_\sigma T'\|_0 &\leq \|v\|_0 + 2\|T\|_0 \\ &\leq \|f\|_0 + 2\|T\|_0 \\ &\leq \|f\|_0 + 2C_{0L}\|T\|_1. \end{aligned} \tag{15.79}$$

Now recall (15.40), basically claiming that, when $\alpha = 0$,

$$\|T\|_1 \leq 2J_+ \|R_\sigma T'\|_0 \tag{15.80}$$

and combine with (15.79) to get (15.76) and (15.77). □

At this point we are able to pass to analyze the linearized Molodensky problem, that we shall consider as written in the perturbative form (15.27).

Theorem 3. *The solution of the linearized Molodensky problem (15.27) exists is unique in $HH^{1,2}(S)$ and such that*

$$\|T\|_1 \leq C_{2L} \|f\|_0 \tag{15.81}$$

with

$$C_{2L} = C_{1L}[1 - C_{1L}(\varepsilon_+ + C_{0L}\eta_+)]^{-1}, \tag{15.82}$$

where $\varepsilon = |\boldsymbol{\varepsilon}|$, if the condition

$$2C_{0L} < \frac{1 - 2\varepsilon_+ J_+}{J_+(2 + \eta_+)} \tag{15.83}$$

is satisfied.

Furthermore, under the same condition (15.83), the simple iterative sequence

$$\begin{aligned} rT'_{(n+1)} + 2T_{(n+1)} = f + \sum_{\ell,m=0}^{L} a_{\ell m}^{(n+1)} Z_{\ell m} \\ + r\boldsymbol{\varepsilon} \cdot \nabla T_{(n)} + \eta T_{(n)} \end{aligned} \tag{15.84}$$

converges to the solution of (15.27) in $HH^{1,2}(S)$.

Proof. From the equation

$$rT' + 2T = f + \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} + r\boldsymbol{\varepsilon} \cdot \nabla T + \eta T \tag{15.85}$$

and (15.76) of Theorem 2 we derive

$$\begin{aligned} \|T\|_1 &\leq C_{1L} \|f + r\boldsymbol{\varepsilon} \cdot \nabla T + \eta T\|_0 \\ &\leq C_{1L} \|f\|_0 + C_{1L}\varepsilon_+ \|R_\sigma |\nabla T|\|_0 + C_{1L}\eta_+ \|T\|_0. \end{aligned} \tag{15.86}$$

With the help of (15.62) and (15.86) becomes

$$\|T\|_1 \leq C_{1L} \|f\|_0 + C_{1L}\varepsilon_+ \|T\|_1 + C_{1L}\eta_+ C_{0L} \|T\|_1. \tag{15.87}$$

Reordering, we see that if condition

$$C_{1L}(\varepsilon_+ + \eta_+ C_{0L}) < 1 \tag{15.88}$$

is satisfied, then (15.81) and (15.82) hold true.

Recalling (15.77) we verify that (15.88) is equivalent to (15.83).

Moreover, let us re-write (15.85) in the form

$$-A_{1S}T - \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} \equiv rT' + 2T - \sum_{\ell,m=0}^{L} a_{\ell m} Z_{\ell m} \tag{15.89}$$
$$= f + D_1 T \equiv f + r\boldsymbol{\varepsilon} \cdot \nabla T + \eta T$$

By means of Theorem 2, if L is such as to satisfy (15.75) and (15.89) can be written, after multiplying by the projection operator $P_{V_L^\perp}$ (cf. (15.57)) and noting that it has to be $P_{V_L^\perp} T = T$ because of the third of (15.27), as well as $P_{V_L^\perp} A_{1S} T = A_{1S} P_{V_L^\perp} T = A_{1S} T$

$$T = -A_{1S}^{-1} P_{V_L^\perp} f - A_{1S}^{-1} P_{V_L^\perp} D_1 T. \tag{15.90}$$

Equation 15.90 is meaningful because A_{1S} is indeed invertible if we restrict its range to $V_L^\perp$.

As we can easily understand the condition (15.88) implies, for the operator $A_{1S}^{-1} P_{V_L^\perp} D_1$, which transforms $HH^{1,2}(S)$ into $HH^{1,2}(S)$,

$$\| A_{1S}^{-1} P_{V_L^\perp} D_1 \| \leq C_{1L}(\varepsilon_+ + \eta_+ C_{0L}) < 1. \tag{15.91}$$

So (15.88) becomes the condition that $A_{1S}^{-1} P_{V_L^\perp} B_1$ is a contraction, which is well-known to be solvable by simple iteration. This proves (15.84). □

It is interesting to observe that the updating at each step of the constants $\{a_{\ell m}\}$ is necessary to implement the action of $P_{V_L^\perp}$, i.e. to guarantee that the known term of (15.84) at step n is depurated of the component on V_L, so that $T_{(n+1)}$ is known to keep on the correct asymptotic behaviour $T_{(n+1)} = O\left(\frac{1}{r^{L+2}}\right)$.

15.4 The Analysis of the Linearized Fixed Boundary BPV

We can now switch to the discussion of the fixed-boundary BVP in linearized form, where the observation equations on the boundary S are as in (15.6).

In analogy with Definition 3, we can introduce here too the linearized problem in spherical approximation.

Definition 4 (simple Hotine's problem). The *simple Hotine problem*, or *fixed boundary problem in spherical approximation*, is to find T such that

$$\begin{cases} \Delta T = 0 & \text{in } \Omega \\ rT' = f & \text{on } S \\ T = O\left(\frac{1}{r}\right) & r \to \infty. \end{cases} \tag{15.92}$$

Theorem 4. *If $f \in L^2(S)$, the simple Hotine problem has one and only one solution in $HH^{1,2}(S)$ satisfying the inequality*

$$\|T\|_1 \leq 2J_+ \|f\|_0. \tag{15.93}$$

Proof. In analogy to what we did for the simple Molodensky problem, we first transform (15.92) into an equivalent Dirichlet problem

$$\begin{cases} \Delta v = 0 & \text{in } \Omega \\ v = f & \text{on } S \\ v = O\left(\frac{1}{r}\right) & r \to \infty. \end{cases} \tag{15.94}$$

where

$$v = rT'. \tag{15.95}$$

We note that (15.95) can indeed be inverted providing T as

$$T(r,\sigma) = -\int_r^{+\infty} \frac{1}{s} v(s,\sigma) ds. \tag{15.96}$$

Since there is one and only one v, solution of (15.94), there is one and only one T solution of (15.92), given by (15.96).

Moreover, this solution satisfies (15.93), which is nothing but the energy integral theorem (see (15.40)) applied in this case with $\alpha = 0$. □

Theorem 5. *If $f \in L^2(S)$, the linearized fixed boundary BVP (15.33) has one and only one solution T in $HH^{1,2}(S)$, satisfying*

$$\|T\|_1 \leq C_{3L} \|f\|_0, \tag{15.97}$$

where

$$C_{3L} = 2J_+(1 - 2\varepsilon_+ J_+)^{-1},$$

on condition that

$$2\varepsilon_+ J_+ < 1. \tag{15.98}$$

Proof. We apply (15.77)–(15.74) obtaining

$$\|T\|_1 \leq 2J_+\|f\|_0 + 2J_+\varepsilon_+\|T\|_1 \,; \tag{15.99}$$

if condition (15.98) is satisfied, (15.97) is proved with

$$C_{3L} = 2J_+(1 - 2\varepsilon_+ J_+)^{-1}.$$

□

Remark 1. There is no need to say that defining $A_{2S} = -\frac{\partial}{\partial r}$ and $D_2 = r\boldsymbol{\varepsilon} \cdot \nabla$, the condition (15.98) guarantees that, similarly to (15.84), the iterative scheme

$$-A_{2S}T_{n+1} = f + D_2 T_n$$

is convergent in $HH^{1,2}(S)$.

Remark 2. Already the ease with which we produce the result of Theorem 5 as compared with the difficulty, or at least the complicacy, in proving Theorem 3, is a clear symptom of the superiority of the linearized fixed boundary BVP, with respect to the linearized Molodensky problem. As a matter of fact, the conditions under which we are able to guarantee the well-posedness (existence, uniqueness and stability of the solution) of the latter are more demanding than for the former.

In fact, remember that to prove Theorem 3 we assumed to know already the harmonic coefficients of the asymptotic development of T up to degree L.

In Exercises 2–4 the reader is invited to relate the conditions of validity of Theorems 3 and 5 to the geometry of the boundary.

15.5 From Least Squares to Galerkin's Method

Now that existence and stability of problems (15.27) and (15.32) have been studied, we would like to implement a numerical method to approximate the solution.

This can be done by constructing some finite dimensional subspace of $HH^{1,2}(S)$, where we can look for a model potential $T_M(r,\sigma)$ in such a way that the corresponding boundary values $f_M(\sigma)$ do approximate the data $f(\sigma)$ in $L^2(S)$; this is basically what can be called the least squares method in a Hilbert space.

Note that if we can reasonably guarantee that $f_M \to f$ in $L^2(S)$, Theorems 3 and 5 tell us that $T_M \to T$ in $HH^{1,2}(S)$.

Yet, as already observed, the harmonic coefficients expressing T_M from L to M will vary, as one can see for instance from the plot of the degree variances of EGM96, EGM08 (see Fig. 3.6 in Part I).

The two models in fact, although using different data, have been computed with quite the same methodology which, as we shall see in the rest of this section and in the next, can be reconducted to an approximation of a least squares solution.

A natural subspace useful to our purpose is that generated by the outer spherical harmonics $S_{\ell m}(r,\sigma)$. In fact let us put

$$H_{LM} = \text{Span}\{S_{\ell m}(r,\sigma),\ |m| \le \ell,\ L \le \ell \le M\},$$
$$S_{\ell m}(r,\sigma) = \left(\frac{R}{r}\right)^{\ell+1} Y_{\ell m}(\sigma) \tag{15.100}$$

In (15.100) naturally it is not strictly necessary to use the radius R, yet it is numerically convenient and in particular it is convenient to put R equal to the mean value of R_σ in such a way that $\delta\overline{R} = R_\sigma - R$ is as small as possible in the average. It has to be noted that by suitably choosing $L > 0$ we see that the harmonic functions in H_{LM} are all $O\left(\frac{1}{r^{L+1}}\right)$ at infinity; this fact can be used to automatically account for the case that we have an a-priori knowledge of coefficients up to degree $L-1$. This means that we shall look for a potential T_M of the form

$$T_M(r,\sigma;\mathbf{T}) = \sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell} T_{\ell m} S_{\ell m}(r,\sigma), \tag{15.101}$$

depending on the vector of unknown parameters $\mathbf{T} = \{T_{\ell m}\}$, and we shall compute from it the data model

$$\begin{aligned} f_M(\sigma) &= A_1 T_M - \sum_{\ell=0}^{L-1}\sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m} \\ &= \sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell} T_{\ell m} A_1(S_{\ell m}) - \sum_{\ell=0}^{L-1}\sum_{m=-\ell}^{\ell} a_{\ell m} Z_{\ell m}, \end{aligned} \tag{15.102}$$

with

$$A_1 = r\left(-\frac{\partial}{\partial h} + \frac{\gamma'}{\gamma}\right) \tag{15.103}$$

for the linearized Molodensky problem, and

$$f_M(\sigma) = A_2 T_M = \sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell} T_{\ell m} A_2(S_{\ell m}) \tag{15.104}$$

with

$$A_2 = r\left(-\frac{\partial}{\partial h}\right) \tag{15.105}$$

for the linearized fixed boundary BVP.

Note that $f_M(\sigma)$ is a linear function of the unknown vector $\mathbf{T}$ too. The idea is that we should use the coefficients $\mathbf{T}$ (and $\{a_{\ell m}\}$ in the case of (15.102)) at our disposal to produce by means of $f_M(\sigma)$ the best approximation of $f(\sigma)$ in the sense of $L^2(S)$; Theorems 3 and 4 then tell us that we are meanwhile approximating T in $HH^{1,2}(S)$. In other words we have to solve the least squares problem

$$\min_{\mathbf{T}} \| f(\sigma) - f_M(\sigma) \|_0^2 ;$$

the minimization extends to $\{a_{\ell m}\}$ in the case (15.102).

Noting that the operator A_1 is defined by (15.102), we obtain for such a problem the linear system

$$\begin{cases} \displaystyle\sum_{j=L}^{M} \sum_{k=-j}^{j} \langle A_1 S_{\ell m},\ A_1 S_{jk} \rangle_0 \, T_{jk} + \\ \displaystyle -\sum_{j=0}^{L-1} \sum_{k=-j}^{j} \langle A_1 S_{\ell m},\ Z_{jk} \rangle_0 \, a_{jk} = \langle A_1 S_{\ell m},\ f \rangle_0 \\ \displaystyle -\sum_{j=L}^{M} \sum_{k=-j}^{j} \langle Z_{\ell m},\ A_1 S_{jk} \rangle_0 \, T_{jk} + \\ \displaystyle +\sum_{j=0}^{L-1} \sum_{k=-j}^{j} \langle Z_{\ell m},\ Z_{jk} \rangle_0 \, a_{jk} = -\langle Z_{\ell m},\ f \rangle_0 . \end{cases} \tag{15.106}$$

In this system the first equations hold for all compatible m and $L \le \ell \le M$, while the second equations hold for $0 \le \ell \le L-1$.

Similarly, but in a simpler form, for the problem (15.104) one arrives at

$$\sum_{j=L}^{M} \sum_{k=-j}^{j} \langle A_2 S_{\ell m},\ A_2 S_{jk} \rangle_0 \, T_{jk} = \langle A_2 S_{\ell m},\ f \rangle_0 . \tag{15.107}$$

The numerical solution of (15.106) or (15.107) cannot be obtained by direct methods, when M is as large as several thousands, because the normal matrices implied are fully populated, since the integrals implicit in the scalar products are performed on functions like $S_{\ell m}(R_\sigma, \sigma)$ or $Z_{\ell m}(R_\sigma, \sigma)$, which are not orthogonal in $L^2(S)$.

Nevertheless they can be solved iteratively by taking the perturbative structure of A_1, A_2 into account.

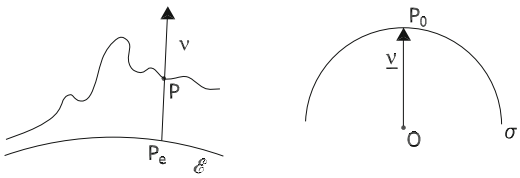


Fig. 15.1 The mapping $P \to P_e \to P_0$ from S to σ

In order to simplify our reasoning we shall for the moment forget the $\{a_{jk}\}$ unknowns and we shall rather treat the principal part of (15.106), and, in a perfectly parallel way, of (15.107). We leave to the last part of this section the task of introducing back $Z_{\ell m}$ into the perturbative scheme.

Remember that $\| \ \|_0$ and the scalar product in (15.106) should be consistent with (15.11); however in the present context we choose to map the points P of the surface S on the unit sphere σ as shown in Fig. 15.1.

Subsequently we decide to write the L^2 scalar product in an equivalent form as

$$\|f(P)\|_0^2 = \frac{1}{4\pi} \int f^2(P) d\sigma_{P_0}. \tag{15.108}$$

With this proviso, (15.106) yields the solution of the least squares principle

$$\min_{T_M} \|r\Delta g - r\Delta g_M\|^2 = \min_{T_M} \frac{1}{4\pi} \int [\Delta g(P) - \Delta g_M(P)]^2 r_P^2 d\sigma_{P_0}; \tag{15.109}$$

where P is mapped to P_0 as explained above.

Also, to simplify the notation, in the following formulas we don't write formally the range of summation of the indexes ℓ, m or j, k which however are assumed to run over $-\ell \le m \le \ell$, $\ell = L, \ldots, M$, and so forth. Therefore we can write more explicitely the main part of (15.106) as

$$\begin{aligned} &\sum_{\ell,m} \langle A_1 S_{jk}, A_1 S_{\ell m} \rangle_0 T_{\ell m} \\ &= \sum_{\ell,m} \left\{ \frac{1}{4\pi} \int (A_1 S_{jk}(P)) \frac{GM}{R} (A_1 S_{\ell m}(P)) d\sigma_{P_0} \right\} T_{\ell m} \\ &= \frac{1}{4\pi} \int (A_1 S_{jk}(P)) \Delta g(P) r_P \, d\sigma_{P_0} = \langle A_1 S_{jk}, r\Delta g \rangle_0 \end{aligned} \tag{15.110}$$

where the point P is restricted to the surface S, i.e. we set

$$r_P = R + \delta R(\vartheta, \lambda)$$

so that we have

$$S_{\ell m}(P) = \left[\frac{R}{R + \delta R(\vartheta, \lambda)}\right]^{\ell+1} Y_{\ell m}(\vartheta, \lambda). \tag{15.111}$$

Remark 3. Since we don't have really observations $\Delta g(P)$ for every point of the boundary, (15.110) can be conveniently discretized in the following way.

Let B_{rs} denote the geographic square

$$\begin{cases} B_{rs} = \{r\Delta \le \vartheta < (r+1)\Delta,\ s\Delta \le \lambda \le (s+1)\Delta\} \\ r = 0, 1, \dots N-1,\ s = 0 \dots 2N-1,\ \Delta = \frac{360^\circ}{N} \end{cases} \tag{15.112}$$

and put

$$\begin{cases} (\overline{f})_{rs} = \frac{1}{|B_{rs}|} \int_{B_{rs}} f(\vartheta, \lambda) d\sigma \\ |B_{rs}| = \text{ area of } B_{rs} = \Delta[\cos r\Delta - \cos(r+1)\Delta] \cong \sin \vartheta_r \cdot \Delta^2 \\ \vartheta_r = (r + \frac{1}{2})\Delta \end{cases} \tag{15.113}$$

If we have N_{rs} observations (or point values) of f in B_{rs} we can put, instead of (15.113),

$$(\overline{f})_{rs} = \frac{1}{N_{rs}} \Sigma_n f(P_n)\ ;\ P_n \in B_{rs}, \tag{15.114}$$

with an error which tends to zero if f is smooth enough and $N_{rs} \to \infty$, i.e. the data density tends to infinity.

Accordingly the elements of the normal matrix of (15.110) can be approximated by

$$\frac{1}{4\pi} \Sigma_{rs} (\overline{A_1 S_{jk}})_{rs} \frac{GM}{\overline{R}} (\overline{A_1 S_{\ell m}})_{rs} |B_{rs}| = D_{jk,\ell m}, \tag{15.115}$$

where the summation is over all the blocks B_{rs} covering the unit sphere σ; the known term (15.110) can be written as

$$\frac{1}{4\pi} \Sigma_{rs} (\overline{A_1 S_{jk}})_{rs} (\overline{r\Delta g})_{rs} |B_{rs}| = F_{jk}, \tag{15.116}$$

so that (15.110) becomes the algebraic system

$$\sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} D_{jk,\ell m} T_{\ell m} = F_{jk}. \tag{15.117}$$

Remark 4. In spite of its deceiving simplicity one has to consider that (15.117) written for instance for $M = 2{,}160$, $L = 20$ implies the solution of a system with no particular structure with 4,665,200 unknowns. This is a formidable numerical

task which at present can be tackled only by suitable sequential techniques, to get at least an approximate solution. This is exactly what we shall illustrate, guided by the geodetic intuition.

We shall obtain a simple solution of (15.117) through some steps:

(a) First of all we put our basic equation into a perturbative form, namely, recalling (15.30) and (15.31),

$$A_1 T = A_{1S} T + D_1 T = r\Delta g \tag{15.118}$$

with

$$A_{1S} = -r\frac{\partial}{\partial r} - 2 \tag{15.119}$$

and

$$\begin{cases} D_1 = r(\boldsymbol{\varepsilon} \cdot \nabla + \eta) \\ \boldsymbol{\varepsilon} = (\mathbf{e}_r - \boldsymbol{\nu}),\ \eta = \frac{2}{r} + \frac{\gamma'}{\gamma}. \end{cases} \tag{15.120}$$

Since (see Proposition (3))

$$|\boldsymbol{\varepsilon}| \leq \frac{1}{2}e^2 \quad |\eta| \leq 2e^2 \tag{15.121}$$

we can conveniently write (15.118) as

$$A_{1S} T = \left(-r\frac{\partial}{\partial r} - 2\right) T = r\Delta g - r\boldsymbol{\varepsilon} \cdot \nabla T - r\eta T. \tag{15.122}$$

If we have a good prior model we can compute $\Delta g_c = \Delta g - \boldsymbol{\varepsilon} \cdot \nabla T - \eta T$ from it, or at most we can iterate to get a better solution.
Moreover, since

$$A_{1S} S_{jk} = (j-1) S_{jk}, \tag{15.123}$$

(15.110) simplifies to

$$\langle S_{jk},\ A_{1S} T_M \rangle_0 = \langle S_{jk},\ r\Delta g_c \rangle_0 . \tag{15.124}$$

Remark 5. Equation 15.124 has a nice functional interpretation. In fact let us put

$$V_{L,M} = \text{Span}\{S_{\ell m}|_S\ ;\ |m| \leq \ell,\ \ell = L, \ldots M\},$$

i.e. the subspace generated by linear combinations of functions $\{S_{\ell m}|_S, S_{\ell m} \in H_{LM}\}$ where, as in (15.111), we mean $S_{\ell m}|_S = S_{\ell m}(R + \delta R(\vartheta, \lambda), \vartheta, \lambda)$. If we call P_M the orthogonal projector of L^2 on $V_{L,M}$, we find T_M by solving (recall (15.100))

$$P_M(A_{1S}T_M) = P_M(r\Delta g_c),\ T_M \in H_{L,M}, \tag{15.125}$$

i.e. by projecting the original equation onto $V_{L,M}$.

In fact one way to express that $(A_{1S}T_M)$ and $r\Delta g$ have the same projection on $V_{L,M}$ is exactly to claim that they must have the same scalar product with a base of $V_{L,M}$, which in our case is (15.124).

Such a method of simple projection is known in functional analysis as Galerkin's method. The interested reader can find much more material in the mathematical literature, e.g. in Mikhlin (1964) and Kirsch (1996). So by switching from the general form of the operator A_1 to its spherical approximation we find that least squares equations become identical to Galerkin's equations.

To perform the next step (b) we need to understand more clearly how Galerkin's method works. Basically we could say that given two Hilbert spaces X, Y and a bounded operator $A : X \to Y$, of which we already know that there is a bounded inverse $A^{-1} : Y \to X$, we want to solve the infinite dimensional equation

$$Ax = y. \tag{15.126}$$

In order to make (15.126) finite dimensional, we first select two sequences of subspaces W_M and V_M in X and Y respectively

$$W_M = \text{Span}\{\xi_n, n = 1 \dots M\}$$
$$V_M = \text{Span}\{\eta_n, n = 1 \dots M\},$$

such that $\{\xi_n\}$ and $\{\eta_n\}$ are complete (generally non-orthonormal) bases, each in its own space. Then we substitute (15.126) with the finite dimensional square system

$$\sum_{m=1}^{M} \langle \eta_n, A\xi_m \rangle \lambda_m = \langle \eta_n, y \rangle, \tag{15.127}$$

where

$$x_M = \sum_{m=1}^{M} \lambda_m \xi_m. \tag{15.128}$$

gives an approximation of x.

When

$$\eta_n = A\xi_n$$

the system (15.127) is the same as that of least squares and the convergence of x_M given by (15.128) to the right solution is guaranteed, otherwise it has to be studied case by case.

So up to the level of the system (15.124) we are on a fully justified theoretical ground. Specifically in this case we have

$$\{\xi_n\} \equiv \{S_{\ell m}(r, \vartheta, \lambda)\} \tag{15.129}$$

in the space $X \equiv HH^{1,2}(S)$, i.e. the space of potentials T, while

$$\{\eta_n\} = \{(\ell - 1)S_{\ell m}(R + \delta R(\vartheta, \lambda), \vartheta, \lambda)\}. \tag{15.130}$$

Note has to be taken that the use of the same functions $S_{\ell m}$ for ξ_n and η_n should not be misunderstood; in fact $\{\xi_n\}$ are potentials defined in Ω, while η_n are surface functions in $L^2(S)$, i.e defined on S only.

The next step (b) then is done on the basis of the remark that $S_{\ell m}(R + \delta R, \vartheta, \lambda)$ are quite close to $Y_{\ell m}(\vartheta, \lambda)$ because $O(\delta R/R) = 10^{-3}$ at most,

(b) We decide now to use $Y_{jk}(\vartheta, \lambda)$ instead of $S_{jk}|_S$ in (15.124), namely to substitute (15.124) with

$$\langle Y_{jk},\ A_{1S}T_M\rangle_0 = \langle Y_{jk}, r\Delta g_c\rangle_0 . \tag{15.131}$$

Numerical experiments fully support this choice.

Finally a third step has to be taken to come to a handable solution. We concentrate on the first member of (15.131) and first of all we set up the following identity:

$$\begin{aligned}
&\langle Y_{jk}, A_{1S}T_M\rangle_0 \\
&= \Sigma_{\ell,m} T_{\ell m}\frac{GM}{R}(\ell - 1)\cdot\frac{1}{4\pi}\int_\sigma Y_{jk}(\vartheta,\lambda)Y_{\ell m}(\vartheta,\lambda)\cdot\left(\frac{R}{R+\delta R}\right)^{\ell+1} d\sigma \\
&= \frac{GM}{R}T_{jk}(j-1) \qquad (15.132)\\
&+\frac{GM}{R}\Sigma_{\ell,m}T_{\ell m}(\ell-1)\cdot\frac{1}{4\pi}\int_\sigma Y_{jk}(\vartheta,\lambda)Y_{\ell m}(\vartheta,\lambda)\cdot\left[\left(\frac{R}{R+\delta R}\right)^{\ell+1} - 1\right] d\sigma \\
&= \frac{GM}{R}T_{jk}(j-1) + \frac{GM}{R}\Sigma_{\ell,m}T_{\ell m}(\ell-1)\langle Y_{jk}, Y_{\ell m}W_\ell\rangle \\
&= \frac{GM}{R}T_{jk}(j-1) + \langle Y_{jk}, r\Delta g_M(r,\vartheta,\lambda) - R\Delta g_M(R,\vartheta,\lambda)\rangle_0, \quad (r = R + \delta R).
\end{aligned}$$

Notice that $W_\ell = \left[\left(\frac{R}{R+\delta R}\right)^{\ell+1} - 1\right]$ are weights depending on (ϑ, λ) because δR is function of these variables and generally small; as a matter of fact Table 15.1 gives an idea of the behaviour of W_ℓ as functions of δR at very high frequencies ($\ell = 2{,}000$).

Table 15.1 weights at degree 2,000 at various heights

δR (m)	$W_{2,000}$
6,400	6.38
3,200	1.72
1,000	0.37

So we can take the final step:

(c) We write (15.131), using (15.132) too, in the form

$$\begin{aligned} T_{jk} &= \frac{1}{j-1}\left(\frac{GM}{R}\right)^{-1}\left\langle Y_{jk}, r\Delta g_c\right\rangle - \frac{1}{j-1}\left(\frac{GM}{R}\right)^{-1} \\ &\quad \cdot\left\langle Y_{jk}, [r\Delta g_M(r,\vartheta,\lambda) - R\Delta g_M(R,\vartheta,\lambda)]\right\rangle_0 \qquad (15.133) \\ &= \frac{1}{j-1}\left(\frac{GM}{R}\right)^{-1}\left\langle Y_{jk}, \{r\Delta g_c - [r\Delta g_M(r,\vartheta,\lambda) \right. \\ &\quad \left. - R\Delta g_M(R,\vartheta,\lambda)]\}\right\rangle_0 . \end{aligned}$$

If we remember what Δg_c is in terms of the original free air anomalies, we can ultimately rewrite (15.133) in the perturbative form

$$\begin{aligned} T_{jk} &= \frac{1}{j-1}\left(\frac{GM}{R}\right)^{-1}\left\langle Y_{jk}, \{r\Delta g - r\boldsymbol{\varepsilon}\cdot\nabla T_M - r\eta T_M \right. \\ &\quad \left. - [r\Delta g_M(r,\vartheta,\lambda) - R\Delta g_M(R,\vartheta,\lambda)]\}\right\rangle_0 \qquad (15.134) \end{aligned}$$

where in the second member T_M, Δg_M do depend on $\{T_{jk}\}$ and r means $r = R + \delta R(\vartheta,\lambda)$.

In principle (15.134), with the above remarks, is still an exact way of writing the Galerkin system, to find an approximate solution $\{T_{jk}, |k| \le j,\ j = L,\dots M\}$ from the datum Δg on S.

Its true numerical implementation naturally passes through a discretization of integrals similar to that presented in (15.115).

We note that if we think that all the terms appearing in the right hand side of (15.134) can be computed as "corrective terms" from some prior model $(T_M)_{\text{prior}}$, then indeed (15.134) will give us straightforwardly the sought solution. Otherwise we still have to work on the right hand side as explained in next section.

Remark 6. We want to call the attention on a point that has not been treated in this section. Namely we have just used a sphere of radius R as reference for the topography, while it would have been more suitable to use directly the ellipsoid $\mathcal{E}$. This can be done, without much difficulty, by using the ellipsoidal harmonics representation of Part I, Sect. 3.9, with the difference that now it is not anymore

true that $rAS^e_{\ell m} = \lambda_{\ell m} S^e_{\ell m}$, for some constant $\lambda_{\ell m}$, as it happened in (15.123), so one further approximation has to be introduced or the already approximated "radial" functions Part I, (3.197) are to be used.

What has been done for the Molodensky boundary operator A_1 can be repeated step by step for the Hotine operator A_2. The only difference is in the perturbative decomposition (cfr. Sect. 14.4, Remark 1)

$$A_2 = A_{2S} + D_2 \tag{15.135}$$

where now

$$\begin{cases} A_{2S} = -r\frac{\partial}{\partial r} \\ D_2 = r\boldsymbol{\varepsilon} \cdot \nabla. \end{cases} \tag{15.136}$$

Naturally the known term is in this case $r\delta g$, and instead of (15.123) we have now

$$A_{2S} S_{jk} = (j+1) S_{jk}. \tag{15.137}$$

Accordingly the analogous of the perturbative system (15.134) becomes

$$\begin{aligned} T_{jk} = \frac{1}{j+1}\left(\frac{GM}{R}\right)^{-1} \big\langle Y_{jk}, \{r\delta g - r\boldsymbol{\varepsilon} \cdot \nabla T_M - [r\delta g_M(r\vartheta, \lambda) + \\ - R\delta g_M(R, \vartheta, \lambda)]\}\big\rangle, \end{aligned} \tag{15.138}$$

with $r = R + \delta R(\vartheta, \lambda)$.

As promised we return finally to the true normal system (15.106) for Molodensky's problem.

Before doing that we simplify somewhat our equations, at least in notation, by recognizing that our perturbative scheme can be reduced to writing the following identities

$$\begin{aligned} A_1 S_{\ell m} &= (\ell - 1) S_{\ell m} + r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m} + r\eta S_{\ell m} \qquad (15.139) \\ &= (\ell - 1) Y_{\ell m} + (\ell - 1)[S_{\ell m} - Y_{\ell m}] + r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m} + r\eta S_{\ell m} \\ &= (\ell - 1) Y_{\ell m} + \Psi_{\ell m} \\ A_2 S_{\ell m} &= (\ell + 1) S_{\ell m} + r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m} \qquad (15.140) \\ &= (\ell + 1) Y_{\ell m} + (\ell + 1)[S_{\ell m} - Y_{\ell m}] + r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m} \\ &= (\ell + 1) Y_{\ell m} + \Phi_{\ell m} \end{aligned}$$

and considering $\Psi_{\ell m}, \Phi_{\ell m}$ as perturbations of the principal terms $(\ell \pm 1) Y_{\ell m}$. In (15.139) and (15.140) $S_{\ell m}$ means $S_{\ell m}(R + \delta R(\vartheta, \lambda), \vartheta, \lambda)$.

Now we return to (15.106) and in order to make more transparent our solution we do that in an Example, assuming that S itself is a sphere, so that $Y_{\ell m} - S_{\ell m}|_S \equiv 0$.

Example 1. Assume S to be sphere. We note immediately that in this case, according to the definition of $Z_{\ell m}$, (15.14), we can take directly

$$Z_{\ell m}(R, \vartheta, \lambda) = Y_{\ell m}(\vartheta, \lambda). \tag{15.141}$$

Accordingly (15.106) becomes

$$\begin{aligned}(\ell-1)^2 T_{\ell m} = &-\left\langle \Psi_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} (j-1) T_{jk} Y_{jk} \right\rangle_0 + \\ &-(\ell-1)\left\langle Y_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} T_{jk} \Psi_{jk} \right\rangle_0 + \\ &-\left\langle \Psi_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} T_{jk} \Psi_{jk} \right\rangle_0 + \sum_{j=0}^{L-1} \sum_{k=-j}^{j} \left\langle \Psi_{\ell m}, Y_{jk} \right\rangle_0 a_{jk} \\ &+ \langle A_1 S_{\ell m}, r\Delta g \rangle_0 , \end{aligned} \tag{15.142}$$

$$a_{\ell m} = \left\langle Y_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} T_{jk} \Psi_{jk} \right\rangle_0 - \langle Y_{\ell m}, r\Delta g \rangle_0 . \tag{15.143}$$

Note that in deriving (15.142) and (15.143) one has carefully to exploit the fact that $\sum_{j,k=0}^{L-1} a_{jk} Z_{jk}$, in this context is always L^2 orthogonal to all $Y_{\ell m}$, with $\ell \geq L$.

15.6 Two Geodetic Solutions of Galerkin's System

To the knowledge of the author only two methods have been applied to produce high resolution ($M > 10^3$) global models from Galerkin's equations (15.133): one is the finite dimensional version of the change of boundary method (Sansò and Sona 1995) and has been implemented by Wenzel (see Wenzel 1998); the other one is the so-called *downward continuation method*, implemented by Rapp (1997a) and developed with his co-workers Pavlis et al. (2008). This second method is described with some variants and much more detail in the second part of the book, in Chap. 6.

(a) *Change of boundary*

The concept can be illustrated for the Dirichlet problem, which is the only one for which we have a theoretical result. In any case remember that by means of the change of unknown

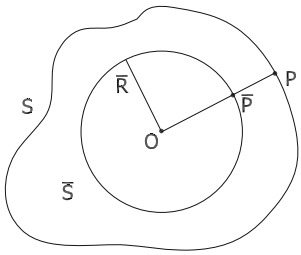


Fig. 15.2 Illustration of the pullback operator $f(P) \to \widetilde{f}(\overline{\sigma})$

$$u = -r\frac{\partial T}{\partial r} - 2T = r\Delta g, \tag{15.144}$$

we can transform the BVP's (15.27) and (15.33) in spherical approximation, into a Dirichlet problem (see on Proposition 4), so this last is not at all useless for a set up of geodetic relevance.

The idea is to take the boundary value of $u = f$, given on the true surface S and to shift it to a Bjerhammar sphere $\overline{S}$ (see Fig. 15.2) be means of the pull-back correspondence (radial projection) $P \to \overline{P}$. In other words we substitute the original Dirichlet BVP

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u(P)|_S = f(P) & \text{on } S \end{cases} \tag{15.145}$$

with the new Dirichlet problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u(\overline{P}) = f(P) & \text{on } \overline{S}. \end{cases} \tag{15.146}$$

The problem (15.146) can be easily solved by means of the Poisson integral, because it refers to a sphere. The operator that defines the new function on $\overline{S}$

$$\overline{f}(\overline{P}) = f(P) \tag{15.147}$$

is called here the pull-back operator and denoted as

$$PB : L^2(S) \to L^2(\overline{S}). \tag{15.148}$$

The function $\overline{u}$ which is the solution of (15.146) can then be evaluated back at the surface S where it takes values $\overline{u}|_S$ which are indeed different from $f(P)$.

For the sake of definiteness we call Π the Poisson operator that gives $\overline{u}$ from $\overline{f}$ and L the "lift" operator such that

$$L(\overline{f}) = (\Pi\overline{f})|_S = \overline{u}(P) \tag{15.149}$$

With f we can form residuals

$$\delta f(P) = f(P) - \overline{u}(P), \tag{15.150}$$

which provide a new (hopefully smaller) boundary function on S. Now we pull back δf from S to $\overline{S}$ and we iterate. The scheme is known to converge with continuous data in the sense of uniform convergence on S (Sansò and Sona 1995). How it works in $L^2(S)$ is not known, yet if we implement a finite dimensional version of it by introducing the $L^2(\overline{S})$ projector

$$P_{LM} f = \sum_{\ell,m=L}^{M} \left(\frac{1}{4\pi} \int Y_{\ell m}(\vartheta,\lambda) f(\vartheta,\lambda) d\sigma \right) Y_{\ell m}(\vartheta,\lambda) \tag{15.151}$$

we get the iterative scheme of Fig. 15.3 that can be ultimately transformed back into a corresponding scheme for T by inverting (15.144). Since (15.144) is solved, for a finite dimensional potential, by

$$\begin{cases} T_M = \displaystyle\sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} \left(\frac{1}{\ell - 1} \right) u_{\ell m} S_{\ell m}(r,\vartheta,\lambda), \\ T_{\ell m} = (\ell - 1)^{-1} u_{\ell m} \end{cases} \tag{15.152}$$

we end up with an iterative scheme which is exactly the simple iterative solution of the Galerkin equations (15.133); therefore we can continue our reasoning on Fig. 15.2 and at the end transform the result back to the anomalous potential T by (15.152).

Naturally we would like to know whether a scheme like that is convergent and, in case of a positive answer, whether it converges to the right solution.

For this purpose we can examine more closely Fig. 15.3. First we note that all functions on the left are defined on S while those on the right are defined on $\overline{S}$, $(\delta \overline{f}_k)$, or in $\overline{\Omega}(r \geq \overline{R})$, $(\overline{u}_k)$. We note too that if the points P on S (see Fig. 15.2) are already expressed in terms of spherical coordinates, i.e. as function of (ϑ,λ) of $\overline{P}$, then the pull-back operator PB is just the identity.

Moreover, we observe that while $f_k, \delta f_k$ are general functions in L^2, $\delta \overline{f}_k, \overline{u}_k$ are on the contrary always finite dimensional functions with maximum degree M. This is achieved when we move horizontally to the right in Fig. 15.3, because we first apply the pull-back to δf_k and then we truncate the resulting functions in $L^2(\overline{S})$

$$\delta \overline{f}_k = \frac{GM}{\overline{R}} \sum_{\ell=L}^{M} \sum_{m=-\ell}^{\ell} (\delta \overline{f}_k)_{\ell m} Y_{\ell m}(\vartheta,\lambda). \tag{15.153}$$

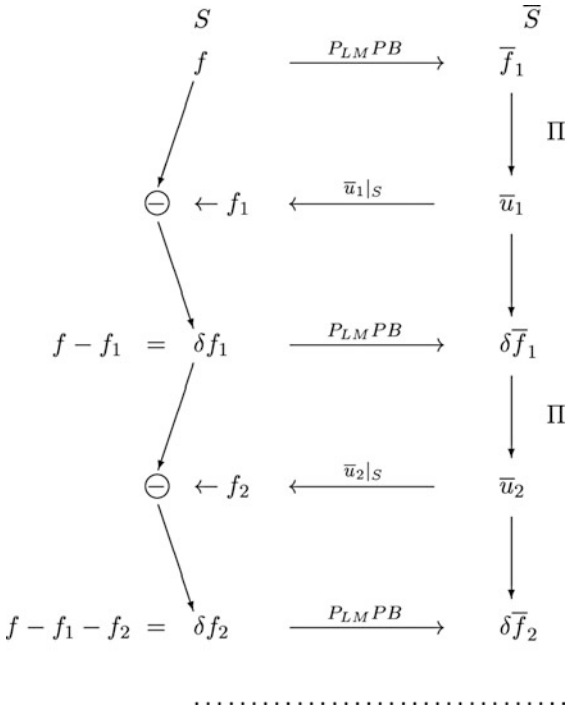


Fig. 15.3 The finite dimensional change of boundary iterative scheme

The potentials corresponding to (15.153) are then

$$\overline{u}_k = \Pi\delta\overline{f}_k = \frac{GM}{\overline{R}}\sum_{\ell=L}^{M}\sum_{m=-\ell}^{\ell}(\delta\overline{f}_k)_{\ell m}S_{\ell m}(r,\vartheta,\lambda). \tag{15.154}$$

Since we evaluate the size of $\overline{u}_k$ in $HL^2(\overline{S})$ as the size of $\delta\overline{f}_k$ in $L^2(\overline{S})$, we have identically

$$\|\overline{u}_k\|^2 = \|\delta\overline{f}_k\|^2 = \left(\frac{GM}{\overline{R}}\right)^2\Sigma_{\ell,m}(\delta\overline{f}_k)^2_{\ell m}. \tag{15.155}$$

Please notice that the factor $\frac{GM}{\overline{R}}$ in front of (15.153) and (15.154) is conventional and introduced to make (15.152) consistent with (15.133).

Furthermore, always on Fig. 15.3, we read

$$\delta f_k = f - \sum_{n=1}^{k} f_n \tag{15.156}$$

$$f_{k+1} = L\delta\overline{f}_k \equiv \overline{u}_k|_S, \tag{15.157}$$

$$\delta f_{k+1} = \delta f_k - L\delta\overline{f}_k, \tag{15.158}$$

$$\delta\overline{f}_{k+1} = P_{LM}PB\delta f_{k+1} = \delta\overline{f}_k - P_{LM}PBL\delta\overline{f}_k, \tag{15.159}$$

Since $\delta\overline{f}_k$ are just the residuals δf_k, pulled back to $\overline{S}$ and projected by P_{LM}, we expect that convergence means $\delta\overline{f}_k \to 0$ in $L^2(\overline{S})$. As all the functions $\{\delta\overline{f}_k\}$ are in Span $\{Y_{\ell m}, L \leq \ell \leq M\}$, we see that (15.159) can be written in the equivalent form (remember that $P_{LM}^2 = P_{LM}$)

$$\delta\overline{f}_{k+1} = (P_{LM} - P_{LM}PBLP_{LM})\delta\overline{f}_k. \tag{15.160}$$

Therefore a sensible sufficient conditions for $\delta\overline{f}_k$ to tend to zero is just

$$\chi_{LM} = \|P_{LM} - P_{LM}PBLP_{LM}\| < 1; \tag{15.161}$$

in (15.161) the norm can be understood in the sense of the L^2 operator norm or, considering that after all (15.160) is a finite dimensional relation, it can be cast in the form of the norm of the matrix that implements (15.160) as a transformation between the harmonic coefficients of $\delta\overline{f}_k$ into those of $\delta\overline{f}_{k+1}$. In any event (15.161) puts a bound on the topography (remember that in this section for the sake of simplicity the topography is directly attached to a sphere instead of the ellipsoid, however without changing the basic nature of the problem).

Yet condition (15.161) still has to be studied in detail, though present numerical experiments say that up to degree $2 \cdot 10^3$ convergence is verified. So we shall make the conjecture that (15.161) is satisfied by a realistic topography and we try to answer to the second question. We first note that $\delta\overline{f}_k \to 0$ implies $f_k \to 0$ because of (15.157) as well as $\overline{u}_k \to 0$ because of (15.155). Even more, from (15.161) we see that

$$\|\delta\overline{f}_{k+1}\| \leq \chi_{LM}^k \|\delta\overline{f}_1\| \tag{15.162}$$

so that the series

$$\sum_{k=1}^{+\infty} \delta\overline{f}_k$$

has to be $L^2(\overline{S})$ convergent. Accordingly the two series

$$g = \sum_{k=1}^{+\infty} f_k, \ \overline{u} = \sum_{k=1}^{+\infty} \overline{u}_k \tag{15.163}$$

have to be convergent too. So if we use (15.156) we are justified to write

$$\lim_{k\to\infty} \delta f_k = f - \sum_{n=1}^{+\infty} f_n = f - g \tag{15.164}$$

as well as

$$\begin{aligned} 0 &= \lim_{k\to\infty} \delta \overline{f}_k = \lim_{k\to\infty} P_{LM} P B \delta f_k \\ &= P_{LM} P B f - P_{LM} P B g. \end{aligned} \tag{15.165}$$

Just another way of writing (15.165) is

$$\langle Y_{\ell m}, f \rangle = \langle Y_{\ell m}, g \rangle \,, \; |m| \le \ell, L \le \ell \le M. \tag{15.166}$$

On the other hand

$$\begin{aligned} g(\vartheta,\lambda) &= \sum_{k=1}^{+\infty} f_k = \sum_{k=1}^{+\infty} \overline{u}_k(r,\vartheta,\lambda)|_S \\ &= \overline{u}(r,\vartheta,\lambda)|_S = \overline{u}(\overline{R} + \delta R(\vartheta,\lambda),\vartheta,\lambda), \end{aligned} \tag{15.167}$$

so that $\overline{u}(r,\vartheta,\lambda)$ is a potential with maximum degree M satisfying the Galerkin equations

$$\langle Y_{\ell m}(\vartheta,\lambda),\, f(\vartheta,\lambda)\rangle = \langle Y_{\ell m}(\vartheta,\lambda), \overline{u}(r,\vartheta,\lambda)|_S\rangle \tag{15.168}$$

for all orders and all appropriate degrees. In other words the series obtained by adding all $\overline{u}_k$ provides the solution of our problem and answers to our question.

Remark 7. We have to note that the series (15.163) defining the potential $\overline{u}$ is added on the *iteration index* k, but it provides in any event a sum which is still a finite degree potential. This should not be confused with the possibility of defining a convergent series with infinite degrees representing the solution of our BVP. Such a series in fact does not exist in general as we have already pointed out in Part I, Sect. 3.5.

(b) *The downward continuation approach*

This approach is seemingly completely different form the previous one, because it goes back to the Galerkin system (15.133) and tries to transform it in such a way that the solution comes out without many iterations. This is more easily done in terms of the harmonic function (see (15.144))

$$u = r\Delta g \tag{15.169}$$

and the corresponding approximation (with degrees from L to M).

$$u_M = r\Delta g_M. \tag{15.170}$$

As a matter of fact, recalling also (15.152), and putting $f(\vartheta,\lambda) = u|_S$, (15.133) can be written as

$$(u_M)_{jk} = \left(\frac{GM}{\overline{R}}\right)^{-1} \left\langle Y_{jk}(\vartheta,\lambda),\, f - [u_M(\overline{R}+\delta R,\vartheta,\lambda) - u_M(\overline{R},\vartheta,\lambda)]\right\rangle$$
$$|k| \le j,\ L \le j \le M\ ; \tag{15.171}$$

in (15.171) δR, the topography, is a function of (ϑ,λ). Now instead of iterating on the coefficients, which in the right hand side are hidden in u_M, we rather make a kind of Taylor development of $u_M(\overline{R}+\delta R,\vartheta,\lambda) - u_M(\overline{R},\vartheta,\lambda)$, i.e.

$$u_M(\overline{R}+\delta R,\vartheta,\lambda) \cong u_M(\overline{R},\vartheta,\lambda) - \sum_{\alpha=1}^{a} \frac{\partial^\alpha u_M(\overline{R}+\frac{1}{2}\delta R,\vartheta,\lambda)}{\partial r^\alpha}\ \frac{(-\delta R)^\alpha}{\alpha!} \tag{15.172}$$

As we can see the development is done at the level of the midpoint between the topography and the ellipsoid, because this guarantees the best performance of the Taylor formula. We note as well that the development (15.172) is performed for the model u_M and not for the true potential; in this way we overcome the objection that there cannot be downward continuation for $u(r,\vartheta,\lambda)$ in general.

The experience is that with one iteration at most (15.172) used in (15.171) can provide the right answer with an accuracy compatible with the order of magnitude of the coefficients at degree $2\cdot 10^3$. As such the use of (15.172) can be considered as an accelerator of the iteration procedure, i.e., by computing a certain number of derivatives from the first iteration step, one avoids (or reduces) the subsequent steps.

This phenomenon is well illustrated by the next elementary example where some features of the two methods are highlighted.

Example 2. We consider a situation in which the surface S has equation

$$r = (1-\varepsilon\cos\vartheta)^{-1}$$
$$= 1+\varepsilon\cos\vartheta + 0(\varepsilon^2)$$

and $\overline{S}$ is just the sphere with unit radius. The potential we want to retrieve is

$$u = \frac{1}{r}$$

so that the corresponding true boundary values are given by

$$f = u|_S = 1-\varepsilon\cos\vartheta.$$

Method (a): we just consider f as given on $\overline{S}$ and we note that, denoting $\cos\vartheta = t$,

$$f = 1 - \varepsilon\cos\vartheta = P_0 - \varepsilon P_1,\ (O(f) = O(1)),$$

so that the corresponding potential is, with the same notation of Fig. 15.3,

$$\overline{u}_1 = \frac{P_0}{r} - \varepsilon\frac{P_1}{r^2}.$$

If we now compute

$$\begin{aligned} f_1 = \overline{u}_1|_S &= 1(1-\varepsilon t) - \varepsilon t(1-\varepsilon t)^2 \\ &= 1 - 2\varepsilon t + O(\varepsilon^2), \end{aligned}$$

we can put

$$\delta f_1 = f - f_1 = \varepsilon t + O(\varepsilon^2)\ \ (O(\delta f_1) = 0(\varepsilon)).$$

Therefore

$$\overline{u}_2 = \frac{\varepsilon t}{r^2} + O(\varepsilon^2) = \varepsilon\frac{P_1(t)}{r^2} + O(\varepsilon^2)$$

and

$$\begin{aligned} f_2 = \overline{u}_2|_S &= \varepsilon t(1-\varepsilon t)^2 + O(\varepsilon^2) \\ &= \varepsilon t + O(\varepsilon^2), \end{aligned}$$

It is then clear that, if we iterate, we get for δf_k an approximation of the order $O(\varepsilon^k)$.

It is useful to observe that indeed $\delta f_k \to 0$ in this case because u already belongs to a finite dimensional space. We note as well that u has the degree zero component which we usually don't have in T; this is because to solve the BVP for free air anomalies one usually assumes that the degree present in T start at least from $\ell = 2$. Yet one can always think to solve the BVP for the gravity disturbance δg which in principle is in biunivocal correspondence with $u = r\delta g$. After all one should keep in mind that the example is built on Dirichlet problem and it is done so simple that one can grasp immediately the type of convergence of the iteration scheme.

Method (b): we aim to prove that in the present example implementing the downward continuation, by means of the first vertical derivative only, speeds up convergence to the $O(\varepsilon^3)$ level of approximation in one step.

Note that in our case

$$\delta R = r - 1 = (1-\varepsilon t)^{-1} - 1 = \varepsilon t + \varepsilon^2 t^2 + O(\varepsilon^3)\ ;$$

moreover $\frac{\partial u}{\partial r} = -\frac{1}{r^2}$ so that

$$\frac{\partial u}{\partial r}\left(1 + \frac{1}{2}\delta R\right) = -\left(\frac{1}{1 + \frac{1}{2}\varepsilon t + O(\varepsilon^2)}\right)^2 = -(1 - \varepsilon t + O(\varepsilon^2)).$$

Therefore the downward continued boundary value is

$$\begin{aligned} & f - \frac{\partial u(1 + \frac{1}{2}\delta R)}{\partial r} \cdot \delta R \\ & = 1 - \varepsilon t + (1 - \varepsilon t + O(\varepsilon^2))(\varepsilon t + \varepsilon^2 t^2 + O(\varepsilon^3)) \\ & = 1 - \varepsilon t + \varepsilon t - \varepsilon^2 t^2 + \varepsilon^2 t^2 + O(\varepsilon^3) = 1 + O(\varepsilon^3). \end{aligned}$$

Accordingly the approximate potential is

$$\overline{u} = \frac{1}{r} + O(\varepsilon^3)$$

as announced.

15.7 New Data Sets from Spatial Gravity Surveying

The introduction of accelerometers on satellites and the possibility of accurately tracking orbits in continuous from the GNSS constellation, has opened a new era of a direct measurement of functionals of the gravity potential. Indeed every measurement of spatial geodesy is in one way or another related to the gravity field through the dynamics of the satellite. Yet what we are achieving now is something different, namely *localized* (i.e. referring to a point) functionals of W (and therefore of T) which can then be treated very much in the same way as we treat gravity observations on the earth surface. As already mentioned, this can be particularly useful, for instance, in areas with data gaps. Naturally an overall analysis of such data is usually performed so as to produce a set of harmonic coefficients up to some maximum degree. Yet another way of synthesizing the results of a specific mission is to produce at satellite altitude, or little below, grids of various functionals of T. This is the so-called spacewise approach to the analysis of satellite missions (cf. Rummel R et al. 1993)

To fix the ideas we shall shortly describe what one can do with a satellite mission at $\sim$250 km, bearing on board a GPS receiver and a cluster of accelerometers linked to form a gradiometer. To fix the ideas we shall assume that we are able to retrieve the position of the satellite every second (i.e. every 8 km along the orbit) with 1 cm error. The accelerometers, that feel the same gravimetric acceleration as the barycenter of the satellite, will measure only accelerations of non-gravitational forces $\mathbf{f}$ with an accuracy in the range of a fraction of nGal (1 nGal=

10^{-9} Gal), for instance $O(\mathbf{f}) \sim 0.5$ nGal. Therefore they are sensitive to variations in the gravitational acceleration, from one position to another one meter apart, of $5 \cdot 10^{-12}\,\mathrm{s}^{-2}$ (5 mEU; 1 Eotvos Unit $= 10^{-9}\,\mathrm{s}^{-2}$).

A mission of this type, GOCE (Global gravity field and Ocean Circulation Experiment), has been launched by ESA in 2009.

Since the inclination of the orbit is not exactly polar, for the purpose of keeping the attitude of the satellite constant with respect to the sun, but has an inclination of $\sim 96°$, there are two small caps over the poles that are never visited by the satellite. All the rest of a sphere at satellite altitude is more or less covered with a dense irregular net of data points. Let us see shortly what are the observables derived by the satellite. These are of two types: one is from satellite tracking from GPS, the others are gradiometric.

We shall not enter into the detailed analysis of the observables but we shall rather give the principles from which we can derive the observation equations.

Tracking data and energy balance. We assume that GPS data can provide positions $\mathbf{X}_I(t)$ of the satellite in a quasi-inertial system I (cf. Part I, Sect. 1.4) as well as the velocity $\dot{\mathbf{X}}_I(t)$ of the satellite. In the inertial system we can write the dynamic equation

$$\ddot{\mathbf{X}}_I = \nabla_{\mathbf{X}} V + \mathbf{g}_P + \mathbf{f}, \tag{15.173}$$

where V is the purely gravitational part of the earth gravity potential (remember that I is non-rotating with respect to stars), $\mathbf{g}_P$ is the set of perturbative gravitational accelerations (luni-solar attraction, tides, etc.), which can be assumed to be known, $\mathbf{f}$ includes all the non-gravitational accelerations acting on the satellite (atmospheric drought, light pressure, albedo, etc.)

Essential is that $\mathbf{f}$ is observed by the accelerometers. The only warning in using (15.173) is that if $V(\mathbf{x})$ is the gravitational potential at the earth-fixed position $\mathbf{x}$, then the function of $\mathbf{X}_I$ to be used in (15.173) is

$$V = V(R^t(t)\mathbf{X}_I) \tag{15.174}$$

where $R(t)$ is the rotation matrix that brings $\mathbf{x}$ into $\mathbf{X}_I = R(t)\mathbf{x}$ and $R^t(t)$ its inverse. Accordingly the acceleration $\mathbf{g}_I(\mathbf{X}) = \nabla_{\mathbf{X}} V$ is given by

$$\begin{aligned} \mathbf{g}_I(\mathbf{X}) &= R(t)\nabla_{\mathbf{x}} V(\mathbf{x})|_{\mathbf{x}=R^t\mathbf{X}} \\ &= R(t)\mathbf{g}(R^t(t)\mathbf{X}_I), \end{aligned} \tag{15.175}$$

where $\mathbf{g}$ is the pure gravitational part of the earth fixed gravity acceleration vector.

From (15.173) multiplying by $\dot{\mathbf{X}}_I$ we find the (specific) power balance equation

$$\frac{1}{2}\frac{d}{dt}|\dot{\mathbf{X}}_I|^2 = \frac{1}{2}\frac{d}{dt}v_I^2 = \frac{d}{dt}V(\mathbf{X}_I) + (\mathbf{g}_P + \mathbf{f}) \cdot \dot{\mathbf{X}}_I. \tag{15.176}$$

Integrating along the orbit from 0 to t we find

$$\frac{1}{2}v_I^2(t) - \frac{1}{2}v_I^2(0) = V(\mathbf{X}_I(t)) - V(\mathbf{X}_I(0)) + \int_0^t (\mathbf{g}_P + \mathbf{f}) \cdot \dot{\mathbf{X}}_I \, dt$$

If we consider that $V(\mathbf{X}_I) = V_e(\mathbf{X}_I) + T(\mathbf{X}_I)$ and we put all together the constants

$$E_0 = \frac{1}{2}v_I(0) - V(\mathbf{X}_I(0)) \tag{15.177}$$

we see that (15.177) can be written as

$$T(\mathbf{X}_I) + E_0 = -V_e(\mathbf{X}_I(t)) + \frac{1}{2}v_I^2(t) + \int_0^t (\mathbf{g}_P + \mathbf{f}) \cdot \dot{\mathbf{X}}_I \, dt, \tag{15.178}$$

where to the LHS we have the unknown functional $T(\mathbf{X}_I)$ and the unknown parameter E_0, while to the right we have only known or observed quantities.

So (15.178) is a localized observation equation for T, with the unknown parameter E_0.

Gradiometric observations. They are just derivatives of the vector of gravitational acceleration. These are obtained by differentiating the signals of the accelerometers, considering that the common part, namely the external forces, act in a similar way on all the accelerometers.

So what is left is a matrix of second derivatives M_I. This is observed in the system I, but has to be the related to the matrix of second derivatives $M = \left[\frac{\partial^2 V}{\partial \mathbf{x} \partial \mathbf{x} t}\right]$ in the earth fixed system. Such a relation is known to be simply

$$M_I = R(t) M R^t(t). \tag{15.179}$$

Accordingly, if we know $R(t)$, we can retrieve the matrix of second derivatives, which, in a spherical local triad have observation equations (in spherical approximation).

Among them the observation of $T_{rr}(P)$ is particularly easy to handle and we have already invited the reader to compute its covariance function in Part I, Chap. 5, Exercise 2.

Naturally the true analysis of data and realistic observation equations are much more complicated than that. Yet the principles are presented and the observation equations, after a first filtering along the orbit taking into account the strong

correlation of the accelerometers noise, can be used to predict by collocation suitable functionals on a sphere, more or less at satellite altitude.

It turns out that a good solution is to predict a regular grid of $10' \times 10'$ size on a sphere at about 150 km altitude; at each knot T is predicted, with an error, in terms of $\frac{T}{\gamma}$, of the order of 2 mm, and T_{rr} with an error of 0,6 mEU.

Such grids constitute an entirely new data sets covering most of the earth at 150 km altitude and their use, in conjunction with ground data for an optimal estimate of high resolution geoid, for instance along the lines illustrated in Part I, Sect. 5.11, is a challenge that will keep geodesists busy for some years.

15.8 Exercises

Exercise 1. Refer to the notation introduced in Sect. 3.1.

When S is a sphere we have $I_+ = 0, J_+ = 1$. Prove directly the inequality (15.40) for this case, when $\alpha = 0$, by using the explicit representation

$$T = \sum_{\ell=0}^{+\infty} \sum_{m=-\ell}^{\ell} T_{\ell m} \left(\frac{R}{r}\right)^{\ell+1} Y_{\ell m}(\sigma), \; v = \sum_{\ell,m=0}^{+\infty} v_{\ell m} \left(\frac{R}{r}\right)^{\ell+1} Y_{\ell m}(\sigma),$$

and showing that (15.40) is reduced to the algebraic inequality

$$(\ell + 1)(\ell + 2) \leq 2(\ell + 1)^2, \; \ell \geq 0.$$

Observe that this can be an equality if the degree zero only is present in the harmonic series.

(**Hint**: remember that $|\nabla T|^2 = (u')^2 + \frac{1}{R^2}|\nabla_\sigma T|^2$. Furthermore use the surface Gauss theorem, namely

$$\int_\sigma |\nabla_\sigma T|^2 d\sigma = \int_\sigma (-\Delta_\sigma T) T d\sigma,$$

and remember that the spherical harmonics $\{Y_{\ell m}\}$ are eigenfunctions of the Laplace Beltrami operator Δ_σ, i.e.

$$-\Delta_\sigma Y_{\ell m} = \ell(\ell + 1) Y_{\ell m}.)$$

Exercise 2. Prove that the condition (15.75) can be cast in the form

$$L \geq \frac{4J_+^2}{1 - 4J_+^2 \frac{\delta R}{R_+}} - 2.$$

Verify that with $I_+ = 60°$, $J_+ = 2$, $\frac{\delta R}{R_+} = \frac{30}{6{,}400}$; the simple Molodensky problem has a regular solution in $HH^{1,2}(S)$ if $L \geq 33$, i.e. if the first 33 degrees of T are considered as known. The estimate is not strict.

Exercise 3. Note that recalling the estimates (15.25) and (15.26) one can cast (15.83) into the form

$$2C_{0L} < \frac{1 - e^2 J_+}{2J_+(1 + e^2)} \cong \frac{1}{2J_+}[1 - e^2(J_+ + 1)].$$

Verify that if we want to be able to handle a geometry where $I_+ = 60°$ and, $\delta R = 30$ km (i.e. mountains up to 7,200 m) we must have

$$L \geq 34,$$

which is not very different from the case of Exercise 2.

Exercise 4. Use the estimate (15.25) to prove that condition (15.98) is satisfied if

$$I_+ \leq 89°.6.$$

Exercise 5. Consider the normal systems (15.106) and (15.107) in case that S is directly a sphere of radius R and note that in such a case one has $S_{\ell m}|_S \equiv Y_{\ell m}, Z_{\ell m} \equiv Y_{\ell m}$ and

$$\begin{aligned} \Psi_{\ell m} &= -r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m} - \eta S_{\ell m} \\ \Phi_{\ell m} &= -r\boldsymbol{\varepsilon} \cdot \nabla S_{\ell m}, \end{aligned}$$

as in Example 1. Prove that in this case the normal systems can be solved by the iterative schemes

$$\begin{aligned} (\ell - 1)^2 T_{\ell m}^{(N+1)} = &- \left\langle \Psi_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} (i - j) Y_{jk} T_{jk}^{(N)} \right\rangle_0 + \\ &- (\ell - 1) \left\langle Y_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} \Psi_{jk} T_{jk}^{(N)} \right\rangle_0 + \\ &- \left\langle \Psi_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} \Psi_{jk} T_{jk}^{(N)} \right\rangle_0 \\ &+ \sum_{j=0}^{L-1} \sum_{k=-j}^{j} \left\langle \Psi_{\ell m}, Y_{jk} \right\rangle_0 a_{jk}^{(N)} + \langle A_1 S_{\ell m}, f \rangle_0 , \ L \leq \ell \leq M, \end{aligned}$$

$$a_{\ell m}^{(N)} = \left\langle Y_{\ell m}, \sum_{j=L}^{M} \sum_{k=-j}^{j} \Psi_{jk} T_{jk}^{(N)} \right\rangle_0 - \langle Y_{\ell m}, f \rangle_0 , \; 0 \le \ell \le L-1,$$

for (15.106), and

$$\begin{aligned}(\ell+1)^2 T_{\ell m}^{(N+1)} = (\ell+1) &\left\langle Y_{\ell m}, \sum_{j=0}^{M} \sum_{k=-j}^{j} \Phi_{jk} T_{jk}^{(N)} \right\rangle_0 \\ &+ \left\langle \Phi_{\ell m}, \sum_{j=0}^{M} \sum_{k=-j}^{j} Y_{jk}(j+1) T_{jk}^{(N)} \right\rangle_0 + \\ &- \left\langle \Phi_{\ell m}, \sum_{j=0}^{M} \sum_{k=-j}^{j} \Phi_{jk} T_{jk}^{(N)} \right\rangle_0 + \langle A_2 S_{\ell m}, f \rangle_0\end{aligned}$$

for (15.107).

Exercise 6. Consider the situation of Example 1 and write the perturbative system for $\{T_{jk}\}$, when δg is given on the spherical boundary S.

(**Hint:** verify that (15.90) becomes

$$\begin{aligned}(\ell+1)^2 T_{\ell m} = &-(\ell+1) \sum_{j,k=2}^{M} \langle Y_{\ell m}, \Phi_{\ell m} \rangle_0 T_{jk} + \\ &- \sum_{j,k=2}^{M} (j+1) \left\langle \Phi_{\ell m}, Y_{jk} \right\rangle_0 T_{jk} + \\ &- \sum_{j,k=2}^{M} \left\langle \Phi_{\ell m}, \Phi_{jk} \right\rangle_0 T_{jk} + \langle A_2 S_{\ell m}, r\delta g \rangle_0 .)\end{aligned}$$

References

Abd-Elmotaal H (1995) Attraction of the topographic masses. Bull Geod 69(4):304–307

Albertella A, Sacerdote F, Sansò F (1992) From harmonic to Fourier analysis on the sphere. In: Holota P, Vermeer M (eds) Proceedings of the 1st workshop on geoid in Europe. Institute of geodesy topography and cartography, Prague

Andersen OB (1995) Global Ocean tides from ERS-1 and TOPEX/POSEIDON altimetry, J Geophys Res 100(C12):25,249–25,260

Andersen OB (1999) Shallow water tides on the northwest European shelf from TOPEX/POSEIDON altimetry. J Geophys Res 104:7729–7741

Andersen OB, Hinderer J (2005) Global inter-annual gravity changes from GRACE: early results. Geophys Res Lett 32(1):L01402. doi:10.1029/2004GL020948

Andersen OB (2008) Marine gravity and geoid from satellite altimetry. In: Lecture notes, international school for the determination and use of the geoid, Como

Andersen OB (2010) Marine gravity and geoid from satellite altimetry. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg

Andersen OB, Knudsen P (2000) The role of satellite altimetry in gravity field modelling in coastal areas. Phys Chem Earth A 25(1):17–24

Andersen OB, Knudsen P (1998) Global marine gravity field from the ERS-1 and GEOSAT geodetic mission altimetry. J Geophys Res 103(C4):8129–8137

Andersen OB, Knudsen P (2008) The DNSC08 global mean sea surface and bathymetry. Presented EGU-2008, Vienna, Apr 2008

Andersen OB, Knudsen P (2009) The DNSC08 mean sea surface and mean dynamic topography. J Geophys Res 114(C11). doi:10.1029/2008JC005179

Andersen OB, Scharroo R (2011) Range and geophysical corrections in coastal regions. In: Vignudelli S et al (eds) Coastal altimetry, Springer, Berlin/Heidelberg. ISBN: 978-3-642-12795-3

Andersen OB, Woodworth PL, Flather RA (1995) Intercomparison of recent global ocean tide models. J Geophys Res 100(C12): 25,261–25,282

Andersen OB, Seneviratne SI, Hinderer J, Viterbo P (2005a) GRACE-derived terrestrial water storage depletion associated with the 2003 European heatwave. Geophys Res Lett 32:18 doi:10.1029/2005GL023574

Andersen OB, Knudsen P, Trimmer RG (2005b) Improved high resolution gravity field mapping (the KMS02 global marine gravity field). In: Sanso F (ed) A window on the future of geodesy, IAG symposium, vol 128, Springer-Verlag, Berlin Heidelberg, pp 326–331

Andersen OB, Egbert G, Erofeeva L, Ray R (2006) Mapping Non linear shallow water tides, a look at the past and future, Ocean Dyn: 1–17. Springer, doi:10.1007/s10236-006-0060-7

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0,

Andersen OB, Knudsen P, Berry P, Kenyon S, Factor JK (2010a) Recent development in high resolution global altimetric gravity field modeling. Lead Edge 29(5):540–545. ISSN: 1070-485X

Andersen OB, Knudsen P, Berry PAM (2010b) The DNSC08GRA global marine gravity field from double retracked satellite altimetry. J Geod 84(3):191–199. doi:10.1007/s00190-009-0355-9

Anzenhofer M, Gruber T, Rentsch M (1996) Global high resolution mean sea surface based on ERS 1 35 and 168 day cycles and TOPEX data. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations, IAG symposia, vol 116. Springer, Berlin/Heidelberg

Arabelos D, Tscherning CC (1990) Simulation of regional gravity field recovery from satellite gravity gradiometer data using collocation and FFT. In: Proceedings of the 1st international geoid commission symposium, Milan, 11–13 June 1990

Ardalan AA, Safari A (2004) Ellipsoidal terrain correction based on a multi-cylindrical equal-area map projection of the reference ellipsoid. J Geod 78:114–123

Arnold VI (1978) Mathematical methods of classical mechanics. Springer, Berlin

Awange JL, Grafarend EW (2005) Solving computational problems in geodesy and geoinformatics. Springer, Berlin

Axler S, Bourdon P, Ramey W (2001) Harmonic function theory. Graduate texts in mathematics, vol 137. Springer, Berlin

Bajracharya S (2003) Terrain effects on geoid determination. UCGE Report 20180, Department of Geomatics Engineering, University of Calgary, Calgary

Bajracharya S, Sideris MG (2004) The Rudzki inversion gravimetric reduction scheme in geoid determination. J Geod 78:272–282

Bajracharya S, Sideris MG (2005) Terrain-aliasing effects on gravimetric geoid determination. Geod Cart 54(1):3–16

Bajracharya S, Kotsakis C, Sideris MG (2002) Aliasing effects on terrain correction computation using constant and lateral density variations. Int Geoid Serv Bull 12:38–47

Baker EM (1988) A finite element model of the Earth's anomalous gravitational potential. Reports of the department of geodesy science and survey, vol 391. Ohio State University, Columbus

Balmino G, Moynot, B, Sarrailh M, Valès N (1987) Free air gravity anomalies over the oceans from Seasat and Geos 3 altimeter data. Eos Trans AGU 68(2):17–18

Bamler R (1999) The SRTM mission: a world-wide 30 m resolution DEM from SAR interferometry in 11 days. In: Fritsch D, Spiller R (eds) Photogrametric week 99. Wichmann, Heidelberg, pp 145–154

Barzaghi R, Forsberg R, Tscherning, CC (1988) A comparison between SEASAT, GEOSAT and gravimetric geoids computed by FFT and collocation in the central Mediterranean sea. In: Proceedings of the Chapman conference on progress in the determination of the Earth's gravity field, Fort Lauderdale, 13–16 Sept 1988, pp 96–99

Barzaghi R, Tselfes N, Tziavos IN, Vergos GS (2009) Geoid and high resolution sea surface topography modelling in the Mediterranean from gravimetry, altimetry and GOCE data: evaluation by simulation. J Geod 83:751–772

Bendat JS, Piersol AG (1980) Engineering applications of correlation and spectral analysis. Wiley, New York

Bendat JS, Piersol AG (2000) Random data analysis and measurement procedures, 3rd edn. Wiley, New York

Berry PAM, Garlick, JD, Freeman JA, Mathers EL (2005) Global inland water monitoring from multi-mission altimetry. Geophys Res Lett 32(16):L16401. doi:10.1029/2005GL022814

Biagi L, Sansò F (2001) TcLight: a new technique for fast RTC computation. In: Sideris MG (ed) Gravity, geoid and geodynamics 2000, IAG symposia, vol 123. Springer, Berlin/Heidelberg/New York

Bjerhammar A (1987) Discrete physical geodesy. Reports of the department of geodesy science and survey, vol 38. Ohio State University, Columbus

Bottoni GP, Barzaghi R (1993) Fast Collocation. Bulletin geodesique, 67:119–126

Bold GEJ (1985) A comparison of the time involved in computing fast Hartley and fast Fourier transforms. Proc IEEE 73(12):1863–1864

Borisenko AI, Tarapov IE (1979) Vector and tensor analysis with applications. Dover, New York
Bosch W (1993) A rigorous least squares combination of low and high degree spherical harmonics. Presented at the IAG general meeting, Beijing
Bosch W (2008) EOT08a model performances near coasts. In: Second coastal altimetry workshop, Pisa, Italy. http://www.coastalt.eu/pisaworkshop08
Bracewell RN (1984) The fast Hartley transform. Proc IEEE 72(8):1010–1018
Bracewell RN (1986a) The Fourier transform and its applications, 2nd edn. McGraw-Hill, New York
Bracewell RN (1986b) The Hartley transform. Oxford engineering science series, vol 19, 2nd edn. Oxford University Press, New York
Bracewell RN, Buneman O, Hao H (1986) Fast two-dimensional Hartley transform. Proc IEEE 74(9):1282–1283
Brigham EO (1988) The fast Fourier transform and its applications. Prentice Hall, Englewood Cliffs
Brown GS (1977). The average impulse response of a rough surface and its applications. IEEE Trans Antennas Propag 25(1):67–74
Challenor PG, Srokosz MA (1989) The extraction of geophysical parameters from radar altimeter return from a nonlinear ocean surface. In: Brooks SR (ed) Mathematics in remote sensing. Institute of Mathematics and Its Applications, Clarendon Press, Oxford, pp 257–268
Chan JC, Pavlis NK (1995) An efficient algorithm for computing high degree gravity field models using "wing block" diagonal technique. Internal Report HSTXG&G 9501, Hughes STX Corporation, Greenbelt
Chelton DB, Schlax MG (1994) The resolution capability of an irregularly sampled dataset: with application to geosat altimeter data. J Atmos Oceanic Tech 11:534–550
Cheng MK, Shum CK, Tapley BD (1997) Determination of long-term changes in the Earth's gravity field from satellite laser ranging observations. J Geophys Res 102(B10):22377–22390. doi:10.1029/97JB01740
Childers V, McAdoo D, Brozena J, Laxon S (2001) New gravity data in the Arctic Ocean: comparison of airborne and ERS gravity. J Geophys Res 106:8871–8886
Cimmino G (1952) Sulle equazioni lineari alle derivate parziali di tipo ellittico. Seminario matematico e teorico di Milano 23:183–203
Cimmino G (1955) Spazi hilbertiani di funzioni armoniche e questioni connesse. Equazioni lineari alle derivate parziali, UMI, pp 76–85
Colombo OL (1979) Optimal estimation from data regularly sampled on a sphere with applications in geodesy. Reports of the department of geodesy science and survey, vol 291. Ohio State University, Columbus
Colombo OL (1981a) Numerical methods for harmonic analysis on the sphere. Reports of the department of geodesy science and survey, vol 310. Ohio State University, Columbus
Colombo OL (1981b) Global geopotential modelling from satellite-to-satellite tracking. Reports of the department of geodesy science and survey, vol 317. Ohio State University, Columbus
Courant R, Hilbert D (1962) Methods of mathematical physics, vol 2. Wiley, New York
Cruz JY (1986) Ellipsoidal corrections to potential coefficients obtained from gravity anomaly data on the ellipsoid. Reports of the department of geodesy science and survey, vol 371. Ohio State University, Columbus
Davis RE, Foote FS, Anderson JM, Mikhail EM (1981) Surveying theory and practice, 6th edn. McGraw-Hill, New York
Deng X, Featherstone W, Hwang C, Berry PAM (2003) Waveform retracking of ERS-1. Mar Geod 25(4):189–204
Denker H (2005) Evaluation of SRTM3 and GTOPO30 terrain data in Germany. In: Jekeli C, Bastos L, Fernandes J (eds) Proceedings of the international association of geodesy symposia "gravity geoid and space missions", vol 129. Springer, Berlin/Heidelberg/New York, pp 218–223
Denker H, Rapp RH (1990) Geodetic and oceanographic results from the analysis of 1 year of Geosat data. J Geophys Res 95(C8):13151–13168

Denker H, Tziavos IN (1998) Investigation of the Molodensky series terms for terrain reduced gravity field data. Boll Geofis Teor Appl 40(3–4):195–203

Denker H, Barriot J-P, Barzaghi R, Fairhead D, Forsberg R, Ihde J, Kenyeres A, Marti U, Sarrailh M, Tziavos IN (2009) The development of the European gravimetric geoid model EGG07. In: Sideris MG (ed) Proceedings of the international association of geodesy symposia "observing our changing Earth", vol 133. Springer, Berlin/Heidelberg/New York, pp 177–185

Dermanis A, Rossikopoulos D (1991) Statistical inference in integrated geodesy. Presented at the IUGG XXth general assembly, international association of geodesy, Vienna, 11–24 Aug 1991

Doufexopoulou M, Czompo J (1988) Application of FFT to free air anomalies with lithospheric signal for modeling the disturbing gravity potential. Paper presented at the 6th international symposium on geodesy and physics of the Earth, Potsdam, 22–30 Aug 1988

Dowson M, Berry PAM (2006) Global analysis of multi-mission echoes over the earth's land surface from 15 years of altimeter missions. In: Proceedings of the Symposium of 15 years of progress in radar altimetry, ESA SP-614, ESA Publications Division, European Space Agency, Noordwijk, The Netherlands

Driscoll JR, Healy DM (1994) Computing Fourier transforms an deconvolutions on the 2-sphere. Adv Appl Math 15:202–250

Dudgeon DE, Mersereau RM (1984) Multidimensional digital signal processing. Prentice-Hall, Englewood Cliffs

Dunford N, Schwarz L (1958) Linear operators. Wiley, New York

Dziewonsky AM, Anderson DL (1981) Preliminary reference Earth model. Phys Earth Planet 25:297–356

Ellmann A, Vaníček P (2007) UNB application of Stokes-Helmert's approach to geoid computation. J Geodyn 43:200–213

Eren K (1980) Spectral analysis of GEOS-3 altimeter data and frequency domain collocation. Reports of the department of geodesy science and survey, vol 297. Ohio State University, Columbus

ESA SP-1233 (1) (1999) The four candidate Earth explorer core missions – Gravity field and steady-state ocean circulation mission. ESA Publications Division, ESTEC, Noordwijk

ESA (2000) From Eötvös to mGal – Final report, ESA/ESTEC Contract No. 13392/98/NL/GD, Graz

ETOPO (2008) The 2-minute gridded global relief data (ETOPO2v2). http://www.ngdc.noaa.gov/mgg/fliers/01mgg04.html. Accessed Mar 2008

Eymard L, Obligis E (2006) The altimetric wet troposphere correction: progress since the ERS-1 mission. In: Danesy D (ed) 15 years of progress in satellite altimetry, Venice, Italy, ESA SP-614. ISBN: 92-9092-925-1

Fairhead JD, Green CM, Fletcher KMU (2004) Global mapping deep-water hydrocarbon plays of the continental margins. ASEG 17th geophysical conference and exhibition, Sydney. doi:10.1071/ASEG2004ab041

Farr TG, Kobrick M (2000) Shuttle radar topography mission produces a wealth of data. EOS Trans Amer Geophys Un 81:583–585

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J (2007) The shuttle radar topography mission. Rev Geophys 45:RG2004. doi:10.1029/2005RG000183

Featherstone WE, Dentith MC (1997) A geodetic approach to gravity data reduction for geophysics. Comput Geosci 23(10):1063–1070

Featherstone WE, Kirby JF (2000) The reduction of aliasing in gravity anomalies and geoid heights using digital terrain data. Geophys J Int (141):204–212

Featherstone WE, Kirby JF, Hirt C, Filmer MS, Claessens SJ, Brown N, Hu G, Johnston GM (2011) The AUSGeoid2009 model of the Australian height datum. J Geod 85(3):133–150

Fichera G (1948) Teoremi di completezza sulla frontiera di un dominio per alcuni sistemi di funzioni. Ann Mat Pura Appl 27:1–28

Fischbach E, Talmadge CL (1999) The search of non-Newtonian gravity. Springer, New York

Forsberg R (1984) A study of terrain corrections, density anomalies and geophysical inversion methods in gravity field modelling. Reports of the department of geodesy science and survey, vol 355. Ohio State University, Columbus

Forsberg R (1985) Gravity field terrain effect computation by FFT. Bull Géod 59:342–360

Forsberg R (1986) Spectral poperties of the gravity field in the Nordic countries. Boll Geod Sc Aff 45:361–384

Forsberg R (1987) A new covariance model for intertial gravimetry and gradiometry. J Geophys Res 92(NB2):1305–1310

Forsberg R (1993) Modelling the fine-structure of the geoid: methods, data requirements and some results. Surv Geophys 14(4–5):403–418

Forsberg R (2008) Terrain effects in geoid computations. In: Lecture notes, international school for the determination and use of the geoid, Como, 15–19 Sept 2008

Forsberg R (2010) Terrain effects in geoid computations. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg, 28 June–2 July 2010

Forsberg R, Sideris MG (1989) On topographic effects in gravity field approximation. In: Festschrift to Torben Krarup. Danish Geodetic Institute, Copenhagen, pp 129–148

Forsberg R, Sideris MG (1993) Geoid computations by the multi-banding spherical FFT approach. Man Geod 18:82–90

Forsberg R, Solheim D (1988) Performance of FFT methods in local gravity field modeling. In: Proceedings of the Chapman conference on progress in the determination of the Earth's gravity field, Fort-Lauderdale, 13–16 Sept 1988, pp l00–103

Forsberg R, Tscherning CC (1981) The use of height data in gravity field approximation by collocation. J Geophys Res 86(B9):7843–7854

Forsberg R, Tscherning CC (1997) Topographic effects in gravity field modelling for BVP. In: Sansò F, Rummel R (eds) Geodetic boundary value problems in view of the one centimetre geoid. Lecture notes in Earth sciences, vol 65. Springer, Berlin/Heidelberg/New York, pp 241–272

Forsberg R, Tscherning CC (2008) An overview manual for the GRAVSOFT geodetic gravity field modelling programs, 2nd edn. Contract Report to JUPEM, Aug 2008

Förste C, Schmidt R, Stubenvoll R, Flechtner F, Meyer U, König R, Neumayer H, Biancale R, Lemoine J-M, Bruinsma S, Loyer S, Barthelmes F, Esselborn S (2008) The GeoForschungsZentrum Potsdam/Groupe de Recherche de Gèodésie Spatiale satellite-only and combined gravity field models: EIGHENGL04S1 and EIGEN-GL04C, J Geod 82(6):331–346. doi:10.1007/s00190-007-0183-8

Fotopoulos G (2003) An analysis on the optimal combination of geoid, orthometric and ellipsoidal height data. PhD Thesis, University of Calgary, Department of Geomatics Engineering, Report No. 20185, Dec 2003

Fotopoulos G (2005) Calibration of geoid error models via a combined adjustment of ellipsoidal, orthometric and gravimetric geoid height data. J Geod 79(1–3). doi:10.1007/s00190-005-0449-y

Fotopoulos G (2008) Matching the gravimetric geoid to the GPS-leveling undulations. In: Lecture notes, international school for the determination and use of the geoid, Como, 15–19 Sept 2008

Fotopoulos G (2010) Matching the gravimetric geoid to the GPS-leveling undulations. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg, 28 June –2 July 2010

Fotopoulos G, Sideris MG (2005) Spatial modeling and analysis of adjusted residuals over a network of GPS-levelling benchmarks. Geomatica 59(3):251–262

Freeden W, Schreiner M (2009) Spherical functions of mathematical geosciences. In: Solving algebraic impact. Springer, Berlin

Fu LL, Cazenave A (2001) Satellite altimetry and earth sciences, Academic Press, New York, P 486

Fu LL, Christensen E, Yamarone C Jr, Lefebvre M, Ménard Y, Dorrer M, Escudier P (1994) TOPEX/POSEIDON: mission overview. J Geophys Res 99(C12):24369–24381

Ganachaud A, Wunsch C, Kim M-C, Tapley B (1997) Combination of TOPEX/POSEIDON data with a hydrographic inversion for determination of the oceanic general circulation and its relation to geoid accuracy. Geophys J Int 128:708–722

Gleason DM (1988) Comparing ellipsoidal corrections to the transformation between the geopotential's spherical and ellipsoidal spectrums. Man geod 13:114–129

Goiginger H, Hoeck E, Rieser D, Mayer-Guerr T, Maier A, Krauss S, Pail R, Fecher T, Gruber T, Brockmann JM, Krasbutter I, Schuh W-D, Jaeggi A, Prange L, Hausleitner W, Baur O, Kusche J (2011) The combined satellite-only global gravity field model GOCO02S. Presented at the general assembly of the European geosciences union, Vienna, 4–8 Apr 2011

GRACE (1998) Gravity recovery and climate experiment: science and mission requirements document, revision A, JPLD-15928, NASA's Earth system science pathfinder program

Grafarend E (1975) Cartan frames and a foundation of physical geodesy. In: Brosowski E, Martensen F (eds) Methoden un Verfahren der Matematischen Physik Bd 12

Grafarend E (1986) Differential geometry of the gravity field. In: Sansò F (ed) Proceedings of 1st Hotine-Marussi symposium on mathematical geodesy, Ist Topogr Mat Geof, Politecnico di Milano, pp 5–14

Grafarend EW (2006) Linear and nonlinear models. Walter de Gruyter, Berlin

Grafarend EW, Ardalan AA, Sideris MG (1999) The spheroidal fixed-free two-boundary-value problem for geoid determination (the spheroidal Bruns transform). J Geod 73:513–533

Gruber T, Bosch W (1992) A new 360 gravity field model. Presented at the XVII general assembly of the European geophysical society, Edinburgh

Gruber T, Anzenhofer M, Rentsch M (1996) The 1995 GFZ high resolution gravity model. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations, IAG symposia, vol 116. Springer, Berlin/Heidelberg

Haagmans RHN, Van Gelderen M (1991) Error variances-covariances of GEM-T1: their characteristics and implications in geoid computation. J Geophys Res 96(B12):20011–20022

Haagmans R, de Min E, Van Gelderen M (1993) Fast evaluation of convolution integrals on the sphere using 1D FFT, and a comparison with existing methods for Stokes' integral. Man Geod 18:227–241

Hanssen RH (1993) The application of Radon transformation for the analysis of sparsely sampled data, TU Delft, Faculty of geodetic engineering report No. 93.1, Delft

Harris FJ (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE 66(1):51–83

Harrison JC, Dickinson M (1989) Fourier transform methods in local gravity field modelling. Bull Geod 63:149–166

Hartley RVL (1942) A more symmetrical Fourier analysis applied to transmission problems. Proc IRE 30:144–150

Hartung J (1981) Non-negative minimum biased invariant estimation in variance component models. Ann Stat 9(2):278–292

Haxby WF (1983) Gravity field of the worlds oceans (Seasat altimetry), map. National Geophysical Data Center, Boulder

Heck B (1990) An evaluation of some systematic error sources affecting terrestrial gravity anomalies. Bull Geod 64:88–108

Heck B (2003a) Integral equation methods in physical geodesy. In: Grafarend EW, Krumm FW, Schwarze VS (eds) Geodesy: the challenge of the third millenium. Springer, Berlin

Heck B (2003b) On Helmert's methods of condensation. J Geod 77(3–4):155–170

Heck B, Seitz F (2007) A comparison of the tesseroid, prism and point mass approaches for mass reductions in gravity field modelling. J Geod 81(2):121–136

Heiskanen WA, Moritz H (1967) Physical geodesy. Freeman, San Francisco

Helmert FR (1884) Die Mathematischen and physikalisken Theorien der höheren Geodäsie. Vol 2. Leipzig, B. G. Teubner

Helmert FR (1924) Die Ausgleichungsrechnung nach der Methode der Kleinsten Quadratic, 3. Aufl. Teubner, Leipzig/Berlin

Hernandez F, Schaeffer P (2000) Altimetric mean sea surfaces and gravity anomaly maps and intercomparison. AVISO Tech Rep, AVI-NT-011-5242, CLS CNES, Toulouse
Hipkin RG (1988) Bouguer anomalies and the geoid: a reassessment of Stokes's method. Geophys J Int 92:53–66
Hirt C, Flury J (2008) Astronomical-topographic levelling using high-precision astrogeodetic vertical deflections and digital terrain model data. J Geod 82(4–5):231–248
Hirt C, Featherstone WE, Marti U (2010) Combining EGM2008 and SRTM/DTM2006.0 residual terrain model data to improve quasigeoid computations in mountainous areas devoid of gravity data. J Geod 84:557–567
Hobson EW (1955) The theory of spherical and ellipsoidal harmonics. Cambridge Univ Press, New York
Hofmann-Wellenhof B, Moritz H (1986) Introduction to spectral analysis. In: Sünkel H (ed) Mathematical and numerical techniques in physical geodesy. Springer, Berlin
Hofmann-Wellenhof B, Moritz H (2005) Physical geodesy. Springer, Vienna/New York
Holmes SA, Pavlis NK (2006) A FORTRAN program for very-high-degree harmonic synthesis, HARMONIC_SYNTH . http://earth-info.nga.mil/GandG/wgs84/gravitymod/new_egm/new_egm.html
Holmes SA, Pavlis NK (2007) Some aspects of harmonic analysis of data gridded on the ellipsoid. In: Gravity field of the Earth, proceedings of the 1st international symposium of the international gravity field service (IGFS), Istanbul. Harita Dergisi 18(Special issue):151–156
Holota P (1983a) The altimetry gravimetry boundary value problem, I: linearization, Friedrics iequality. Boll Geod Sci Aff 42:14–32
Holota P (1983b) The altimetry gravimetry boundary value problem, II: weak solution, V-ellipticity. Boll Geod Sci Aff 42:70–84
Holota P, Nevadba O (2003) Domain transformation, boundary problems and optimization concepts in the combination of terrestrial and satellite gravity field data. In: IAG Symposia 133. Springer, Berlin/Heidelberg
Hörmander L (1976) The boundary problems of physical geodesy. Arch Ration Mech Anal 62:1–52
Hotine M (1969) Mathematical geodesy. ESSA monographs (Environmental science services Adm – US department of commerce). Mon. N2-1969
Hsu HP (1970) Fourier analysis, Simon and Schuster, New York
Hsu HP (1984) Applied Fourier analysis. Harcourt Brace Jovanovich, San Diego.
Huang J, Vaníček P, Pagiatakis SD, Brink W (2001) Effect of topographical density on geoid in the Canadian Rocky Mountains. J Geod 74:805–815
Huang J, Sideris MG, Vaníček P, Tziavos IN (2003) Numerical investigation of downward continuation techniques for gravity anomalies. Boll Geod Sci Affini 62(1):33–48
Hwang C (1991) Orthogonal functions over the oceans and applications to the determination of orbit error, geoid and sea surface topography from satellite altimetry. Reports of the department of geodesy science and survey, vol 414. Ohio State University, Columbus
Hwang C (1998) Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea. J Geod 72:304–312
Hwang C, Hsu H (2003) Marine gravity anomaly from satellite altimetry; a comparison of methods over shallow waters. In: Proceedings: international workshop on satellite altimetry for geodesy, geophysics and oceanography. IAG symposium vol 126, Springer, Berlin/Heidelberg, pp 59–66
Hwang C, Parsons B (1995) Gravity anomalies derived from seasat, geosat, ERS-1 and TOPEX/POSEIDON altimetry and ship gravity: a case study over the Reykjanes Ridge. Geophys J Int 122:551–568
Hwang C, Hsu H, Jang R (2002) Global mean sea surface and marine gravity anomaly from multi-satellite altimetry: applications of deflection-geoid and inverse Vening Meinesz formulae. J Geod 76(8):407–418
Hwang C, Wang C-G, Hsiao Y-S (2003) Terrain correction computation using Gaussian quadrature. Comput Geosci 29:1259–1268

Hwang C, Hsiao Y-S, Shih H-C, Yang M, Chen K-H, Forsberg R, Olesen AV (2007) Geodetic and geophysical results from a Taiwan airborne gravity survey: data reduction and accuracy assessment. J Geophys Res 112:B04407. doi:10.1029/2005JB004220

IEEE (1967) Special issue of the fast Fourier transform and its applications to digital filtering and spectral analysis. IEEE Trans Audio Electroacoust AU-15(2):43–117

Jekeli C (1982) Optimizing kernels of truncated integral formulas in physical geodesy. In: Proceedings of the lAG general meeting, Tokyo, 7–15 May 1982

Jekeli C (1987) The downward continuation of aerial gravimetric data without density hypothesis. Bull Geod 61:319–329

Jekeli C (1988) The exact transformation between ellipsoidal and spherical harmonic expansions. Man Geod 13:106–113

Jekeli C (1991) The statistics of the Earth's gravity field, revisited. Man Geod 16:313–325

Jekeli C (1996) Spherical harmonic analysis, aliasing and filtering. J Geod 70:214–223

Jekeli C (1999a) The determination of gravitational potential differences from satellite-to-satellite tracking. Celest Mech Dyn Astr 75(2):85–100

Jekeli C (1999b) An analysis of vertical deflections derived from high-degree spherical harmonic models. J Geod 73(1):10–22. doi:10.1007/s001900050213

Jekeli C (2005) Spline representations of functions on a sphere for geopotential modeling. Reports of the department of geodesy science and survey, vol 475. Ohio State University, Columbus

Jekeli C, Serpas JG (2003) Review and numerical assessment of the direct topographical reduction in geoid determination. J Geod 77:226–239

Jekeli C, Zhu L (2006) Comparisons of methods to model the gravitational gradients from topographic data bases. Geophys J Int 166:999–1014

Jensen H, Raney K (1998) Delay Doppler radar altimeter: better measurement precision. In: Proceedings IEEE geoscience and remote sensing symposium IGARSS'98, Seattle. IEEE: 2011–2013

Jiang Z, Duquenne H (1996) On the combined adjustment of a gravimetrically determined geoid and GPS levelling stations. J Geod 70:505–514

Jordan SK (1978) Fourier physical geodesy, Report No. AFGL-TR-78-0056. The Analytical Sciences Corporation, Reading

Kallianpur G (1980) Stochastic filtering theory. Applications of mathematics, vol 13, Springer, New York

Katsambalos KE (1979) The effect of the smoothing operator on potential coefficient determinations. Reports of the department of geodesy science and survey, vol 287. Ohio State University, Columbus

Kaula WM (1966) Tests and combination of satellite determinations of the gravity field with gravimetry. J Geophys Res 71:5303–5314

Kaula WM (2000) Theory of satellite geodesy. Dover, New York

Kearsley AHW, Sideris MG, Krynski J, Forsberg R, Schwarz KP (1985) White sands revisited – a comparison of techniques to predict deflections of the vertical, UCSE Report No. 30007, Department of Surveying Engineering, The University of Calgary, Calgary

Kellog OD (1953) Foundations of potential theory. Dover, New York

Kenyon SC, Forsberg R (2008) New gravity field for the Arctic. EOS 89:32, 289

Kenyon SC, Pavlis NK (1996) The development of a global surface gravity data base to be used in the joint DMA/GSFC geopotential model. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations, IAG symposia, vol 116. Springer, Berlin

Kiamehr R, Sjöberg LE (2005) Effect of the SRTM global DEM on the determination of a high-resolution geoid model: a case study in Iran. J Geod 79:540–551

Kim J-H (1996) Improved recovery of gravity anomalies from dense altimeter data. Rep No 444, Department of Geodetic Sciences and Surveying, The Ohio State University, Columbus, 130pp

Kim MC, Tapley BD, Shum C-K, Ries JC (1995) Center for space research mean sea surface model. Presented at the TOPEX/POSEIDON working team meeting, Pasadena

Kirby JF, Featherstone WE (1999) Terrain correcting the Australian gravity observations using the national digital elevation model and the fast Fourier transform. Aust J Earth Sci 46:555–562

Kirsch A (1996) An introduction to the mathematical theory of inverse problems. Applied mathematical sciences, vol 120. Springer, New York
Klose U, Ilk KH (1992) A solution to the singularity problem occurring in the terrain correction formula. Paper submitted to manuscripta geodaetica
Knudsen P (1987a) Estimation and modelling of the local empirical covariance function using gravity and satellite altimeter data. Bull Geod 61:145–160
Knudsen P (1987b) Adjustment of satellite altimeter data from cross-over differences using covariance relations for the time varying components represented by Gaussian functions. In: Proceedings of IAG symposia. International Association of Geodesy, Paris, pp 617–628
Knudsen P (1991) Simultaneous estimation of the gravity field and sea surface topography from satellite altimeter data by least squares collocation. Geophys J Int 104:307–317
Knudsen P (1993) Geodesy and geophysics. In: KakkuriJ (ed) lecture notes for NKG autumn school. Korpilampi, Finland, pp 87–126
Knudsen P, Brovelli M (1991) Collinear and cross-over adjustment of geosat ERM and seasat altimeter data in the Mediterranean Sea. Surv Geophys, 14(4):449–459, 1993
Knudsen P, Andersen OB, Tscherning CC (1992) Altimetric gravity anomalies in the Norwegian-Greenland Sea - preliminary results from the ERS-1 35 days repeat mission. Geophys Res Lett 19(17):1795–1798
Koch KR (1987) Parameter estimation and hypothesis testing in linear models. Springer, Berlin
Koch KR (1999) Parameter estimation and hypothesis testing in linear models, 2nd edn. Springer, Berlin/Heidelberg
Kotsakis C, Sideris MG (1999): On the adjustment of combined GPS/levelling/geoid networks. J Geod 73(8):412–421
Kotsakis C, Sideris MG (2001) A modified Wiener-type filter for geodetic estimation problems with non-stationary noise. J Geod 75(12):647–660
Krarup T (2006) In: Borre K (ed) Mathematical foundations of geodesy. Springer, Berlin
Kuhn M (2000) Geoidbestimmung unter Verwendung verschiedener Dichtehypothesen. Dissertation. DGK, Reihe C, Nr. 520
Kuhn M (2003) Geoid determination with density hypotheses from isostatic models and geological information. J Geod 77:50–65
Kuhn M, Featherstone WE, Kirby JF (2009) Complete spherical Bouguer gravity anomalies over Australia. Aust J Earth Sci 56:213–223
Kuhtreiber N (1990) Untersuchungen zur gravimetrischen Bestimmung von Lotabmeichungen im Hochgebirge nach Molodensky mittels Fast-Fourier-Transformation. Technische Universitat Graz
Kulhanek O (1976) Introduction to digital filtering in geophysics. Elsevier, Amsterdam
Lauritzen S (1973) The probabilistic background of some statistical methods in physical geodesy. Meddelelse (geodætisk institut (Denmark)), vol 48. Geodætisk Institut, København
Laxon S, McAdoo D (1998) Satellites provide new insights into polar geophysics, EOS. Trans AGU 79(6):69–72
Lemoine FG, Kenyon SC, Factor JK, Trimmer RG, Pavlis NK, Chinn DS, Cox CM, Klosko SM, Luthcke SB, Torrence MH, Wang YM, Williamson RG, Pavlis EC, Rapp RH, Olson TR (1998) The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96. NASA/TP-1998-206861. Goddard Space Flight Center, Greenbelt
Lerch FJ (1991) Optimum data weighting and error calibration for estimation of gravitational parameters. Bull Geod 65:44–52
Lerch FJ, Klosko SM, Laubscher RE, Wagner CA (1979) Gravity model improvement using GEOS 3 (GEM9 and 10). J Geophys Res 84(B8):3897–3916
Lerch FJ, Pavlis NK, Chan JC (1993) High degree gravitational modeling: quadrature formulæversus a block diagonal normal matrix inversion. Presented at the XVIII general assembly of the European geophysical society, Wiesbaden
Levitus S (1982) Climatological atlas of the world ocean. NOAA professional paper vol 13. US Government Print Office, Washington, p 173

Li YC (1993) Optimized spectral geoid determination. UCGE Report No. 20050. Department of Geomatics Engineering, The University of Calgary, Calgary
Li YC, Sideris MG (1992) The fast Hartley transform and its application in physical geodesy. Man Geod 17:381–387
Li YC, Sideris MG (1994) Improved gravimetric terrain corrections. Geophys J Int 119(3):740–752
Lillibridge JL, Smith WHF, Scharrroo R, Sandwell DT (2004) The geosat geodetic mission 20th anniversary data product. AGU, 85(47). Fall Meet Suppl, Abstract SF43A–0786.
Liu QW, Li YC, Sideris MG (1997) Evaluation of deflections of the vertical on the sphere and the plane a comparison of FFT techniques. J Geod 71(8):461–468
Luthcke SB, Rowlands DD, Lemoine FG, Klosko SM, Chinn D, McCarthy JJ (2006) Monthly spherical harmonic gravity field solutions determined from GRACE inter satellite range rate data alone. Geophys Res Let 33:L02402. doi:10.1029/2005GL024846
MacMillan WD (1958) The theory of the potential. In: Theoretical mechanics, vol 2. Dover, New York
Mainville A, Forsberg R, Sideris MG (1992) Global positioning system testing of geoids computed from geopotential models and local gravity data: a case study. J Geophys Res 97(B7):11137–11147
Makhloof AA (2007) The use of topographic-isostatic mass information in geodetic application. Disseratation D98, Institute of Geodesy and Geoinformation, Bonn
Makhloof AA, Ilk KH (2006) Effects of topographic-compensation masses on gravitational functionals at the surface of the Earth, at airborne and satellite altitudes. J Geod 82(2):93–111
Marchenko AN (1998) Parametrizaton of the Earth gravity field. Astronomical and Geodetic Society, Lviv
Marks RJ Jr (1991) Introduction to Shannon sampling and interpolation theory. Springer, New York
Marsh JG, Lerch JF, Koblinsky CJ, Klosko SM, Robbins JW, Williamson RG, Patel GB (1990) Dynamic sea surface topography, gravity, and improved orbit accuracies from the direct evaluation of SEASAT altimeter data. J Geophys Res 95(C8):13129–13150
Marti U (2004) Comparison of SRTM data with national DTMs of Switzerland. Presented in the gravity geoid and space missions, International Association of Geodesy Symposia, Porto
Marti U, Schlatter A, Brockmann E, Wiget A (2000) The way to a consistent national height system for Switzerland. In: Adam J, Schwarz K-P (eds) International association of geodesy symposia, vistas for geodesy in the new millennium, vol 125. Springer, New York
Martinec Z, Matyska C, Grafarend EW, Vanicek P (1993) On Helmert's 2nd condensation method. Man Geod 18:417–421
Martinec Z (1998) Boundary value problems for gravimetric determination of a precise geoid. Lecture notes in Earth sciences, vol 73. Springer, Heidelberg/Berlin
Martinec Z, Vaníček P (1994) The indirect effect of topography in the Stokes-Helmert technique for a spherical approximation of the geoid. Man Geod 19:213–219
Martinec Z, Vaníček P, Mainville A, Vèronneau M (1994) Evaluation of topographical effects in precise geoids computation from densely sampled heights. J Geod 70:746–754
Marussi A (1985) Intrinsic geodesy. Springer, Berlin
Maus S, Green CM, Fairhead D (1998) Improved ocean-geoid resolution from retracked ERS-1 satellite altimeter waveforms. Geophys J Int 134(1):243–253
Mayer-Gürr T (2007) ITG-Grace03s: the latest GRACE gravity field solution computed in Bonn. Presented at the joint international GSTM and SPP symposium, Potsdam, 15–17 Oct 2007. http://www.geod.uni-bonn.de/itg-grace03.html
Mayer-Gürr T, Eicker A, Ilk KH (2007) ITG-Grace02s: a GRACE gravity field derived from range measurements of short arcs. In: Gravity field of the Earth, proceedings of the 1st international symposium of the international gravity field service (IGFS), Istanbul. Harita Dergisi 18(Special issue):193–198
Mazzega P, Houry S (1986) An experiment to invert Seasat altimetry for the Mediterranean and Black Sea mean surface. Geophys J 96:259–272. Rosborough GW (1989) Satellite orbit perturbations due to the geopotential. Rep No CSR-86-1, Center for Space Research, The University of Texas, Austin

Meckelburg HJ (1985) Fast Hartley transform algorithm. Electron Lett 21(8):341–343
Meissl P (1981) The use of finite elements in physical geodesy. Reports of the department of geodesy science and survey, vol 313. Ohio State University, Columbus
Merry CL (2003) DEM-induced errors in developing a quasi-geoid model for Africa. J Geod 77:537–542
Mesko A (1984) Digital filtering: applications in geophysical exploration for oil. Akademiai Kiado, Budapest
Mikhlin SG (1957) Integral equations. Pergamon, Oxford
Mikhlin SG (1964) Variational methods in mathematical physics. Pergamon, Oxford
Miranda C (1970) Partial differential equations of elliptic type. Springer, Berlin
Molodensky MS, Ermeev VF, Yurkina MI (1962) Methods for the study of the gravitational field of the Earth. Translated from Russian (1960), Israel program for scientific translations, Jerusalem
Moritz H (1980) Advanced physical geodesy, 2nd edn. Wichmann, Karlsruhe
Moritz H (1990) The figure of the Earth. Wichman, Karlsruhe
Moritz H (2000) Geodetic reference system 1980. In: The Geodesist's handbook 2000. Springer, Berlin. J Geod 74:128–133
Nagy D (1966) The prism method for terrain corrections using digital computers. Pure Appl Geophys 63:31–39
Nagy D (1980) Gravity anomalies, deflections of the vertical and geoidal heights for a three-dimensional model. Acta Geodet et Montanist Acad Sci Hung 15(1):17–26
Nagy D (2006) The prism method for terrain corrections using digital computers. Pure Appl Geophys 63:31–39
Nagy D, Papp G, Benedek J (2000) The gravitational potential and its derivatives for the Prism. J Geod 74(7–8):552–560. Erratum in J Geod 76(8):475–475
Nahavandchi H (2000) The direct topographical correction in gravimetric geoid determination by the Stokes-Helmert method. J Geod 74:488–496
Nash RA, Jordan SK (1987) Statistical geodesy – an engineering perspective. Proc IEEE 66:532–550
Nerem RS, Lerch FJ, Marshall JA, Pavlis EC, Putney BH, Tapley BD, Eanes RJ, Ries JC, Schutz BE, Shum CK, Watkins MM, Klosko SM, Chan JC, Luthcke SB, Patel GB, Pavlis NK, Williamson RG, Rapp RH, Biancale R, Nouel F (1994) Gravity model development for TOPEX/POSEIDON: joint gravity models 1 and 2. J Geophys Res 99(C12):24421–24447
Nikiforov AF, Uvarov VB (1988) Special functions of mathematical physics. Birkhäuser, Basel/Boston
Novák P, Vaníček P, Martinec P, Vèronneau M (2001) Effects of the spherical terrain on gravity and the geoid. J Geod 75:491–504
Novák P, Kern M, Schwarz K-P, Heck B (2003) Evaluation of band-limited topographical effects in airborne gravimetry. J Geod 76:597–604
Nowell DAG (1999) Gravity terrain corrections – an overview. J Appl Geophys 42:117–134
Olesen AV (2003) Improved airborne scalar gravimetry for regional gravity field mapping and geoid determination. Tech Rep 24, National Survey and Cadastre, Copenhagen, 54pp. ISBN: 87-7866-383-0
Olesen AV, Andersen OB, Tscherning CC (2002) Merging of airborne gravity and gravity derived from satellite altimetry: test cases along the coast of Greenland. Stud Geophys Geod 46:387–394
Olgiati A, Balmino G, Sarrailh M, Green, CM (1995) Gravity anomalies from satellite altimetry: comparison between computation via geoid heights and via deflections of the vertical. Bull Geod 69(4):252–260
Omang ODC, Forsberg R (2000) How to handle topography in practical geoid determination: three examples. J Geod 74:458–466
Oppenheim AV, Schafer RW (1989) Discrete-time signal processing. Prentice Hall, Englewood Cliffs
Pagiatakis SD, Frazer D, McEwen K, Goodacre AK, Vèronneau M (1999) Topographic mass density and gravimetric geoid modelling. Boll di Geofis Teor ed Appl 40:189–194

Pail R, Sansò F, Reguzzoni M, Kütreiber N (2010) On the combination of global and local data in collocation theory. Studia Geoph Geod 54(2):195–218
Papoulis A (1977) Signal analysis. McGraw-Hill, New York
Papoulis A (1990) Probability and Statistics. Prentice-Hall, Inc.
Parker RL (1995) Improved Fourier terrain correction, Part I. Geophysics 60:1007–1017
Parker RL (1996) Improved Fourier terrain correction, Part II. Geophysics 61:365–372
Pavlis NK (1988) Modeling and estimation of a low degree geopotential model from terrestrial gravity data. Reports of the department of geodesy science and survey, vol 386. Ohio State University, Columbus
Pavlis NK (1996) Global geopotential solutions to degree 500: preliminary results. Presented at the XXI general assembly of the European geophysical society, Den Haag
Pavlis NK (1998) Observed inconsistencies between satellite-only and surface gravity-only geopotential models. In: Forsberg R, Feissel M, Dietrich R (eds) Geodesy on the move – Gravity, geoid, geodynamics and Antarctica, IAG symposia, vol 119. Springer, Berlin
Pavlis NK (2000) On the modeling of long wavelength systematic errors in surface gravimetric data. In: Schwarz KP (ed) Geodesy Beyond 2000 – The challenges of the first decade, IAG symposia, vol 121. Springer, Berlin
Pavlis NK (2008) Development and applications of geopotential models. In: Lecture notes, international school for the determination and use of the geoid, Como, 15–19 Sept 2008
Pavlis NK (2010) Development and applications of geopotential models. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg, 28 June–2 July 2010
Pavlis NK, Kenyon SC (2003) Analysis of surface gravity and satellite altimetry data for their combination with CHAMP and GRACE information. In: Tziavos IN (ed) Gravity and geoid 2002, Proceedings of the 3rd meeting of the international gravity and geoid commission, Ziti, Thessaloniki
Pavlis NK, Rapp RH (1990) The development of an isostatic gravitational model to degree 360 and its use in global gravity modelling. Geoph J Int 100:369–378
Pavlis NK, Saleh J (2005) Error propagation with geographic specificity for very high degree geopotential models. In: Jekeli C, Bastos L, Fernandes J (eds) Gravity, geoid and space missions, IAG symposia, vol 129. Springer, Berlin
Pavlis NK, Klosko SM, Rapp RH (1993) Intercomparison of contemporary gravitational models. Presented at the XVIII general assembly of the European geophysical society, Wiesbaden
Pavlis NK, Chan JC, Lerch F (1996a) Alternative estimation techniques for global high degree gravity modeling. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations, IAG symposia, vol 116. Springer, Berlin
Pavlis NK, Chan JC, Rapp RH, Smith DE (1996b) Recent improvements in the high degree representation of the Earth's gravitational potential. Presented at the Spring AGU meeting, Baltimore
Pavlis NK, Holmes SA, Kenyon SC, Schmidt D, Trimmer R (2005) A preliminary gravitational model to degree 2160. In: Jekeli C, Bastos L, Fernandes J (eds) Gravity, geoid and space missions, IAG symposia, vol 129. Springer, Berlin
Pavlis NK, Factor JK, Holmes SA (2007a) Terrain-related gravimetric quantities computed for the next EGM. In: Gravity field of the Earth, proceedings of the 1st international symposium of the international gravity field service (IGFS), Istanbul. Harita Dergisi 18(Special issue):31–323
Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2007b) Earth gravitational model to degree 2160: status and progress. Paper presented at XXIV general assembly of the international union of geodesy and geophysics (IUGG), Perugia, pp 2–13
Pavlis NK, Holmes SA, Kenyon, SC, Factor JK (2008) Earth gravitational model to degree 2160 EGM2008. Paper presented to the European geosciences union general assembly, Vienna, Apr 2008
Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2012), The development and evaluation of the Earth Gravitational Model 2008 (EGM2008), J. Geophys. Res., 117, B04406, doi:10.1029/2011JB008916.

Peng M (1994) Topographic effects on gravity and gradiometry by the 3D FFT and FHT methods. M.Sc., Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary
Peng M, Li YC, Sideris MG (1995) First results on the computation of terrain corrections by the 3D-FFT method. Man Geod 20(6):475–488
Prentice A (1990) Probability and statistics. Prentice-Hall, Englewood Cliffs
Pizzetti P (1894) Sulla espressione della gravità alla superficie del geoide, supposto ellissoidico. Atti Reale Accad Naz Lincei V3:166–172
Poutanen M, Vermeer M, Mäikinen J (1996) The permanent tide in GPS positions. J Geod 70:499–504
Rabiner LR, Rader CM (eds) (1972) Digital signal processing. IEEE Press, New York
Rabus B, Eineder M, Roth A, Bamler A (2003) The shuttle radar topography mission – a new class of digital elevation models acquired by spaceborne radar. Photogramm Rem Sens 57:241–262
Rao CR (1970) Estimation of heterogeneous variances in linear models. J Amer Stat Assoc 65:161–172
Rao CR (1971) Estimation of variance components – MINQUE theory. J Multivariate Stat 1:257–275
Rao CR, Kleffe J (1988) Estimation of variance components and applications. North-Holland series in statistics and probability, vol 3. North Holland, Amsterdam
Rapp RH (1967) The geopotential to (14, 14) from a combination of satellite and gravimetric data. Presented at the XIV general assembly of IUGG/IAG, Lucerne
Rapp RH (1981) The Earth's gravity field to degree and order 180 using Seasat altimeter data, terrestrial gravity data, and other data. Reports of the department of geodesy science and survey, vol 322. Ohio State University, Columbus
Rapp RH (1993) Use of altimeter data in estimating global gravity models. In: satellite altimetry in geodesy and oceanography. Lecture notes in earth sciences, vol 50. Springer, Berlin.
Rapp RH (1997a) Global models for the one-centimeter geoid, present status and near term perspectives. In: Geodetic boundary value problems in view of the one-centimeter geoid. Lecture notes in Earth science, vol 65. Springer, Berlin
Rapp RH (1997b) Use of potential coefficient models for geoid undulation determinations using a spherical harmonic representation of the height anomaly/geoid undulation difference. J Geod 71:282–289
Rapp RH (1998) Past and future developments in geopotential modeling. In: Forsberg R, Feissel M, Dietrich R (eds) Geodesy on the move – gravity, geoid, geodynamics and antarctica, IAG Symposia, vol 119. Springer, Berlin
Rapp RH, Balasubramania N (1992) A conceptual formulation of a world height system. Reports of the department of geodesy science and survey, vol 421. Ohio State University, Columbus
Rapp RH, Basic T (1992) Oceanwide gravity anomalies from GEOS 3, SEASAT and GEOSAT altimeter data. Geophys Res Lett 19(19):1979–1982
Rapp RH, Cruz JY (1986a) Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30′ mean anomalies. Reports of the department of geodesy science and survey, vol 376. Ohio State University, Columbus
Rapp RH, Cruz JY (1986b) The representation of the Earth's gravitational potential in a spherical harmonic expansion to degree 250. Reports of the department of geodesy science and survey, vol 372. Ohio State University, Columbus
Rapp RH, Pavlis NK (1990) The development and analysis of geopotential coefficient models to spherical harmonic degree 360. J Geophys Res 95(B13):21885–21911
Rapp RH, Wang YM, Pavlis NK (1991) The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models. Rep No 410, Department of Geodetic Sciences and Surveying, The Ohio State University, Columbus
Raney, KR (2009) An overview of the future of coastal altimetry. In 3rd coastal altimetry workshop, ESRIN. ESA available from: http://www.congrex.nl/09c32/Talks-Files-PDF/05-Future-Coastal-Altimetry-3rd-CA-WS-Raney.pdf
Ray RD (2001) Inversion of oceanic tidal currents from measured elevations. J Mar Syst 28:1–18
Ray RD (2006) Secular changes of the M_2 tide in the Gulf of Maine. Cont Shelf Res 26:422–427

Reigber C, Bock R, Förste C, Grunwaldt L, Jakowski N, Lühr H, Schwintzer P, Tilgner C (1996) CHAMP Phase-B Executive Summary, STR96/13 . GFZ, Potsdam

Reigber C et al (2001) Global gravity field recovery with CHAMP. Presented at the 2001 IAG scientific assembly, Budapest, 2–7 Sept 2001

Riesz F, Szőkefalvi-Nagy B (1965) Leçons d'analyse fonctionelle. Gauthier-Villars, Paris

Rizos C (1979) An efficient computer technique for the evaluation of geopotential from spherical harmonic models. Aust J Geodesy Photogramm Cartogr 31:161–170

Rodriguez E, Morris CS, Belz JE, Chapin EC, Martin JM, Daffer W, Hensley S (2005) An assessment of the SRTM topographic products. Technical report JPL D-31639, Jet Propulsion Laboratory, Pasadena

Rowlands DD, Ray RD, Chinn DS, Lemoine FG (2002) Short-arc analysis of intersatellite tracking data in a gravity mapping mission. J Geod 76(6–7)307–316. doi:10.1007/s00190-002-0255-8

Rozanov I (1982) Markov random fields. Springer, Berlin

Rummel R (1993) Principle of satellite altimetry and elimination of radial orbit errors. In: Rummel R, Sanso F (eds) Satellite altimetry in geodesy and oceanography. Lecture notes in earth sciences, vol 50. Springer, Berlin/Heidelberg, pp 190–243

Rummel R, Schwarz K-P (1977) On the non-homogeneity of the global coavariance function. Bull Geod 51(2):93–103

Rummel R, Haagmans RHN (1990) Gravity gradients from satellite altimetry. Mar Geod 14:1–12

Rummel R, Rapp RH, Sünkel H, Tscherning CC (1988) Comparisons of global topographic isostaticmodels to the Earth's observed gravity field. Reports of the department of geodesy science and survey, vol 388. Ohio State University, Columbus

Rummel R, Van Gelderen M, Koop R, Schrama E, Sansò F, Brovelli M, Migliaccio F, Sacerdote F (1993) Spherical harmonic analysis of satellite gradiometry. Publications on geodesy, new series, vol 39. Netherlands Geodetic Commission, Delft

Sabadini R, Vermeersen B (2004) Global dynamics of the Earth. Kluwer, Doordrecht

Sacerdote F, Sansò F (1986) The scalar boundary value problem of physical geodesy. Man Geod 11(1):15–28

Sacerdote F, Sansò F (1991) Holes in boundary and out of boundary data. In: Determination of the geoid, present and future, IAG Symposia, vol 106. Springer, Berlin

Sacerdote F, Sansò F (2010) Least squares, Galerkin and BVPs applied to the determination of global gravity field models, IAG Symposia, vol 135. Springer, Berlin/Heidelberg

Saleh J, Pavlis NK (2003) The development and evaluation of the global digital terrain model DTM2002. In: Tziavos IN (ed) Gravity and geoid 2002, proceedings of the 3rd meeting of the international gravity and geoid commission, Ziti, Thessaloniki

Sanchez BV, Cunningham WJ, Pavlis NK (1997) The calculation of the dynamic sea surface topography and the associated flow field from altimetry data: a characteristic function method. J Phys Oceanogr 27(7):1371–1385

Sandwell DT (1992) Antarctic marine gravity field from high-density satellite altimetry. Geophys J Int 109:437–448

Sandwell DT, Smith WHF (1997) Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. J Geophys Res 102(B5):10039–10054

Sandwell DT, Smith WHF (2005) Retracking ERS-1 altimeter waveforms for optimal gravity field recovery. Geophys J Int 163:79–89. doi:10.1111/j.1365-246X.2005.02724

Sandwell DT, Smith WHF (2009) Global marine gravity from retracked Geosat and ERS-1 altimetry: ridge segmentation versus spreading rate. J Geophys Res 114:B01411. doi:10.1029/2008JB006008

Sansò F (1986) Statistical methods in physical geodesy. In: Matematical and numerical techniques in physcial geodesy. Lecture notes in Earth sciences, vol 7. Springer, Berlin

Sansò F (1995) The long road from measurement to boundary value probles in physical geodesy. Man Geod 20:326–344

Sansò F (1997) The hierarchy of geodetic BVP's. In: Geodetic boundary value problems in view of the one-centimeter geoid. Lecture notes in Earth science, vol 65. Springer, Berlin/Heidelberg

Sansò F, Sideris MG (1997) On the similarities and differences between systems theory and leasts-quares collocation in physical geodesy. Bolletino di Geodesia e Scienze Affini Anno LVI 2:173–206
Sansò F, Sona G (1995) Gravity reduction versus approximate BVP's. In: Proceedings of 3rd Hotine-Marussi symposium on mathematical geodesy. IAG Series, vol 114. Springer, Berlin
Sansò F, Vaníček P (2006) The orthometric height and the holonomity problem. J Geod 80:225–232
Sansò, Venuti G (1998) White noise, stochastic BVP's and Cimmino's theory. In: Proceedings of 4th Hotine-Marussi symposium on mathematical geodesy. IAG Series, vol 122. Springer, Berlin, pp 5–3
Sansò F, Venuti G (2002a) The estimation theory for random fields in the Bayesian context: a contribution from geodesy. In: Proceedings of 5th Hotine-Marussi symposium on mathematical geodesy. IAG Series, vol 127. Springer, Berlin/Heidelberg
Sansò F, Venuti G (2002b) The height datum/geodetic problem. Geophys J Int 149:768–775
Sansò F, Venuti G (2008) On the explicit determination of stability constants for linear geodetic boundary value problems. J Geod 82(12):909–916
Sansò F, Tscherning CC, Venuti G (2000) A theorem of insensitivity of the collocation solution to variations of the metric of the interpolation space. In: Geodesy beyond Year 2000. Proceedings of the 35th IAG general assembly. IAG Series, vol 121. Springer, Berlin/Heidelberg
Sansò F, Venuti G, Tziavos IN, Vergos GS, Grigoriadis GN (2008) Geoid and sea surface topography from satellite and ground data in the Mediterranean region. A review and new proposals. Bull Geod Geomat 3:156–201
Schafer RW, Rabiner LR (1973) A digital signal processing approach to interpolation. Proc IEEE 61(6):692–70A10
Schrama EJO (1989) The role of orbit errors in processing of satellite altimeter data. Rep No 33, Netherlands Geodetic Commission, Publications on Geodesy, New series, Delft
Schuh WD (1996) Tailored numerical solution strategies for the global determination of the Earth's gravity field. Technical report, Institute of Theoretical Geodesy, Technical University Graz, Graz
Schwarz K-P (1984) Data types and their spectral properties. In: Schwarz K-P (ed) Proceedings of the international summer school on local gravity field approximation, Beijing, pp 1–66
Schwarz KP, Sideris MG, Forsberg R (1987) Orthometric heights without levelling. J Surv Eng 113(1):28–40
Schwarz K-P, Sideris MG, Forsberg R (1990) The use of FFT tecniques in physical geodesy. Geophys J Int 100:485–514
Schwintzer P, Reigber C, Bode A, Kang Z, Zhu SY, Massmann F-H, Raimondo JC, Biancale R, Balmino G, Lemoine JM, Moynot B, Marty JC, Barlier B, Boudon Y (1997) Long wavelength global gravity field models: GRIM4 S4, GRIM4 C4. J Geod 71(4):189–208
She BB (1993) A PC-based unified geoid for Canada. M.Sc. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary
Shum CK, Woodworth PL, Andersen OB, Egbert G, Francis O, King C, Klosko S, Le Provost C, Li X, Molines JM, Parke M, Ray R, Schlax M, Stammer D, Tierney C, Vincent P, Wunch C (1997) Accuracy assessment of recent ocean tide models. J Geophys Res 102(C11):25173–25194
Sideris MG (1984) Computation of gravimetric terrain corrections using fast Fourier transform techniques. UCGE Report 20007, Department of Geomatics Engineering, The University of Calgary, Calgary
Sideris MG (1985) A fast Fourier transform method of computing terrain corrections. Man Geod 10:(1)66–73
Sideris MG (1987a) Spectral methods for the numerical solution of Molodensky's problem. UCSE Report 20024, Department of Geomatics Engineering, The University of Calgary, Calgary
Sideris MG (1987b) On the application of spectral techniques to the gravimetric problem. In: Proceedings of the XIX JUGG general assembly, Vancouver, 9–22 Aug 1987, pp 428–442
Sideris MG (1990) Rigorous gravimetric terrain modeling using Molodensky's operator. Man Geod 15:97–106

Sideris MG (1994) Regional geoid determination. In: Vaníček P, Christou NT (eds) Geoid and its geophysical Interpretations. CRC Press, Boca Raton, pp 77–94
Sideris MG (1995) Fourier geoid determination with irregular data. J Geod 70(1):2–12
Sideris MG (1996) On the use of heterogeneous noisy data in spectral gravity field modeling methods. J Geod 70(8):470–479
Sideris MG (2008) Geoid determination by FFT techniques. In: Lecture notes, international school for the determination and use of the geoid, Como, 15–19 Sept 2008
Sideris MG (2010) Geoid determination by FFT techniques. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg, 28 June –2 July 2010
Sideris MG, Forsberg R (1990) Review of geoid prediction methods in mountainous regions. In: Proceedings of the 1st international geoid commission symposium, Milan, 11–13 June 1990
Sideris MG, Li YC (1992) Improved geoid determination for levelling by GPS. In: Proceedings of the Sixth international geodetic symposium on satellite positioning, Columbus, 17–20 March 1992
Sideris MG, Li YC (1993) Gravity field convolutions without windowing and edge effects. Bull Geod 67:107–118
Sideris MG, Schwarz K-P (1986) Solving Molodensky's series by fast Fourier transform techniques. Bull Geod 60:51–63
Sideris MG, Schwarz K-P (1988) Recent advances in the numerical solution of the linear Molodensky problem. Bull Geod 62:59–69
Sideris MG, She BB (1994) A new, high-resolution geoid for Canada and part of the US by the ID-FFT method, Accepted for publication in Bulletin Geodesique
Sideris MG, Tziavos IN (1988) FFT-evaluation and applications of gravity-field convolution integrals with mean and point data, Bull Geod 62:521–540
Sjöberg LE (1986) Comparison of some methods of modifying Stokes' formula. In: Proceedings of the international symposium on the definition of the geoid, Florence, 26–30 May 1986
Sjoberg LE (2000) Topographic effects by the Stokes Helmert method of geoid and quasi-geoid determination. J Geod 74:255–268
Sjöberg LE (2005) A discussion of the approximations made in the practical implementation of the remove-compute-restore technique in regional geoid modeling. J Geodyn 78:645–653
Sjöberg LE (2007) The topographic bias in physical geodesy. J Geodyn 81:345–350
Sjöberg LE (2009) The terrain correction in gravimetric geoid computation-is it needed? Geophys J Int 176:14–18
Smith DA (2000) The gravitational attraction of any polygonally shaped vertical prism with inclined top and bottom faces. J Geod 74(5):414–420
Smith WHF, Sandwell DT (1997) Global seafloor topography from satellite altimetry and ship depth soundings. Science 277:1957–1962
Sneeuw N, Ilk KH (1997) The status of spaceborne gravity field mission concepts: a comparative simulation study. In: Segawa J, Fujimoto H, Okubo S (eds) Gravity, geoid and marine geodesy, IAG Symposia, vol 117. Springer, Berlin
Somigliana C (1929) Teoria generale del campo gravitazionale dell'ellissoide di rotazione. Mem Soc Astr Ital 4:541–599
Sona G (1995) Numerical problems in the computation of ellipsoidal harmonics. J Geod 70:117–126
Stokes GG (1849) On the variation of gravity on the surface of the Earth. Trans Camb Phil Soc 8:672–695
Strang van Hees G (1990) Stokes' formula using fast Fourier techniques. Man Geod 15:235–239
Strykowski G, Boschetti F, Papp G (2005) Estimation of the mass density contrasts and the 3D geometrical shape of the source bodies, in the Yilgarn area, Eastern Goldfields, Western Australia. J Geodyn 39:444–460
Sun W (2002) A formula for gravimetric terrain corrections using powers of topographic height. J Geodesy 76(8):399–406
Sünkel H (1981a) Cardinal interpolation. Reports of the department of geodesy science and survey, vol 312. Ohio State University, Columbus

Sünkel H (1981b) Point mass models and the anomalous gravitational field. Reports of the department of geodesy science and survey, vol 328. Ohio State University, Columbus
Sünkel H (1983) The generation of a mass point model from surface gravity data. Reports of the department of geodesy science and survey, vol 353. Ohio State University, Columbus
Sünkel H (1984) Splines: their equivalence to collocation. Reports of the department of geodesy science and survey, vol 357. Ohio State University, Columbus
Sünkel H (1986) Global topographic isostatic models. In: Sünkel H (ed) Mathematical and numerical techniques in physical geodesy. Lecture notes in Earth sciences, vol 7. Springer, Berlin/Heidelberg/New York
Sünkel H, Tscherning CC (1981) A method for the construction of spherical mass distributions consistent with the harmonic part of the Earth's gravity field. Man Geod 6:131–156
Szegö G (1948) Orthogonal polynomials. American Mathematical Society, Providence
Tapley BD, Watkins MM, Ries JC, Davis GW, Eanes RJ, Poole SR, Rim HJ, Schutz BE, Shum CK, Nerem RS, Lerch FJ, Marshall JA, Klosko SM, Pavlis NK, Williamson RG (1996) The joint gravity model-3. J Geophys Res 101(B12):28029–28049
Tapley BD, Shum CK, Ries JC, Poole SR, Abusali PAM, Bettadpur SV, Eanes RJ, Kim MC, Rim HJ, Schutz BE (1997) The TEG 3 geopotential model. In: Segawa J, Fujimoto H, Okubo S (eds) Gravity, geoid and marine geodesy, IAG Symposia, vol 117. Springer, Berlin/New York
Tapley BD, Bettadpur S, Watkins M, Reigber C (2004) The gravity recovery and climate experiment: mission overview and early results. Geophys Res Let 31:L09607. doi:10.1029/2004GL019920
Tapley BD, Ries J, Bettadpur S, Chambers D, Cheng M, Condi F, Gunter B, Kang Z, Nagel P, Pastor R, Pekker T, Poole S, Wang F (2005) GGM02 – An improved Earth gravity field model from GRACE. J Geod 79(8):467–478. doi:10.1007/s00190-005-0480-z
Taylor AE (1958) Introduction to functional analysis. Wiley, New York
Tenzer R, Vaníček P, Novák P (2003) Far-zone contributions to topographical effects in the Stokes-Helmert method of the geoid determination. Stud Geophys Geod 47:467–480
Tenzer R, Vaníček P, Santos M, Featherstone WE, Kuhn M (2005) The rigorous determination of orthometric heights. J Geod 79:82–92
Tenzer R, Novák P, Moore P, Kuhn M, Vaníček P (2006) Explicit formula for the geoid-quasigeoid separation. Stud Geod Geofys 50:607–618
Teunissen PJG, Amiri-Simkoowei AR (2008) Least-squares variance component estimation. J Geod 82(2):65–82
Tocho C, Vergos GS, Sideris MG (2007) Evaluation of the SRTM 90m DTM over Argentina and its implications to gravity field and geoid modelling. In: Forsberg R, Kiliǫğlu A (eds) 1st international symposium of the international gravity field service "gravity field of the Earth". General Command of Mapping 18(Special issue):324–329
Todhunter I (1873) A history of the mathematical theory of attraction and the figure of the Earth from the time of Newton to that of Laplace. MacMillan, London
Torge W (1989) Gravimetry. Walter de Gruyter, Berlin/New York
Torge W (2001) Geodesy, 3rd edn. de Gruyter, Berlin/New York
Trimmer RG, Manning DM (1996) The altimetry derived gravity anomalies to be used in computing the joint DMA/NASA Earth gravity model. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations, IAG Symposia, vol 116. Springer, Berlin/Heidelberg
Tscherning CC (1974) A FORTRAN IV program for the determination of the anomalous potential using stepwise least squares collocation. Reports of the department of geodesy science and survey, vol 212. Ohio State University, Columbus
Tscherning CC (1976) Covariance expressions for second and lower order derivatives of the anomalous potential. Reports of the department of geodesy science and survey, vol 225. Ohio State University, Columbus
Tscherning CC (1983) The role of high degree spherical harmonic expansions in solving geodetic problems. In: Proceedings of the international association of geodesy symposia, IUGG XVII

general assembly, vol 1. Reports of the department of geodesy science and survey. Ohio State University, Columbus, pp 431–441
Tscherning CC (1985) Local approximation of the gravity potential by least squares collocation. In: Schwarz K-P (ed) Proceedings of the international summer school on local gravity field approximation, Beijing, 21 Aug–4 Sept 1984. Publ 60003
Tscherning CC (1991) The use of optimal estimation for gross-error detection in databases of spatially correlated data. Bull Inf 68:79–89
Tscherning CC (1993) Computation of covariances of derivatives of the anomalous gravity potential in a rotated reference frame. Man Geod 18:(3)115–123
Tscherning CC (2008) Geoid determination by 3D least squares collocqation. In: Lecture notes, international school for the determination and use of the geoid, Como, 2008 Sept 15–19
Tscherning CC (2010) Geoid determination by 3D least squares collocqation. In: Lecture notes, international school for the determination and use of the geoid, Saint Petersburg, 28 June–2 July 2010
Tscherning CC, Forsberg R (1992) Harmonic continuation and gridding effects on geoid height prediction. Man Geod 66:41–53
Tscherning CC, Poder K (1981) Some geodetic applications of Clenshow summation. In: Proceedings of 5th Hotine-Marussi symposium on mathematical geodesy. IAG Series, vol 127. Springer, Berlin
Tscherning CC, Rapp RH (1974) Closed covariance expressions for gravity anomalies, geoid undulations, and deflections of the vertical implied by anomaly degree variances. Rep No 208, Department of Geodetic Sciences and Surveying, The Ohio State University, Columbus
Tscherning CC, Veicherts M (2007) Optimization of gradient prediction. GOCE-TN-HPF-GS-0214, 2007. http://cct.gfy.ku.dk/publ_cct/cct1912.pdf
Tscherning CC, Forsberg R, Knudsen P (1992) The GRAVSOFT package for geoid determination. In: Holota P, Vermeer M (eds) 1st continental workshop on the geoid in Europe, Prague, pp 327–334
Tscherning CC, Knudsen P, Ekholm S, Andersen OB (1993) An analysis of the gravity field in the Norwegian sea using ERS-1 altimeter measurements. In: Proceedings of the first ERS-1 symposium. European Space Agency Special Publication ESA SP-359, ESA Publications Division, European Space Agency, Noordwijk, The Netherlands, pp 413–418
Tscherning CC, Radwan A, Tealeb AA, Mahmoud SM, Abd El-Monum M, Hassan R, El-Syaed I, Saker K (2001) Local geoid determination combining gravity disturbances and GPS/levelling: a case study in the Lake Nasser area, Aswan, Egypt. J Geod 75(7–8):343–348
Tsoulis D (1999) Analytical and numerical methods in gravity field modelling of ideal and real masses. Dissertation, DGK, Reihe C, Nr. 510, Munchen
Tsoulis D, Tziavos IN (2003) A comparison of some existing methods for the computation of terrain corrections in local gravity field modelling. In: Tziavos N (ed) Gravity and geoid 2002. Ziti, Thessaloniki, pp 156–160
Tsoulis D, Grigoriadis VN, Tziavos IN (2007) The utilization of global digital crustal databases in regional applications of forward gravity field modeling. In: Forsberg R, Kiliçğlu A (eds) 1st international symposium of the international gravity field service "gravity field of the Earth". General Command of Mapping 18(Special issue):348–353
Turcotte DL, Schubert G (2001) Geodynamics. CCambridge University Press, New York
Tziavos IN (1992) Alternative numerical techniques for the efficient computation of terrain corrections and geoid undulations. In: 1st continental workshop on the geoid in Europe – "Towards a precise pan European reference geoid for the nineties", Prague
Tziavos IN (1993) Numerical considerations of FFT methods in gravity field modelling, Wissenschaftliche Arbeiten der Facbrichtung Vennessungswesen der Universitiit Hannover Nr. 188, Hannover
Tziavos IN (1996) Comparisons of spectral techniques for geoid computations over large regions. J Geod 70:357–373
Tziavos IN, Featherstone WE (2001) First results of using digital density data in gravimetric geoid computation in Australia. In: Sideris MG (ed) Proceedings of international asso-

ciation of geodesy symposia “gravity, geoid and geodynamics 2000”, vol 123. Springer, Berlin/Heidelberg, pp 335–340
Tziavos IN, Sideris MG, Forsberg R, Schwarz K-P (1988) The effect of the terrain on airborne gravity and gradiometry. J Geophys Res 93(B8):9173–9186
Tziavos IN, Sideris MG, Schwarz K-P (1992) A study of the contribution of various gravimetric data types on the estimation of gravity field parameters in the mountains. J Geophys Res 97(B6):8843–8852
Tziavos IN, Sideris MG, Sünkel H (1996) The effect of surface density variation on terrain modeling – A case study in Austria. In: Tziavos IN, Vermeer M (eds) Techniques for local geoid determination, Report No. 96(2). Finish Geodetic Institute, Masala, pp 99–110
Tziavos IN, Andritsanos VD (1998) Recent advances in terrain correction computations. In: Vermeer M, Adam J (eds) Proceedings of the 2nd continental workshop on the geoid in Europe, Budapest, 10–14 March 1998, pp 169–176
Tziavos IN, Vergos GS, Grigoriadis VN (2010) Investigation of topographic reductions and aliasing effects on gravity and the geoid over Greece based on various digital terrain models. Surv Geophys 31:23–67
USGS (2008) Shuttle Radar Topography Mission digital topographic data. United states geological survey. ftp://e0srp01u.ecs.nasa.gov/srtm. Accessed March 2008
Uotila UA (1986) Notes on adjustment computations, Part I. Reports of the department of geodesy science and survey. Ohio State University, Columbus
Vaníček P (1991) Vertical datum and NAVD88. Surv land inform syst 51(2):83–86
Vaníček P, Krakiwsky EJ (1986) Geodesy the concepts. North-Holland, Amsterdam
Vaníček P, Sjöberg LE (1989) Kernel modification in generalized Stokes’ technique for geoid determination. In Proceedings of the lAG general meeting, Edinburgh, 3–1 Aug 1989
Vaníček P, Najafi M, Martinec Z, Harrie L, Sjöberg LE (1995) Higher-degree reference field in the generalized Stokes-Helmert scheme for geoid computation. J Geod 70:176–182
Vaníček P, Huang J, Novák P, Pagaiatakis S, Veronneau M,, Martinec Z, Featherstone WE (1999) Determination of the boundary values for the Stokes-Helmert problem. J Geod 73:180–192
Vaníček P, Novák, Martinec Z (2001) Geoid, topography, and the Bouguer plate or shell. J Geod 75:210–215
Vaníček P, Tenzer R,, Sjöberg LE, Martinec Z, Featherstone WE (2004) New views of the spherical Bouguer gravity anomaly. Geophys J Int 159:460–472
Vapnik V (1982) Estimation of dependencies base on empirical data. Springer, New York
Vergos GS, Tziavos IN, Andritsanos VD (2005a) On the determination of marine geoid models by least squares collocation and spectral methods using heterogeneous data. In: Sansò F (ed) Proceedings of international association of geodesy symposia “a window on the future of geodesy”, vol 128. Springer, Berlin/Heidelberg, pp 332–337
Vergos GS, Tziavos IN, Andritsanos VD (2005b) Gravity database generation and geoid model estimation using heterogeneous data. In: Jekeli C, Bastos L, Fernandes J (eds) Proceedings of international association of geodesy symposia “gravity geoid and space missions”, vol 129. Springer, Berling/Heidelberg, pp 155–160
Vermeer M (1992) A Frequency domain approach to optimal geophysical data gridding. Man Geod (17):141–154
Vermeer M (2008) Comment on Sjöberg (2006) “The topographic bias by analytical continuation in physical geodesy”. J Geod 81(5):345–350
Vermeer M, Forsberg R (1992) Filtered terrain effects: a frequency domain approach to terrain effect evaluation. Man Geod (17):215–226
Wang YM (1987) Numerical aspects of the solution of Molodensky’s problem by analytical continuation. Man Geod 12:290–295
Wang YM (1988) Downward continuation of the free air gravity anomalies to the ellipsoid using the gradient solution, Poisson’s integral and terrain correction – numerical comparison and the computations. Reports of the department of geodesy science and survey, vol 393. Ohio State University, Columbus

Wang YM (2001) GSFC00 mean sea surface, gravity anomaly, and vertical gravity gradient from satellite altimeter data. J Geoph Res 106(C12):31167–31174
Wang YM, Rapp RH (1990) Terrain effects on geoid undulation computations. Man Geod 15:23–29
Wenzel HG (1985) Hochauflösende Kugelfunktionsmodelle für das Gravitationspotential der Erde, Wiss. Arb. 137, Fachrichtung Vermess. der Univ. Hannover, Hannover
Wenzel HG (1998) Ultra high degree geopotential models GPM98A, B and C to degree 1800. http://www.gik.uni-karlsruhe.de/~wenzel/gpm98abc/gpm98abc.htm
Wenzel G (1999) Schwerefeldmodellierung durch ultra-hochaufll"osende Kugelfunktionsmodelle. Z Vermess 124(5):144–154
Werner J (1974) Potential theory. Lecture notes in mathematics, vol 408. Springer, Berlin
Werner M (2001) Shuttle radar topography mission (SRTM), mission overview. J Telecom 55:75–79
Wesolowsky GO (1976) Multiple regression and analysis of variance. Wiley, New York
Wichiencharoen C (1982) The indirect effects on the computation of geoids undulations. Reports of the department of geodesy science and survey, vol 336. Ohio State University, Columbus
Wild-Pfeiffer F (2008) A comparison of different mass elements for use in gravity gradiometry. J Geod 83: 637–653
Wolff M (1969) Direct measurements of the Earth's gravitational potential using a satellite pair. J Geoph Res 74(22):5295–5300
Wunch C (1993) Physics of the ocean circulation. In: Geodesy and oceanography. Lecture notes in earth sciences, vol 50. Springer, Berlin
Wunsch C, Zlotnicki V (1984) The accuracy of altimétric surfaces. Geophys J R Astron Soc 78:795–808. doi:10.1111/j.1365-246X.1984.tb05071.x
Yale MM, Sandwell DT, Smith WHF (1995) Comparison of along-track resolution of stacked Geosat, ERS 1 and TOPEX satellite altimeters. J Geophys Res 100(8):15117–15127
Yamaguchi Y, Kahle AB, Tsu H, Kawakami T, Pniel M (1998) Overview of advanced spaceborne thermal emission and reflection radiometer (ASTER). IEEE Tans Geosci Remote Sens 36(4):1062–1071
Yi Y (1995) Determination of gridded mean sea surface from TOPEX, ERS-1 and GEOSAT altimeter data. Rep No 434, Department of Geodetic Sciences and Surveying, The Ohio State University, Columbus
Yosida K (1978) Functional analysis. Springer, Berlin/Heidelberg
Zhang B, Sideris MG (1996) Oceanic gravity by analytical inversion of Hotine's formula. Mar Geod 19:115–136
Zhu L, Jekeli C (2009) Gravity gradient modeling using gravity and DEM. J Geod 83:557–567
Zlotnicki V (1984) On the accuracy of gravimetric geoids and the recovery of oceanographic signals from altimetry. Mar Geod 8:129–157
Zwally HJ, Schutz B, Abdalati W, Abshire J, Bentley C, Brenner A, Bufton J, Dezio J, Hancock D, Harding D, Herring T, Minster B, Quinn K, Palm S, Spinhirne J, Thomas R (2002) ICESat's laser measurement of polar ice, atmosphere, ocean and land. J Geodyn 34(3–4):405–445

Index

F. Sansò and M.G. Sideris (eds.), *Geoid Determination*, Lecture Notes in Earth System Sciences 110, DOI 10.1007/978-3-540-74700-0,