Chapter 5
Problem Posing in Modelling from Data Series

Creation of models on the basis of observations and investigation of their properties is, in essence, the main contents of science.

"System Identification. Theory for the User" (Ljung, 1991)

5.1 Scheme of Model Construction Procedure

Despite infinitely many situations, objects and purposes of modelling from observed data, one can single out common stages in a modelling procedure and represent them with a scheme (Fig. 5.1). The procedure is started with consideration of available information about an object (including previously obtained experimental data from the object or similar ones, a theory developed for the class of objects under investigation and intuitive ideas) from the viewpoint of modelling purposes, with acquisition and preliminary analysis of data series. It ends with the use of an obtained model for solution to a concrete problem. The procedure is usually iterative, i.e. it is accompanied by multiple returns to a starting or an intermediate point of the scheme and represents a step-by-step approach to a "good" model.

Model structure[1] is formed at the key stage 2. This stage is often called structural identification. Firstly, one selects a type of equations. Below, we speak mainly of finite-dimensional deterministic models in the form of discrete maps

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n, \mathbf{c}) \tag{5.1}$$

or ordinary differential equations

$$\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{f}(\mathbf{x}, \mathbf{c}), \tag{5.2}$$

where $\mathbf{x}$ is the D-dimensional state vector, $\mathbf{f}$ is the vector-valued function, $\mathbf{c}$ is the P-dimensional parameter vector, n is the discrete time and t is the continuous time. Secondly, a kind of functions (scalar components of $\mathbf{f}$) is specified. Thirdly, one establishes a relationship between dynamical variables (components of $\mathbf{x}$) and

[1] A model structure is a parameterised set of models (Ljung, 1991).

B.P. Bezruchko, D.A. Smirnov, *Extracting Knowledge From Time Series*, Springer Series in Synergetics, DOI 10.1007/978-3-642-12601-7_5,

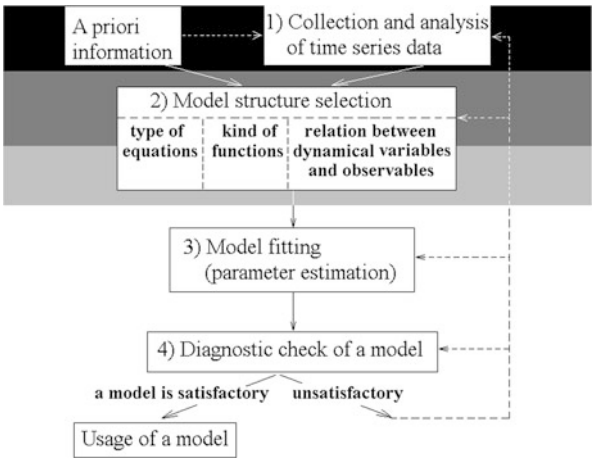

Fig. 5.1 Typical scheme of an empirical modelling procedure

observed quantities $\boldsymbol{\eta}$. Dynamical variables may coincide with the components of $\boldsymbol{\eta}$. In a more general case, the relationship is specified in the form $\boldsymbol{\eta} = \mathbf{h}(\mathbf{x})$, where $\mathbf{h}$ is called a *measurement function*. Moreover, one often introduces a random term $\boldsymbol{\zeta}$ ($\boldsymbol{\eta} = \mathbf{h}(\mathbf{x}) + \boldsymbol{\zeta}$) to allow for a *measurement noise*. To make a model more realistic, one often incorporates random terms called *dynamical noise* into the Eq. (5.1) or (5.2).

Specification of a model structure is the most complicated and creative stage of the modelling procedure. After that, it remains just to determine concrete values of the parameters $\mathbf{c}$ (the stage 3). Here, one often speaks of *parameter estimation* or *model fitting*. To estimate parameters, one usually searches for an extremum of a *cost function*, e.g. minimum of the sum of squared deviations of a model realisation from the observed data. If necessary, one performs preliminary processing of the observed data series, such as filtering, numerical differentiation or integration. This is mainly a technical stage of numerical calculations. However, it requires the choice of an appropriate principle for parameter estimation and a relevant technique for its realisation.

One must also make a decision at the stage of model testing (the stage 4). Typically, model "quality" is checked with the use of a test part of an observed series specially reserved for this purpose. Depending on modelling purposes, one distinguishes between two types of problems: "cognitive identification" (the purpose is to get an adequate[2] model) and "practical identification" (there is a practical task to be solved with a model, e.g. a forecast). Accordingly, one either performs *validation (verification) of a model* in respect of the object properties of interest or checks *model efficiency* for achievement of a practical goal. If a model is found satisfactory (adequate or efficient), it is used further. Otherwise, it is returned to previous stages of the scheme (Fig. 5.1) for revision.

[2] *Adequacy* is understood as a correspondence between properties of a model and an object, i.e. a correct reproduction of original properties by a model (Philosophic dictionary, 1983, p. 13).

In particular, a researcher can return even to the stage of data collection (the stage 1) and ask for new data. It is appropriate to note that the data may not only be collected and analysed at this stage but also be pre-processed in different ways, e.g., to reduce measurement errors and fill possible gaps. This is especially important in geophysics, where the large field of *data assimilation* has appeared and evolved into a mature discipline (Anderson and Willebrand, 1989; Brassieur and Nihoul, 1994; Ghil and Malanotte-Rizzoli, 1991; Ide et al., 1997; Malanotte-Rizzoli, 1996). Data assimilation combines different estimation and modelling methods, where the Kalman filters occupy one of the central places (Bouttieur and Courtier, 1999; Evensen, 2007; Houtekamer and Mitchell, 1998; Robinson and Lermusiaux, 2000a, b).

5.2 Systematisation in Respect of A Priori Information

Background under the scheme in Fig. 5.1 changes from black (darkness of ignorance) to white reflecting a degree of prior uncertainty in modelling. The least favourable is a situation called "black box" when information about possibly adequate model structure is lacking so that one must start a modelling procedure from the very top of the above scheme. The more is known about how the model should look like (i.e. the lower is a "starting position" at the scheme), the more probable is a success. A "box" becomes "grey" and even "transparent". The latter means that a model structure is completely known a priori.

Methods for preliminary analysis of a time series, which can give useful information about a model structure and simplify a modelling problem, are discussed in Chap. 6. One can never avoid problems of the lower levels of the scheme, which are inevitably faced by a researcher overcoming the structural identification stage. Therefore, Chaps. 7 and 8 deal with the simplest situation, where everything is known about a model except for the concrete values of its parameters. It corresponds to the white background in Fig. 5.1.

Depending on the origin of a time series, two qualitatively different situations emerge in respect of the formulation of the modelling problem. The first one is when observations are a realisation of a certain mathematical model (a set of equations) obtained with a numerical technique. It is this situation where the term "reconstruction of equations" is completely appropriate. In such a case, model validation is much simpler since one can compare modelling results with the "true" original equations and their solutions. Besides, one may formulate theoretical conditions for the efficiency of modelling techniques for different classes of systems. The second situation is when a time series results from measurements of a certain real-world process so that a unique true model does not exist (Chap. 1) and one cannot assure success of modelling. One can only wonder at "inconceivable efficiency of mathematics" if a "good" model is achieved.[3]

[3] In addition, one can consider laboratory experiments as an intermediate situation. Strictly speaking, they represent real-world processes (not numerical simulations), but they can be manipulated

Further, we consider various techniques for model construction. To explain and illustrate them, we use mainly the former of the two situations, i.e. we reconstruct equations from their numerical solutions with various noises introduced to make a situation more realistic. We discuss different modelling problems in Chaps. 8, 9 and 10 according to the following "hierarchy":

(i) *Parameter estimation*. Structure of model equations is specified completely from physical or other conceptual considerations. Only parameter values are unknown. This problem setting is the simplest one and can be called "transparent box". However, essential difficulties can still arise due to a big number of unknown parameters and unobserved (hidden) dynamical variables.
(ii) *Reconstruction of equivalent characteristics*. A model structure is known to a significant degree from physical considerations. Therefore, one does not have to search for a multivariate function **f**. One needs to find only some of its components, which are univariate or bivariate functions (possibly, non-linear). We call them "equivalent characteristics". The term is borrowed from radiophysics but often appropriate in physics, biology and other fields.
(iii) *Black box reconstruction*. Since a priori information is lacking, one looks for a function **f** in a universal form. Solution to this problem is most often called "reconstruction of equations of motion". This is the most complicated situation.

Transition from the setting (i) to the setting (iii) is gradual. Many situations can be ordered according to the degree of prior uncertainty (darkness of the grey tone or non-transparency). However, it is not a linear ordering since not all situations can be readily compared to each other and recognised as "lighter" or "darker". For instance, it is not easy to compare information about the presence of a sinusoidal driving and a certain symmetry of orbits in a phase space.

5.3 Specific Features of Empirical Modelling Problems

5.3.1 Direct and Inverse Problems

It is often said that getting model equations from an observed time realisation belongs to the class of *inverse problems*. The term "inverse problem" is used in many mathematical disciplines. Inverse problems are those where input data and sought quantities switch their places as compared with some habitual basic problems which are traditionally called "direct".

---

in different ways and even constructed so as to correspond well to some mathematical equations. Thus, one may know a lot about appropriate mathematical description of such processes. It extends one's opportunities of successful modelling and acute model validation.

As a rule, direct problems are those whose formulations arise first (logically) after creation of a certain mathematical apparatus. Usually, one has regular techniques to solve direct problems, i.e. they are often relatively simple. An example of a direct problem is the Cauchy problem (a direct problem of dynamics): to find a particular solution to an ordinary differential equation, given initial conditions. Then, getting a set of ODEs whose particular solution is a given function is an inverse problem of dynamics.

It is appropriate to speak of inverse problems in experimental data analysis when one must determine parameters of a mathematical model of an investigated process from measured values of some variables. For instance, in spectroscopy and molecular modelling one considers the following problem: to determine a geometrical configuration of molecules of a certain substance and compute the corresponding parameters from an observed absorption spectrum. This is an inverse problem, while a direct one is to compute an absorption spectrum, given a model of a molecule.

5.3.2 Well-posed and Ill-posed Problems

Inverse problems are often ill-posed in the strict sense described below.

Let us express a problem formulation as follows: to find an unknown quantity Y (a solution) given some input data X. A problem is called *stable with respect to input data* if its solution depends on input data continuously $Y = \Phi(X)$, i.e. a solution changes weakly under a weak variation in X. A problem is called *well-posed according to Hadamard* if the three conditions are fulfilled: (i) a solution exists, (ii) a solution is unique, (iii) the problem is stable with respect to input data.

The first two conditions do not require comments. The third one is important from a practical point of view, since data are always measured with some errors. If a weak variation in the input data leads to a drastic change in a solution to a problem, then one cannot assure reliability of the solution so that usefulness of such a solution is doubtful. The third condition requires an answer which changes weakly under a weak variation in input data.

If at least one of the three conditions is violated, a problem is called *ill-posed according to Hadamard*. Of course, one should always tend to formulate well-posed problems. Therefore, ill-posed problems were out of mathematicians' interests for a long time (Tikhonov and Arsenin, 1974). It was thought that they did not make "physical sense", had to be reformulated, etc. However, as time went by, such problems more and more often emerged in different fields of practice. Therefore, special approaches to their solutions began to be developed such as regularisation techniques and construction of quasi-solutions.

Below, we often discuss well-posedness of modelling problems. Ill-posedness of a problem is not a "final verdict". It does not mean that "everything is bad". Thus, even the problem of differentiation is ill-posed[4] but differentiation is widely used.

[4] Significant difficulties in numerical estimation of derivatives from a time series (Sect. 7.4.2) are related to ill-posedness of the differentiation problem (Tikhonov and Arsenin, 1974). A concrete

Though it would be preferable to deal with well-posed problems, even a model obtained under ill-posed formulation can appear quite appropriate for practical purposes. It is possible, e.g., if ill-posedness are based on the non-uniqueness of a solution, but one manages to select a certain solution from a multitude of them, which gives satisfactory results. Thus, in molecular modelling the authors of Gribov et al. (1997) stress principal ill-posedness of the arising problems (namely, getting model parameters from an absorption spectrum or a diffraction pattern) and the necessity to consider models obtained under various experimental settings as mutually complementary and only partially reflecting object properties.

There is another widespread approach to solve an ill-posed problem. Let a solution Y^* exist for input data X^*. Let the data X^* be known with a certain error and denote such a "noise-corrupted" input as X. Strictly speaking, the problem may not have a solution for the input data X. Then, one seeks for a quantity $Z = \Phi(X)$, which is close to a solution in some sense, while the mapping Φ is continuous and $\Phi(X) \to Y^*$ for $\|X - X^*\| \to 0$. Such a quantity Z is called a *quasi-solution*.

In practice, ill-posedness of a problem of modelling from a time series is often eliminated due to ad hoc additional assumptions or a priori information about a model structure, which helps to choose appropriate basis functions (Chaps. 7, 8, 9 and 10). Also, there are more general procedures providing well-posedness of a problem. They consist of construction of the so-called regularising functional (Tikhonov and Arsenin, 1974; Vapnik, 1979, 1995).

A typical example of the latter situation is the problem of approximation of a dependence $Y(X)$ based on a finite "learning" sample (Sect. 7.2.1). It is quite easy to construct a curve passing via each experimental point on the plane (X, Y). However, there is an infinite multitude of such curves differing from a constructed one by arbitrary oscillations *between the points*. Each of the curves provides a minimal (zero) value of an empirical mean-squared approximation error and represents in this sense an equitable solution to the problem. The number of solutions is reduced if one imposes constraints on the acceptable value of inter-point oscillations. It is done with a regularising functional (Sect. 7.2.3). Thus, a check for ill-posedness of a problem is important, since it helps to select the most efficient way for its solution or even change the entire problem setting.

---

example is following. There is a continuously differentiable function $x_0(t)$, whose derivative is denoted as $\mathrm{d}x_0(t)/\mathrm{d}t = y_0(t)$. Let $x_0(t)$ be known with a certain error, i.e. the input data is a function $x(t) = x_0(t) + \xi(t)$, where $\xi(t)$ is a continuously differentiable function with $|\xi(t)| \le \delta$. The input data $x(t)$ are very close to $x_0(t)$ in the sense of metrics L_∞. A derivative of the "noise-corrupted" function $x(t)$ is $\mathrm{d}x(t)/\mathrm{d}t = y_0(t) + \mathrm{d}\xi(t)/\mathrm{d}t$. Its deviation ε from $y_0(t)$ can be arbitrarily large in the same metrics. Thus, $\mathrm{d}x(t)/\mathrm{d}t = y_0(t) + \omega\delta\cos(\omega t)$ for $\xi(t) = \delta\sin(\omega t)$ so that the error $\varepsilon = \omega\delta$ can be arbitrarily large for arbitrarily small δ if ω is sufficiently large. In other words, the differentiation error can be arbitrarily large for arbitrarily low amplitude of "quickly oscillating noise $\xi(t)$". However, it is important how closeness in the space of input data and in the space of solutions is understood. If one regards close such input data that their difference $\xi(t)$ satisfies simultaneously two conditions – $|\xi(t)| \le \delta$ and $|\mathrm{d}\xi(t)/\mathrm{d}t| \le \delta$ (i.e. $\xi(t)$ is a slowly varying function) – then the differentiation problem gets well-posed.

5.3.3 Ill-conditioned Problems

For practical applications, it is important to mention a kind of problem which is well-posed from a theoretical viewpoint but whose solution is "quite sensitive" to weak variations in input data. "Quite sensitive" is not a rigorous concept. It is determined by a specific problem, but in general it means that for a small relative error δ in input data, a relative error ε in a solution is much greater: $\varepsilon = K \cdot \delta$, where $K >> 1$. Though a problem is theoretically stable with respect to input data (i.e. an error in a solution is infinitesimal for an infinitesimal error in input data), a solution error may appear very large for a *finite* small error in input data. Such problems are called *weakly* stable with respect to input data or *ill-conditioned* (Kalitkin, 1978; Press et al., 1988).

From a practical viewpoint, they do not differ from the ill-posed problems, given that all numerical calculations and representations of numbers are of finite precision. An example of an ill-conditioned problem is a set of linear algebraic equations, whose matrix is close to a degenerate one. Such a matrix is also often called ill-conditioned (Golub and Van Loan, 1989; Kalitkin, 1978; Press et al., 1988; Samarsky, 1982). Ill-conditioned problems often arise in construction of a "cumbersome" mathematical model from a time series. To solve them, one needs the same ideas and techniques as for the ill-posed problems.

References

Anderson, D., Willebrand, J. (eds.): Oceanic Circulation Models: Combining Data and Dynamics. Kluwer, Dordrecht (1989)

Bouttier, F., Courtier, P.: Data Assimilation Concepts and Methods. Lecture Notes, ECMWF. http://www.ecmwf.int/newsevents/training/lecture_notes (1999)

Brasseur P., Nihoul, J.C.J. (eds.): Data dissimilation: tools for modelling the ocean in a global change perspective. NATO ASI Series. Series I: Global Environ. Change **19** (1994)

Evensen, G.: Data Assimilation. The Ensemble Kalman Filter. Springer, Berlin (2007)

Ghil, M., Malanotte-Rizzoli, P.: Data assimilation in meteorology and oceanography. Adv. Geophys. **33**, 141–266 (1991)

Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. Johns Hopkins University Press, Baltimore (1989)

Gribov, L.A., Baranov, V.I., Zelentsov, D.Yu.: Electronic-Vibrational Spectra of Polyatomic Molecules. Nauka, Moscow (in Russian) (1997)

Houtekamer, P.L., Mitchell, H.L.: Data assimilation using an ensemble Kalman filter technique. Mon. Wea. Rev. **126**, 796–811 (1998)

Ide, K., Courtier, P., Ghil, M., Lorenc, A.: Unified notation for data assimilation: Operational, sequential and variational. J. Meteor. Soc. Jpn. **75**, 181–189 (1997)

Kalitkin, N.N.: Numerical Methods. Nauka, Moscow, (in Russian) (1978)

Ljung, L.: System Identification. Theory for the User. Prentice-Hall, Engle Wood Cliffs, NJ (1991)

Malanotte-Rizzoli, P. (ed.): Modern Approaches to Data Assimilation in Ocean Modeling. Elsevier Science, Amsterdam (1996)

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C. Cambridge University Press, Cambridge (1988)

Robinson, A.R., Lermusiaux, P.F.J.: Overview of data assimilation. Harvard Rep. Phys. Interdisciplinary Ocean Sci. **62**, (2000a)

Robinson, A.R., Lermusiaux, P.F.J.: Interdisciplinary data assimilation. Harvard Reports in Physical. Interdisciplinary Ocean Sci. **63**, (2000b)
Samarsky, A.A.: Introduction to Numerical Methods. Nauka, Moscow, (in Russian) (1982)
Tikhonov, A.N., Arsenin, V.Ya. Methods for Solving Ill-Posed Problems. Nauka, Moscow (1974). Translated into English: Wiley, New York (1977)
Vapnik, V.N.: Estimation of Dependencies Based on Empirical Data. Nauka, Moscow (1979). Translated into English: Springer, New York, (1982)
Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)