J Geod (2007) 81:759–780
DOI 10.1007/s00190-007-0140-6

ORIGINAL ARTICLE

# Best prediction in linear models with mixed integer/real unknowns: theory and application

**P. J. G. Teunissen**

Received: 19 September 2006 / Accepted: 2 February 2007 / Published online: 28 February 2007


**Abstract** In this contribution, we extend the existing theory of minimum mean squared error prediction (best prediction). This extention is motivated by the desire to be able to deal with models in which the parameter vectors have real-valued and/or integer-valued entries. New classes of predictors are introduced, based on the principle of equivariance. Equivariant prediction is developed for the real-parameter case, the integer-parameter case, and for the mixed integer/real case. The best predictors within these classes are identified, and they are shown to have a better performance than best linear (unbiased) prediction. This holds true for the mean squared error performance, as well as for the error variance performance. We show that, in the context of linear model prediction, best predictors and best estimators come in pairs. We take advantage of this property by also identifying the corresponding best estimators. All of the best equivariant estimators are shown to have a better precision than the best linear unbiased estimator. Although no restrictions are placed on the probability distributions of the random vectors, the Gaussian case is derived separately. The best predictors are also compared with least-squares predictors, in particular with the integer-based least-squares predictor introduced in Teunissen (J Geodesy, in press, 2006).

**Keywords** Minimum mean squared prediction error · Best prediction and estimation · Linear unbiased prediction · Least-squares prediction · Equivariant prediction · Integer equivariant prediction

P. J. G. Teunissen (✉)
Delft Institute for Earth Observation and Space Systems (DEOS), Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
e-mail: P.J.G.Teunissen@tudelft.nl

## 1 Introduction

The prediction of spatially and/or temporally varying variates based on observations of these variates (or functionals thereof) at some locations in space and/or instances in time, is an important topic in various spatial and Earth science disciplines. This topic has been extensively studied, albeit under different names. In physical geodesy, where it is used to predict spatially varying variates, it is known as least-squares collocation (LSC). Fundamental contributions to this field have been made by Krarup (1969) and Moritz (1973), also see Rummel (1976), Dermanis (1980), Sanso (1980, 1986), Grafarend and Rapp (1980), Moritz and Suenkel (1978), Tscherning (1978). The underlying model of LSC is the so-called trend-signal-noise model. This model is quite general and it encompasses many of the conceivable geodetic measurements (Moritz 1980). It also forms the basis of the concept of integrated geodesy as introduced in Eeg and Krarup (1973), also see Krarup (1980) and Hein (1986).

Prediction of spatially varying variates was also developed in meteorology, where it was originally referred to as objective analysis (Gandin 1963). Furthermore, least-squares prediction finds its analogue in Baarda's (1968) $x^R$-variates, which show how correlated, but free or constituent, variates are adjusted.

The trend-signal-noise model also forms the basis of prediction in geostatistics, where optimal linear prediction is called Kriging, named after Krige (1951) and further developed by Matheron (1970), also see, e.g., Blais (1982), Journel and Huijbregts (1991), Reguzzoni et al. (2005). When the trend is unknown it is referred to as universal Kriging and when the trend is absent or set to zero, it is called simple Kriging. Although collocation

and Kriging have been developed for spatially varying variates, they are closely connected with the fundamental work of Kolmogorov (1941) and Wiener (1948) on the interpolation, extrapolation and smoothing of stationary time-series. In the absence of a trend, collocation and simple Kriging become the spatial analogue of Kolmogorov–Wiener prediction (Grafarend 1976; Moritz 1980).

All of the above methods of prediction can be cast in the framework of either least-squares prediction or of best linear (unbiased) prediction. In a statistical context, we speak of prediction if a function of an observable random vector $y$ is used to guess the outcome of another random, but unobservable, vector $y_0$. We speak of 'best' prediction if the predictor minimizes the mean squared prediction error.

In the present contribution, the minimization of the mean squared prediction error will be the leading principle. Since the current theory of best prediction is restricted to models in which the parameter vectors are real-valued, no best predictors yet exist that can take advantage of the possible integer nature of the parameters. This is a serious shortcoming of the present theory and implies that it is not optimally applicable to such models as used in, e.g., Global Navigation Satellite Systems (GNSS) or Interferometric Synthetic Aperture Radar (InSAR).

The goal of the present contribution is therefore to extend the current theory of best prediction, so as to be able to accomodate models in which the parameter vector is of the mixed type, i.e. having integer-valued as well as real-valued entries. As a result, we will introduce new predictors that can be shown to outperform some of the best predictors of the current theory. We will also show the link with integer-based least-squares prediction, the theory of which has been developed in Teunissen (2006). The principle of integer-based least-squares prediction is intrinsically different from that of best prediction, the difference of which becomes particularly apparant if one has to deal with integer parameters.

This contribution is organized as follows. In Sect. 2, we first present some general properties of best predictors. They are very useful for studying the properties of specific best predictors treated in the following sections. Also, a brief review of best linear prediction is given, which will serve as reference for some of the new predictors that will be introduced. In Sect. 3, we introduce the linear model of prediction. It forms the basis of our extention of the current theory of best prediction. This model is quite versatile and can be shown to cover various prediction problems. In this context, it is also shown that prediction is a more general concept than estimation. We will take advantage of this in the sections following by simultaneously identifying the best estimators as well. Since one can minimize the mean squared error within different classes of predictors, there are different predictors that one can call 'best'. All the best predictors treated in the present contribution will be related to one another according to their mean squared error and error variance performance.

In Sect. 3 we consider the class of linear unbiased predictors, of which the weighted least-squares predictor is an example. Linear unbiased prediction forms the stepping stone to the new concept of equivariant prediction, which is introduced in Sect. 4. Since the class of equivariant predictors encompasses the class of linear unbiased predictors, best equivariant prediction outperforms best linear unbiased prediction. The best equivariant predictor is derived and its properties are given.

In Sect. 5, we introduce the concept of integer equivariant prediction. Predictors of this class make an explicit use of the 'integerness' of the parameters. The best integer equivariant predictor is derived and its properties are given. This predictor outperforms the previously treated predictors. The same holds true for the corresponding estimators. Thus, the best integer equivariant estimator can be shown to have a better precision than the well-known best linear unbiased estimator (BLUE).

In Sect. 6, we use the results of Sects. 4 and 5 as building blocks for studying the mixed integer/real parameter case. Although we make no restriction on the probability distribution when deriving the best predictors, the best mixed equivariant predictor is also derived for the Gaussian case. This predictor is also compared to the integer-based weighted least-squares predictor. Finally, it is shown that the best linear unbiased predictor and the integer-based weighted least-squares predictor can be seen as two different limiting cases of the best mixed equivariant predictor.

Various examples are given to illustrate the theory. In order to avoid possible discontinuities in the lines of thought, most of the (longer) proofs are placed in the Appendix. We make use of the following notation: matrices and (multivariate) functions will be denoted by capitals, with the capital $Q$ being reserved for variance-covariance matrices. The matrix inequality $A \leq B$ means that matrix $B - A$ is a positive semi-definite matrix. The $n$-dimensional space of real numbers is denoted as $R^n$ and the $n$-dimensional space of integers is denoted as $Z^n$. $E(.)$ denotes the mathematical expectation operator and the probability density function (PDF) of a random vector $y$ will be denoted as $f_y(.)$. $||.||$ denotes the standard Euclidean norm and $||.||_W$ denotes the weighted norm, in which $W$ is a positive semi-definite weight matrix ($W \geq 0$). Thus $||.||^2_W = (.)^{\mathrm{T}} W(.)$.

We will often need to evaluate $E\left(||.||_W^2\right)$. If there is no reason for confusion, we will write this mean squared value simply as $E||.||_W^2$. Also the conditional mean will often be used. As is usual, we will write the mean of a random vector $y_0$, conditioned on another random vector $y$, as $E(y_0|y)$. The conditional mean $E(y_0|y)$ is again a random vector when considered as function of $y$. Sometimes, however, we will have the need to consider it just as a function, using a different symbol, say $v$, as argument. To make clear that this particular function is based on the conditional PDF $f_{y_0|y}(.,.)$, we will then write $E_{y_0|y}(y_0|v)$ instead of $E(y_0|v)$.

## 2 Minimum mean squared error prediction

### 2.1 Classes of best predictors

In this subsection, we present some general lemmas for best predictors. They will be useful for studying the properties of best predictors from different classes. We speak of prediction if a function of an observable random vector $y \in R^m$ is used to guess the outcome of another random, but unobservable, vector $y_0 \in R^{m_0}$. If the function is given as $G$, then $G(y)$ is said to be the predictor of $y_0$ (we call it a prediction of $y_0$ if the function is taken of an outcome of $y$).

If $G(y)$ is a predictor of $y_0$, then $e_0 = y_0 - G(y)$ is its prediction error. We will use the mean squared error (MSE) $E(||e_0||^2)$ to judge the performance of a predictor. Note, since both $y_0$ and $y$ are random, that the mean is taken with respect to their joint PDF. Thus, $E(||e_0||^2) = \int\int ||y_0 - G(y)||^2 f_{y_0 y}(y_0, y)\mathrm{d}y_0\mathrm{d}y$.

A predictor is called 'best' if it succeeds in minimizing the MSE. Since one can minimize the MSE over different classes of functions, there are different predictors that one can call 'best'.

**Definition 1** (Best predictor of certain class) $\hat{G}(y)$ is said to be the best predictor of $y_0$ within class $\Omega$ if

$$E||y_0 - \hat{G}(y)||^2 = \min_{G\in\Omega} E||y_0 - G(y)||^2 \tag{1}$$

In this contribution, different classes of functions are considered, some of which are subsets of others. By knowing the relation among the different classes of functions, one can often already infer which of the minimum MSEs will be largest or smallest. It will be clear that the minimum MSE (MMSE) can not get smaller if one puts more restrictions on the class of functions $\Omega$ over which the minimization takes place. We therefore have the following lemma.

**Lemma 1** (MMSE versus predictor class) *Let $\Omega_1, \Omega_2$ be two classes of functions. If $\Omega_1 \subset \Omega_2$, then*

$$\min_{G\in\Omega_2} E||y_0 - G(y)||^2 \le \min_{H\in\Omega_1} E||y_0 - H(y)||^2 \tag{2}$$

Usually one will have a strict inequality in Eq. (2). Lemma 1 can also be used to show that the inclusion of more data will never deteriorate the performance of a 'best' predictor (and in case of a strict inequality, it will improve the performance). Let $G(y) = G(y_1, y_2)$, in which $y_1$ represents the 'old' data and $y_2$ represents the 'new' data. Then the stated claim follows if $\Omega_1$ in Eq. (2) is taken as the subset of functions $G(y_1, y_2)$ of $\Omega_2$ for which the outcome does not depend on $y_2$.

So far, we have taken the MSE with respect to the Euclidean norm $||.||$. One can also decide, however, to weight the prediction errors in the MSE and thus take the MSE with respect to the weighted norm $||.||_W$, in which $W$ is a positive semi-definite weight matrix.

As we will see, all 'best' predictors treated in the present contribution will be invariant for this choice of norm. That is, the choice of weight matrix in the norm is of no consequence for the 'best' predictors. One of the consequences of this invariance is that the 'best' predictor of a linear function of $y_0$ is equal to the same linear function of the 'best' predictor of $y_0$.

**Lemma 2** (Best prediction of linear functions) *If*

$$\begin{aligned} &E||y_0 - \hat{G}(y)||_W^2 \\ &\quad = \min_{G\in\Omega} E||y_0 - G(y)||_W^2 \quad \text{for any } W \ge 0 \end{aligned} \tag{3}$$

*then $\hat{H}(y) = F^{\mathrm{T}}\hat{G}(y) + f_0$ satisfies*

$$E||z_0 - \hat{H}(y)||^2 = \min_{H\in\Phi} E||z_0 - H(y)||^2 \tag{4}$$

*where $z_0 = F^{\mathrm{T}}y_0 + f_0$ and $\Phi = \{H|H = F^{\mathrm{T}}G + f_0, G \in \Omega\}$.*

*Proof* Since Eq. (3) holds true for any $W \ge 0$, it also holds true for $W = FF^{\mathrm{T}}$. Hence, $E||y_0 - \hat{G}(y)||_W^2 = E||(F^{\mathrm{T}}y_0 + f_0) - (F^{\mathrm{T}}\hat{G}(y) + f_0)||^2 = E||z_0 - \hat{H}(y)||^2$ and $\min_{G\in\Omega} E||y_0 - G(y)||_W^2 = \min_{G\in\Omega} E||(F^{\mathrm{T}}y_0 + f_0) - (F^{\mathrm{T}}G(y) + f_0)||^2 = \min_{H\in\Phi} E||z_0 - H(y)||^2$, from which the result follows. □

The MSE of an arbitrary predictor can often be decomposed into a sum of squares, with one of the squares being the MSE of the 'best' predictor. Lemma 3 states some general conditions under which such a decomposition is made possible. As we will see later, these conditions are satisfied by all 'best' predictors treated in this contribution.

**Lemma 3** (MSE decomposition) *Let* $\hat{G}(y)$ *and* $G(y)$*, both of class* $\Omega$*, be the best predictor and an arbitrary predictor of* $y_0$*, respectively, and let* $\hat{e}_0 = y_0 - \hat{G}(y)$ *be the error of the best predictor. If Eq. (3) holds true and* $\hat{G}(y) + \lambda\left(G(y) - \hat{G}(y)\right) \in \Omega$ *for any* $G \in \Omega$ *and any* $\lambda \in R$*, then*

$$E\left(\hat{e}_0^{\mathrm{T}} W[G(y) - \hat{G}(y)]\right) = 0 \quad \forall G \in \Omega, W \geq 0 \tag{5}$$

*and*

$$E||y_0 - G(y)||_W^2 = E||\hat{e}_0||_W^2 + E||\hat{G}(y) - G(y)||_W^2 \quad \forall G \in \Omega, W \geq 0 \tag{6}$$

*Proof* see Appendix. □

As a consequence of Lemma 3, we have the following result.

**Lemma 4** (Error-predictor covariance) *If the conditions of Lemma 3 apply and the best predictor* $\hat{G}(y)$ *is unbiased, i.e.* $E(\hat{e}_0) = E(y_0 - \hat{G}(y)) = 0$*, then*

$$Q_{\hat{e}_0 \hat{G}(y)} = Q_{\hat{e}_0 G(y)} \quad \forall G \in \Omega \tag{7}$$

*Proof* Substitution of $W = ff^{\mathrm{T}}$ into Eq. (5) gives $0 = E(\hat{e}_0^{\mathrm{T}} ff^{\mathrm{T}}[G(y) - \hat{G}(y)]) = f^{\mathrm{T}} E(\hat{e}_0[G(y) - \hat{G}(y)]^{\mathrm{T}}) f$, and therefore, since $E(\hat{e}_0) = 0$, $f^{\mathrm{T}} Q_{\hat{e}_0[G(y) - \hat{G}(y)]} f = 0$. Since this holds true for any $f \in R^{m_0}$, we have $Q_{\hat{e}_0[G(y) - \hat{G}(y)]} = 0$, from which the result follows. □

Equation (7) states that the covariance between the best prediction error and any predictor of class $\Omega$ is constant and equal to the covariance between the best prediction error and the best predictor. This property will be later used to infer the type of functions the different 'best' prediction errors are uncorrelated with.

The variance matrix of the prediction error is referred to as the *error variance matrix*. Note, since $E(||e_0||^2) = \mathrm{trace} E(e_0 e_0^{\mathrm{T}})$, the MSE is equal to the trace of the error variance matrix if the predictor is unbiased. Thus for unbiased predictors, minimization of the MSE is equivalent to minimization of the trace of the error variance matrix. Under certain conditions, this equivalence can be generalized to the variance of any linear function of the predictor error.

**Lemma 5** (Minimum error variance) *Let* $\hat{e}_0 = y_0 - \hat{G}(y)$ *and* $e_0 = y_0 - G(y)$*. If Eq. (3) holds true and* $\hat{G}(y)$ *is unbiased, then*

$$Q_{\hat{e}_0 \hat{e}_0} \leq Q_{e_0 e_0} \tag{8}$$

*for any unbiased predictor* $G \in \Omega$*.*

*Proof* It follows with $W = ff^{\mathrm{T}}$ from Eq. (3) that $f^{\mathrm{T}} E(\hat{e}_0 \hat{e}_0^{\mathrm{T}}) f \leq f^{\mathrm{T}} E(e_0 e_0^{\mathrm{T}}) f$ for any $f \in R^{m_0}$. Since $E(\hat{e}_0) = E(e_0) = 0$, the result follows. □

Under the conditions stated in Lemma 5, best predictors are *minimum error variance unbiased* predictors. Since these conditions hold true for the best predictors treated in this contribution, they are all minimum error variance unbiased predictors, albeit in different classes $\Omega$.

### 2.2 Best and best linear prediction

In this subsection, we give a brief review of the best predictor and the best linear predictor, together with their properties (e.g., Bibby and Toutenburg 1977; Koch 1980; Rao and Toutenburg 1999; Teunissen et al. 2005). They will serve as reference for the predictors treated in the subsequent sections.

**Theorem 1** (Best predictor) *A predictor* $\hat{y}_0 = \hat{G}(y)$ *is said to be the* best predictor (*BP*) *of* $y_0$ *if it satisfies* $E||y_0 - \hat{y}_0||_W^2 \leq E||y_0 - G(y)||_W^2$ *for any G. The best predictor is given by the conditional mean,*

$$\hat{y}_0 = E(y_0|y) \tag{9}$$

*Proof* see Appendix. □

Note that the BP is generally a nonlinear function of the data and that one needs the conditional PDF $f_{y_0|y}(y_0|y)$ in order to be able to compute the BP. This is, however, not needed in case one restricts the minimization of the MSE to the class of linear functions, which gives the best linear predictor.

**Theorem 2** (Best linear predictor) *A predictor* $\hat{y}_0 = \hat{G}(y)$ *is said to be the* best linear predictor (*BLP*) *of* $y_0$*, if it satisfies* $E||y_0 - \hat{y}_0||_W^2 \leq E||y_0 - G(y)||_W^2$ *for any G which is of the form* $G(y) = L_0 y + l_0$*. The BLP is given as*

$$\hat{y}_0 = \bar{y}_0 + Q_{y_0 y} Q_{yy}^{-1}(y - \bar{y}) \tag{10}$$

*where* $\bar{y}_0 = E(y_0)$ *and* $\bar{y} = E(y)$*.*

*Proof* see Appendix. □

If there is no reason for confusion, we will use the same notation for the BLP as the one used for the BP (this will also be done for the other 'best' predictors treated in the next sections). We will only use a discriminating notation, e.g. $\hat{y}_{0\mathrm{BP}}$ for the BP and $\hat{y}_{0\mathrm{BLP}}$ for the BLP, in case the need arises.

Note, as opposed to the BP, which requires the complete PDF $f_{y_0|y}(y_0|y)$, that the BLP only requires the first- and second-order moments, namely the means $\bar{y}_0$, $\bar{y}$, and the variance–covariance matrices $Q_{y_0 y}$, $Q_{yy}$. We now list the properties of both the BP and the BLP. They are similar, but since the minimization of the MSE is carried out over a more restrictive class of functions in

case of the BLP, one can also expect some of the BLP-properties to be more restrictive. On the other hand, since linearity is imposed, some of the results will be easier to compute.

**Corollary 1** (BP and BLP properties)

(i) Zero-mean error*: The BP and the BLP are both unbiased predictors. Hence, they have zero-mean prediction errors.*

(ii) Error covariance*: The BP prediction error is uncorrelated with any function of the data vector y, whereas the BLP prediction error is uncorrelated with any linear function of the data. Thus*

$$Q_{\hat{e}_0 H(y)} = 0 \tag{11}$$

*for any H in case of the BP and for any linear H in case of the BLP.*

(iii) Error variance*: The error variance matrices of the BP and the BLP are equal to the difference of the variance matrix of $y_0$ and the variance matrix of the BP and the BLP, respectively: $Q_{\hat{e}_0\hat{e}_0} = Q_{y_0y_0} - Q_{\hat{y}_0\hat{y}_0}$. Hence, their error variance matrices are given as*

$$\begin{aligned} Q^{\mathrm{BP}}_{\hat{e}_0\hat{e}_0} &= Q_{y_0y_0} - Q_{E(y_0|y)E(y_0|y)} \quad \text{and} \\ Q^{\mathrm{BLP}}_{\hat{e}_0\hat{e}_0} &= Q_{y_0y_0} - Q_{y_0y}Q^{-1}_{yy}Q_{yy_0} \end{aligned} \tag{12}$$

(iv) Minimum error variance*: The BP is a minimum error variance unbiased predictor, whereas the BLP is a minimum error variance linear unbiased predictor.*

(v) Mean squared error*: The BP and the BLP MSEs are equal to the traces of their error variance matrices. Their MSEs are related as*

$$\mathrm{MSE(BLP)} = \mathrm{MSE(BP)} + E||\hat{y}_{0\mathrm{BLP}} - \hat{y}_{0\mathrm{BP}}||^2 \tag{13}$$

(vi) Predicting a function*: The BP of a linear function of $y_0$ is the linear function of the BP of $y_0$. The same holds true for the BLP.*

(vii) Predicting the observable*: An observable is its own BP and BLP, respectively.*

(viii) Independence*: The BP reduces to the mean of $y_0$, if $y_0$ and y are independent. For the BLP this already happens in case $y_0$ and y are uncorrelated.*

(ix) Gaussian case*: The BP takes the form of the BLP in case $y_0$ and y have a joint normal distribution.*

*Proof* see Appendix. □

Equation (13) shows that the MSE of the BP is never larger than the MSE of the BLP. Similarly, the error variance matrix of the BP is never larger than the error variance matrix of the BLP. These properties are a consequence of the fact that in case of the BLP the MSE is minimized over a more restricted class of functions.

We remark that, instead of using the principle of minimizing the MSE, one can also use the above first two properties, $E(\hat{e}_0) = 0$ and $Q_{\hat{e}_0 H(y)} = 0$, as the defining principle for best prediction. To see this, with $E(\hat{e}_0) = 0$, we can write $Q_{\hat{e}_0 H(y)} = 0$ as $E\left([y_0 - \hat{G}(y)]H(y)\right) = \int[\int(y_0 f_{y_0|y}(y_0|y)\mathrm{d}y_0) - \hat{G}(y)]H(y)f_y(y)\mathrm{d}y = \int[E(y_0|y) - \hat{G}(y)]H(y)f_y(y)\mathrm{d}y = 0$. Since this 'orthogonality' relation needs to hold for any $H$, the optimal predictor follows as $\hat{G}(y) = E(y_0|y)$. Should one restrict $H$ to linear functions of $y$, then $E(\hat{e}_0) = 0$ and $Q_{\hat{e}_0 H(y)} = 0$ leads to the BLP as the best predictor. For the best predictors treated in the following sections, we will also see that $E(\hat{e}_0) = 0$ and $Q_{\hat{e}_0 H(y)} = 0$ hold true, but then for alternative classes of functions $H$.

## 3 Best linear unbiased and weighted least-squares prediction

### 3.1 The linear model for prediction

Although the requirements for the BLP are less stringent than those for the BP, the BLP still requires that the means $\bar{y}_0 = E(y_0)$ and $\bar{y} = E(y)$ be known. In many applications, this information is not available. We therefore now consider the situation where the two means $\bar{y}_0$ and $\bar{y}$ are still unknown, but linked to one another by a known linear relationship. This is a situation that holds true for many applications (see, for instance, the examples given in this and subsequent sections). Consider the following partitioned linear model,

$$\begin{bmatrix} y \\ y_0 \end{bmatrix} = \begin{bmatrix} A \\ A_0 \end{bmatrix} x + \begin{bmatrix} e \\ e_0 \end{bmatrix} \tag{14}$$

with known matrices $A$ and $A_0$ of order $m \times n$ and $m_0 \times n$, respectively, $x$ a nonrandom unknown parameter vector and $[e^{\mathrm{T}}, e_0^{\mathrm{T}}]^{\mathrm{T}}$ a random vector, with expectation and dispersion given as,

$$\begin{aligned} E\begin{bmatrix} e \\ e_0 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and} \\ D\begin{bmatrix} e \\ e_0 \end{bmatrix} &= D\begin{bmatrix} y \\ y_0 \end{bmatrix} = \begin{bmatrix} Q_{yy} & Q_{yy_0} \\ Q_{y_0y} & Q_{y_0y_0} \end{bmatrix} \end{aligned} \tag{15}$$

respectively. Matrix $A$ is assumed to be of full column rank.

As the following examples show, the above formulation of the linear prediction model also allows one to cover other formulations of the prediction problem:

*Example 1* (Prediction of random vector with unknown mean) Let $y = Ax' + e$, in which $x'$ is a random vector with known variance matrix $Q_{x'x'}$ and unknown mean $x$, and $e$ is a zero-mean random vector, uncorrelated with $x'$, with known variance matrix $Q_{ee}$. To set the stage for predicting $x'$, in Eqs. (14) and (15) we set, $e \mapsto A(x' - x) + e$, $y_0 \mapsto x'$, $A_0 \mapsto I$, and $e_0 \mapsto x' - x$ (here the notation '$a \mapsto b$' means 'replace $a$ by $b$').

*Example 2* (Predicting individual error components) Let $e$ in $y = Ax + e$ be given as $e = Dd$, with matrix $D$ known and where $d$ is a zero-mean random vector with known variance matrix $Q_{dd}$. As an application of this formulation, the entries of $d$ can be thought of as being the individual error components that contribute to the overall error vector $e$. To set the stage for predicting $d$, in Eqs. (14) and (15) we set, $e \mapsto Dd$, $y_0 \mapsto d$, $A_0 \mapsto 0$, and $e_0 \mapsto d$. For the special case $D = I$, the prediction of $e$ is covered.

*Example 3* (Trend-signal-noise model) The so-called trend-signal-noise model is another important case of Eqs. (14) and (15). The trend-signal-noise model has found wide-spread application in the spatial and Earth sciences (e.g., Moritz 1980; Stark 1987; Journel and Huijbregts 1991; Cressie 1991; Wackernagel 1995). In this model, the observable vector $y$ is written as a sum of three terms, $y = Ax+s+n$, with $Ax$ a deterministic trend, with an unknown parameter vector $x$, $s$ a zero-mean random signal vector, and $n$ a zero-mean random noise vector.

To predict the signal $s$ and noise $n$, one can proceed as described in Example 2. Often one can extend the trend-signal-noise model so as to hold true for an unobservable vector $y_0 = A_0x+s_0+n_0$, in which $s_0$ and $n_0$ are uncorrelated zero-mean random vectors, and $n_0$ is also uncorrelated with $n$. For instance, $y_0$ could be a functional of the same type as $y$, but evaluated at a different location in space or at a different instant in time. To set the stage for predicting $y_0, s_0$ and $n_0$, in Eqs. (14) and (15) we set, $e \mapsto s + n$, $y_0 \mapsto (y_0^{\mathrm{T}}, s_0^{\mathrm{T}}, n_0^{\mathrm{T}})^{\mathrm{T}}$, $A_0 \mapsto (A_0^{\mathrm{T}}, 0, 0)^{\mathrm{T}}$, and $e_0 \mapsto \big((s_0 + n_0)^{\mathrm{T}}, s_0^{\mathrm{T}}, n_0^{\mathrm{T}}\big)^{\mathrm{T}}$.

Although our goal is to use the observable random vector $y$ to predict the unobservable random vector $y_0$, at this point it is useful to include the concept of estimation into our considerations as well. Recall that we speak of prediction if a function of an observable random vector $y$ is used to guess the outcome of another random, but unobservable, vector $y_0$. We speak of estimation, however, if a function of $y$ is used to guess the value of a deterministic, but unknown, parameter vector $x$, or a function thereof. As Lemma 6 shows, the assumptions of the above linear model, with its known linear relationship between the unknown means of $y$ and $y_0$, respectively, imply that MSE-based estimation may be considered a special case of MSE-based prediction.

**Lemma 6** (Prediction versus estimation) *Let $y$ and $y_0$ satisfy Eqs.* (14) *and* (15)*, and let $G(y)$ be a best predictor of $y_0$ within a certain class $\Omega$. Then $G(y)$ reduces to the best estimator of $A_0x$ within the same class, if $e_0$ is identically zero.*

*Proof* If $e_0$ is identically zero, then $f_{y_0y}(y_0, y) = \delta(y_0 - A_0x)f_y(y)$, in which $\delta(\tau)$ is the Dirac impulse function (with the properties: $\int \delta(\tau)\mathrm{d}\tau = 1$ and $\int g(\tau)\delta(\tau - v)\mathrm{d}\tau = g(v)$). Thus if $e_0$ is identically zero, the MSE of a predictor $G(y)$ of $y_0$ becomes $E||y_0 - G(y)||^2 = \iint ||y_0 - G(y)||^2 f_{y_0y}(y_0, y)\mathrm{d}y_0\mathrm{d}y = \int ||A_0x - G(y)||^2 f_y(y)\mathrm{d}y = E||A_0x - G(y)||^2$, which is the MSE of $G(y)$ as estimator of $A_0x$. □

Thus in the present context, prediction is a more general concept than estimation. As a consequence, best predictors and best estimators come in pairs. We will take advantage of this in the next sections, by also identifying the best estimators.

### 3.2 Best linear unbiased prediction

Given the model of Eqs. (14) and (15), in which both means $\bar{y} = Ax$ and $\bar{y}_0 = A_0x$ are unknown, one cannot use the BLP to predict $y_0$ in a MMSE-sense. This is also apparant if one considers the second term on the right-hand side of Eq. (53) in the Appendix, which - in the present case - reads $||\bar{y}_0 - L_0\bar{y} - l_0||^2_W = ||(A_0 - L_0A)x - l_0||^2_W$. Setting this term equal to zero by choosing the optimal $l_0$ equal to $\bar{y}_0 - L_0\bar{y}$, as was done in the case of the BLP, would now not help as it would give a value for $l_0$ that still depends on the unknown $x$.

To make the dependence on $x$ disappear, the approach taken is to consider only those values for $L_0$ that satisfy $L_0A = A_0$. With this choice and the choice $l_0 = 0$, we again achieve that the second term on the right-hand side of Eq. (53) becomes equal to zero. The consequence of this choice is of course that we are now considering the minimization of the MSE over a restricted class of linear predictors, namely the linear predictors $G(y) = L_0y + l_0$ for which $L_0A = A_0$ and $l_0 = 0$. This class is referred to as the class of linear unbiased predictors, since $E(G(y)) = L_0Ax + l_0 = A_0x = E(y_0)$.

We will now give a useful representation of the class of linear unbiased predictors.

**Lemma 7** (Linear unbiased predictors) *A linear predictor $G(y) = L_0y + l_0$ is said to be a* linear unbiased predictor *(LUP), with respect to the linear model as defined in Eqs.* (14) *and* (15)*, if $L_0A = A_0$ and $l_0 = 0$. Let*

*$G(y)$ be a LUP. Then an $m_0 \times (m-n)$ matrix $H$ exists such that*

$$G(y) = A_0\hat{x} + Ht \tag{16}$$

*where $\hat{x} = (A^{\mathrm{T}}Q_{yy}^{-1}A)^{-1}A^{\mathrm{T}}Q_{yy}^{-1}y$, $t = B^{\mathrm{T}}y$, and $B$ is an $m \times (m-n)$ matrix of which the columns span the null space of $A^{\mathrm{T}}$.*

*Proof* The sought-for representation follows from solving the matrix equation $L_0A = A_0$ or its transposed form $A^{\mathrm{T}}L_0^{\mathrm{T}} = A_0^{\mathrm{T}}$. The general solution of this transposed form is given by the sum of its homogeneous solution and a particular solution. Since $BH^{\mathrm{T}}$ is the homogeneous solution and $Q_{yy}^{-1}A(A^{\mathrm{T}}Q_{yy}^{-1}A)^{-1}A_0^{\mathrm{T}}$ is a particular solution, the general solution for $L_0$ follows as $L_0 = A_0(A^{\mathrm{T}}Q_{yy}^{-1}A)^{-1}A^{\mathrm{T}}Q_{yy}^{-1} + HB^{\mathrm{T}}$. Substitution of this solution into $G(y) = L_0y + l_0$ gives, with $l_0 = 0$, the result Eq. (16). □

The vector $(\hat{x}^{\mathrm{T}}, t^{\mathrm{T}})^{\mathrm{T}}$ in Eq. (16) stands in a one-to-one relation with the data vector $y$. We have

$$\begin{bmatrix}\hat{x}\\ t\end{bmatrix} = \begin{bmatrix}(A^{\mathrm{T}}Q_{yy}^{-1}A)^{-1}A^{\mathrm{T}}Q_{yy}^{-1}\\ B^{\mathrm{T}}\end{bmatrix} y$$
$$\Leftrightarrow\ y = \left[A, Q_{yy}B(B^{\mathrm{T}}Q_{yy}B)^{-1}\right]\begin{bmatrix}\hat{x}\\ t\end{bmatrix} \tag{17}$$

Note that $E(t) = 0$ and that $\hat{x}$ and $t$ are uncorrelated. They are independent when $y$ is normally distributed.

The $(m-n)$-vector $t$, which will be referred to as the redundancy vector of misclosures, is identically zero in the absence of redundancy (the full rank matrix $A$ will then be a square matrix with $m = n$). Thus, in the absence of redundancy, only a single LUP exists, namely $G(y) = A_0\hat{x} = A_0A^{-1}y$. Hence, it is the presence of redundancy ($m > n$) that gives the freedom to select a best predictor from the class of linear unbiased predictors. This best linear unbiased predictor follows from minimizing the MSE over this more restricted class of linear predictors.

**Theorem 3** (Best linear unbiased predictor) *Given the linear model as defined in Eqs.* (14) *and* (15)*, a predictor $\hat{y}_0 = \hat{G}(y)$ is said to be the* best linear unbiased predictor *(BLUP) of $y_0$, if it satisfies $E||y_0 - \hat{y}_0||_W^2 \le E||y_0 - G(y)||_W^2$ for any $G$, which is of the form as given in Eq.* (16)*. The BLUP is given as*

$$\hat{y}_0 = A_0\hat{x} + Q_{y_0t}Q_{tt}^{-1}t$$
$$= A_0\hat{x} + Q_{y_0y}Q_{yy}^{-1}(y - A\hat{x}) \tag{18}$$

*Proof* see Appendix. □

The first expression of Eq. (18) is given to explictly show the LUP-structure (i.e., $H = Q_{y_0t}Q_{tt}^{-1}$; see Eq. 16). Note, with reference to Lemma 6, that Theorem 3 generalizes the classical Gauss–Markov theorem of best linear unbiased estimation (BLUE). That is, the Gauss-Markov theorem on BLUE can be considered a corollary of Theorem 3. If $e_0$ in Eq. (14) is identically zero, then $Q_{y_0t} = 0, Q_{y_0y} = 0$ and Eq. (18) reduces to $\hat{y}_0 = A_0\hat{x}$, which is the expression for the BLUE of $E(y_0) = A_0x$. The BLUE-property of $\hat{y}_0 = A_0\hat{x}$ is a consequence of the minimum error variance property of the BLUP (see *(iv)* of Corollary 4). The minimum error variance of $\hat{e}_0 = y_0 - \hat{y}_0$ becomes, since $y_0$ is now nonrandom, a minimum variance of $\hat{y}_0$.

The following three examples show the BLUP at work.

*Example 4* (Predicting individual error components)

If we use the settings as given in Example 2 and apply Eq. (18), we obtain the BLUP of $d$ as $\hat{d} = Q_{dd}D^{\mathrm{T}}(DQ_{dd}D^{\mathrm{T}})^{-1}(y - A\hat{x})$. Note that for the special case $D = I$, we obtain the BLUP of $e$ as $y - A\hat{x}$, which is also known as the least-squares residual.

*Example 5* (Separation of trend, signal and noise) Consider the problem of separating the trend, signal and noise in $y = Ax + s + n$. If we use the settings $e \mapsto s + n$, $y_0 \mapsto (s^{\mathrm{T}}, n^{\mathrm{T}})^{\mathrm{T}}$, $A_0 \mapsto 0$, $e_0 \mapsto (s^{\mathrm{T}}, n^{\mathrm{T}})^{\mathrm{T}}$ in the linear model of Eqs. (14) and (15), and apply Eq. (18), we obtain

$$\hat{x} = \left(A^{\mathrm{T}}(Q_{ss} + Q_{nn})^{-1}A\right)^{-1}A^{\mathrm{T}}(Q_{ss} + Q_{nn})^{-1}y$$
$$\hat{s} = Q_{ss}(Q_{ss} + Q_{nn})^{-1}(y - A\hat{x})$$
$$\hat{n} = Q_{nn}(Q_{ss} + Q_{nn})^{-1}(y - A\hat{x})$$

Note that $y = A\hat{x} + \hat{s} + \hat{n}$, which reflects the property that the observable is its own BLUP (see also property *(vii)* of Corollary 4).

*Example 6* (Ionospheric prediction) Consider as a trend-signal-noise model, the single-frequency, single epoch, geometry-free GPS equations, based on double-differenced (DD) carrier phase and pseudorange,

$$y_1 = \lambda x_1 + x_2 + s + n_1$$
$$y_2 = \quad\ + x_2 - s + n_2$$

with $x_1$ the unknown integer DD carrier phase ambiguity, $\lambda$ the known wavelength of the carrier phase, $x_2$ the unknown DD range, $s$ the residual ionospheric signal, and $n_1$ and $n_2$ the noise of the carrier phase and the pseudorange, respectively.

Let $\sigma_1^2$ and $\sigma_2^2$ denote the variances of the DD carrier phase and pseudorange, respectively, and let $\sigma_s^2$ denote the variance of the ionospheric signal. Then the BLUE

of $x$ and its variance matrix are given as

$$\begin{bmatrix}\hat{x}_1\\ \hat{x}_2\end{bmatrix}=\begin{bmatrix}(y_1-y_2)/\lambda\\ y_2\end{bmatrix} \text{ and}$$
$$Q_{\hat{x}\hat{x}}=\frac{1}{\lambda^2}\begin{bmatrix}4\sigma_s^2+\sigma_1^2+\sigma_2^2 & -\lambda(2\sigma_s^2+\sigma_2^2)\\ -\lambda(2\sigma_s^2+\sigma_2^2) & \lambda^2(\sigma_s^2+\sigma_2^2)\end{bmatrix}$$

If we want to predict the signal $s_0$ (e.g. the residual ionospheric delay at another time instant), then $s_0$ plays the role of $y_0$ and thus

$$Q_{s_0y}=[\sigma_{s_0s},-\sigma_{s_0s}],\quad Q_{yy}=\begin{bmatrix}\sigma_s^2+\sigma_1^2 & -\sigma_s^2\\ -\sigma_s^2 & \sigma_s^2+\sigma_2^2\end{bmatrix}$$

from which the BLUP $\hat{s}_0=Q_{s_0y}Q_{yy}^{-1}(y-A\hat{x})$ works out as

$$\hat{s}_0=\frac{\sigma_{s_0s}/\sigma_1^2}{1+\sigma_s^2/\sigma_1^2+\sigma_s^2/\sigma_2^2}\left[(y_1-\lambda\hat{x}_1-\hat{x}_2)-\frac{\sigma_1^2}{\sigma_2^2}(y_2-\hat{x}_2)\right]$$

Note that this predictor has not made use of the fact that $x_1$ is integer-valued (see Example 12).

Note that the structure of the BLUP resembles that of the BLP [cf. Eqs. (18) and (10)]. The BLUP is obtained from the expression of the BLP, by replacing the (unknown) means $\bar{y}_0$ and $\bar{y}$ by their BLUEs $A_0\hat{x}$ and $A\hat{x}$, respectively. Since the class of LUPs is a subset of the class of LPs, the error variance performance and the MSE performance of the BLUP will be poorer than that of the BLP. This is made precise in the following corollary.

**Corollary 2** (BLUP and BLP compared)

(i) Error variance*: The error variance matrices of the BLUP and the BLP are related by*

$$Q_{\hat{e}_0\hat{e}_0}^{\text{BLUP}}=Q_{\hat{e}_0\hat{e}_0}^{\text{BLP}}+A_{0|y}Q_{\hat{x}\hat{x}}A_{0|y}^{\text{T}} \tag{19}$$

*where* $A_{0|y}=A_0-Q_{y_0y}Q_{yy}^{-1}A$.

(ii) Mean squared error*: The MSEs of the BLUP and the BLP are related as*

$$\text{MSE(BLUP)}=\text{MSE(BLP)}+E||\hat{y}_{0\text{BLUP}}-\hat{y}_{0\text{BLP}}||^2 \tag{20}$$

*Proof* (*i*) To prove Eq. (19), we first note that the term within the brackets on the right-hand side of $\hat{y}_{0\text{BLUP}}=(y_0-Q_{y_0y}Q_{yy}^{-1}y)+A_{0|y}\hat{x}$ is uncorrelated with $y$, and therefore also uncorrelated with $\hat{x}$. The result follows then from an application of the variance propagation law. (*ii*) Follows from an application of Lemma 3, cf. Eq. (6), with $G$ as the BLUP and $\hat{G}$ as the BLP. □

Note that Eq. (19) clearly shows, with reference to the BLP, that the additional uncertainty in the BLUP prediction error is due to the uncertainty in estimating the unknown parameter vector $x$. More properties of the BLUP will be given in Sect. 4, where we compare the BLUP with the best equivariant predictor. First, however, we will give a deterministic least-squares-based interpretation of the BLUP.

### 3.3 Weighted Least-Squares Prediction

It is well-known that any weighted least-squares estimator of $x$ in Eq. (14) is a member from the class of linear unbiased estimators of $x$. It is also known that the weighted least-squares estimator, which uses the inverse of the variance matrix $Q_{yy}$ as weight matrix, is identical to the BLUE of $x$. In this subsection, we will generalize this equivalence to the problem of prediction.

The objective function that we will work with is given by the positive definite quadratic form,

$$F(y,y_0,x)=\begin{bmatrix}y-Ax\\ y_0-A_0x\end{bmatrix}^{\text{T}}\begin{bmatrix}W_{yy} & W_{yy_0}\\ W_{y_0y} & W_{y_0y_0}\end{bmatrix}\begin{bmatrix}y-Ax\\ y_0-A_0x\end{bmatrix} \tag{21}$$

If $y$ and $y_0$ are observable and $x$ is unknown, then the unique $\hat{x}'$ satisfying $F(y,y_0,\hat{x}')\le F(y,y_0,x)$, for all $x\in R^n$, is said to be a (weighted) least-squares estimator of $x$ based on both $y$ and $y_0$. If $y$ is observable, $x$ is known and $y_0$ is unobservable, then the unique $\hat{y}_0'$ satisfying $F(y,\hat{y}_0',x)\le F(y,y_0,x)$, for all $y_0\in R^{m_0}$, is said to be a (weighted) least-squares predictor of $y_0$. We are interested in the case that is a combination of the previous two problems. Thus, as before, we assume $y$ to be observable, $x$ to be unknown and $y_0$ to be unobservable.

**Theorem 4 (Weighted least-squares predictor)**

*Given the objective function of Eq.* (21)*, the unique pair* $\hat{x}_{\text{WLSE}}$, $\hat{y}_{0\text{WLSP}}$ *satisfying* $F(y,\hat{y}_{0\text{WLSP}},\hat{x}_{\text{WLSE}})\le F(y,y_0,x)$*, for all* $x\in R^n$, $y_0\in R^{m_0}$*, is said to be the* weighted least-squares estimator–predictor *(WLSE–WLSP) pair of x,* $y_0$*. This pair is given as*

$$\hat{x}_{\text{WLSE}}=(A^{\text{T}}W_{yy|y_0}A)^{-1}A^{\text{T}}W_{yy|y_0}y$$
$$\hat{y}_{0\text{WLSP}}=A_0\hat{x}_{\text{WLSE}}-W_{y_0y_0}^{-1}W_{y_0y}(y-A\hat{x}_{\text{WLSE}}) \tag{22}$$
*with* $W_{yy|y_0}=W_{yy}-W_{yy_0}W_{y_0y_0}^{-1}W_{y_0y}$.

*Proof* see Appendix. □

Note, since $\hat{y}_{0\text{WLSP}}$ is unbiased and a linear function of $y$, that $\hat{y}_{0\text{WLSP}}$ is a LUP of $y_0$ for any choice of the weight matrix in Eq. (21). Hence, the MSE property of $\hat{y}_{0\text{WLSP}}$ will, in general, be inferior to that of the BLUP. For a particular choice of the weight matrix, however, the WLSP becomes identical to the BLUP.

**Corollary 3 (BLUP as WLSP)** *If the weight matrix in Eq. (21) is chosen equal to the inverse of the joint variance matrix of $y$ and $y_0$, then $\hat{x}_{\mathrm{WLSE}}$ and $\hat{y}_{0\mathrm{WLSP}}$ become identical to the BLUE of $x$ and the BLUP of $y_0$, respectively.*

*Proof* If the weight matrix of Eq. (21) is equal to the inverse of the variance matrix of Eq. (15), then $W_{yy|y_0} = Q_{yy}^{-1}$ and $W_{y_0y_0}^{-1}W_{y_0y} = -Q_{y_0y}Q_{yy}^{-1}$. With this result, the estimator–predictor pair of Eq. (22) becomes identical to the BLUE–BLUP pair. □

Corollary 3 generalizes the relation that exists between least-squares estimation and BLUE, to that between least-squares prediction and BLUP. Note that a similar least-squares-based interpretation can be given to the BLP. If $x$ is assumed known and the quadratic form of Eq. (21) is minimized as function of $y_0$, then the resulting WLSP reads $\hat{y}'_{0\mathrm{WLSP}} = A_0x - W_{y_0y_0}^{-1}W_{y_0y}(y - Ax)$, which becomes identical to the BLP if $W_{y_0y_0}^{-1}W_{y_0y} = -Q_{y_0y}Q_{yy}^{-1}$.

## 4 Best equivariant prediction

### 4.1 Equivariant prediction

So far, we considered nonlinear and linear predictors. We will now introduce a new class of predictors for the linear model of Eqs. (14) and (15). This class will be larger than the class of linear unbiased predictors. It is quite natural that one ends up with the class of LUPs for the linear model, if one starts from the class of LPs. Thus if one starts from the class of LPs, one has to enforce the unbiasedness condition in order to ensure that predictors are obtained that are independent of the unknown parameter vector $x$.

Starting from the class of LPs is not needed, however, since one can start from a larger class and still do justice to the equivariance that is present in the linear model. The idea is as follows. Assume that $y$ in Eq. (14) is perturbed by $A\alpha$. Then $x$ gets perturbed by $\alpha$ and $y_0$ gets perturbed by $A_0\alpha$. When designing a predictor of $y_0$, it therefore seems reasonable to request that any such predictor, being a function of $y$, behaves in the same way with regard to such perturbations. Predictors that have this property will be called equivariant predictors.

**Definition 2** (Equivariant predictors) The predictor $G(y)$ is said to be an *equivariant predictor* (EP) of $y_0$, with respect to the linear model as defined in Eqs. (14) and (15), if

$$G(y + A\alpha) = G(y) + A_0\alpha \quad \forall y \in R^m, \alpha \in R^n \tag{23}$$

Note that equivariant predictors need not be linear. They only behave linearly with respect to pertubations $A\alpha$.

**Lemma 8** (LUP $\subset$ EP) *A linear predictor (LP) $G(y) = L_0y + l_0$ is an equivariant predictor if and only if $L_0A = A_0$. Hence, any linear unbiased predictor is an equivariant predictor.*

*Proof* We have $G(y + A\alpha) = L_0(y + A\alpha) + l_0 = G(y) + L_0A\alpha$. Hence, $G(y + A\alpha) = G(y) + A_0\alpha$ for all $\alpha \in R^n$, if and only if $L_0A = A_0$. Since this condition on $L_0$ is satisfied by LUPs, any LUP is an equivariant predictor. □

Note, since any LUP is an EP, the BLUP and the WLSP are both EPs. Lemma 8 also shows that the set of EPs and the set of LPs have an overlap in which the subset of LUPs resides. In Eq. (16), we gave a representation of LUPs. We will now give an equivalent representation of EPs.

**Lemma 9** (EP representation) *Let $G(y)$ be an EP of $y_0$. Then a function $H : R^{m-n} \mapsto R^{m_0}$ exists such that*

$$G(y) = A_0\hat{x} + H(t) \tag{24}$$

*Proof* Here we make use of the reparametrization $y = A\hat{x} + Ct$, where $C = Q_{yy}B(B^{\mathrm{T}}Q_{yy}B)^{-1}$. This reparametrization establishes a one-to-one relation between $y$ and $(\hat{x}^{\mathrm{T}}, t^{\mathrm{T}})^{\mathrm{T}}$ (see Eq. 17). First, we prove that any $G$ of the form given in Eq. (24) is an EP. With $y = A\hat{x} + Ct$ and $y' = y + A\alpha$, we have $y' = A(\hat{x} + \alpha) + Ct = A\hat{x}' + Ct$. Therefore, $G(y + A\alpha) = G(y') = A_0\hat{x}' + H(t) = A_0(\hat{x} + \alpha) + H(t) = G(y) + A_0\alpha$. We now prove the converse. If we choose $\alpha = -\hat{x}$ in $G(y + A\alpha) = G(y) + A_0\alpha$, then $G(y) = A_0\hat{x} + G(y - A\hat{x})$, where $G(y - A\hat{x})$ is a function of only $t$, since $y = A\hat{x} + Ct$. □

Compare Eq. (24) with Eq. (16). It shows that the difference between the two classes of predictors, LUP and EP, lies in the way use can be made of the vector of misclosures $t$. In case of the LUPs, only linear functions of $t$ are considered, whereas in case of the EPs, nonlinear functions of $t$ are also permitted. Note that an EP is unbiased if and only if $E(H(t)) = 0$. Also note that, if redundancy is absent and thus $t$ is identically zero, only a single unbiased EP is left, namely $G(y) = A_0A^{-1}y$.

### 4.2 Best equivariant predictor

Now that we have defined and characterized the class of equivariant predictors, we are in the position to select the best equivariant predictor.

**Theorem 5** (Best equivariant predictor) *Given the linear model as defined in Eqs.* (14) *and* (15)*, a predictor* $\hat{y}_0 = \hat{G}(y)$ *is said to be the* best equivariant predictor *(BEP) of* $y_0$*, if* $E||y_0 - \hat{y}_0||^2_W \leq E||y_0 - G(y)||^2_W$ *for any G, which is of the form as given in Eq.* (24)*. The BEP is given as*

$$\begin{aligned}\hat{y}_0 &= A_0\hat{x} + E(y_0 - A_0\hat{x}|t) \\ &= A_0\hat{x} + \int_{R^n} E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v,t) f_{\hat{x}|t}(v|t)\mathrm{d}v \end{aligned} \tag{25}$$

*Proof* see Appendix. □

The second expression for the BEP in Eq. (25) has been included so as to make an easy comparison possible with the 'best' predictor of the next section. Note that the BEP is the sum of $A_0\hat{x}$ and the BP of $y_0 - A_0\hat{x}$ based on the 'data vector' $t$.

*Example 7* Prediction of individual error components) To determine the BEP of the zero-mean random vector $d$ in $y = Ax + Dd$, we take the settings in Example 4 and apply Eq. (25). This gives the BEP as $\hat{d} = E(d|t)$. Similarly, one finds the BEP of $e$ as $\hat{e} = E(e|t)$.

If $e_0$ in Eq. (14) is identically zero, we obtain, from Eq. (25), the best equivariant estimator (BEE) of $E(y_0) = A_0x$ as $A_0\hat{x} + A_0E(x - \hat{x}|t)$. Since the BEP of $e_0$ is given as $E(e_0|t)$, it follows that the BEP of $y_0 = A_0x + e_0$ can be decomposed as the sum of the BEE of $E(y_0) = A_0x$ and the BEP of $e_0$.

We now list and compare the properties of the BEP and the BLUP, respectively.

**Corollary 4** (BEP and BLUP properties)

(i) Zero-mean error*: The BEP and the BLUP are both unbiased predictors. Hence, they have zero-mean prediction errors.*

(ii) Error covariance*: The BEP prediction error is uncorrelated with any function of the vector of misclosures t, whereas the BLUP prediction error is uncorrelated with any linear function of the misclosures. Thus*

$$Q_{\hat{e}_0H(t)} = 0 \tag{26}$$

*for any H in case of the BEP and for any linear H in case of the BLUP.*

(iii) Error variance*: The error variance matrices of the BEP and the BLUP are equal to the difference of the variance matrix of* $y_0 - A_0\hat{x}$ *and the variance matrix of the BEP and the BLUP of* $y_0 - A_0\hat{x}$*, respectively,* $Q_{\hat{e}_0\hat{e}_0} = Q_{(y_0 - A_0\hat{x})(y_0 - A_0\hat{x})} - Q_{(\hat{y}_0 - A_0\hat{x})(\hat{y}_0 - A_0\hat{x})}$*. Hence, their error variance matrices are given as*

$$\begin{aligned} Q^{\mathrm{BEP}}_{\hat{e}_0\hat{e}_0} &= Q_{(y_0 - A_0\hat{x})(y_0 - A_0\hat{x})} \\ &\quad - Q_{E(y_0 - A_0\hat{x}|t)E(y_0 - A_0\hat{x}|t)} \\ Q^{\mathrm{BLUP}}_{\hat{e}_0\hat{e}_0} &= Q_{(y_0 - A_0\hat{x})(y_0 - A_0\hat{x})} - Q_{y_0t}Q_{tt}^{-1}Q_{ty_0} \end{aligned} \tag{27}$$

(iv) Minimum error variance*: The BEP is a minimum error variance equivariant unbiased predictor, whereas the BLUP is a minimum error variance linear unbiased predictor.*

(v) Mean squared error*: The BEP and the BLUP MSEs are equal to the traces of their error variance matrices. Their MSEs are related by*

$$\begin{aligned}\mathrm{MSE(BLUP)} = \mathrm{MSE(BEP)} + E||\hat{y}_{0\mathrm{BLUP}} \\ - \hat{y}_{0\mathrm{BEP}}||^2 \end{aligned} \tag{28}$$

(vi) Predicting a function*: The BEP of a linear function of* $y_0$ *is the linear function of the BP of* $y_0$*. The same holds true for the BLUP.*

(vii) Predicting the observable*: An observable is its own BEP and BLUP, respectively.*

(viii) Independence*: The BEP reduces to the BLUE of* $E(y_0)$*, if both* $y_0$ *and* $\hat{x}$ *are independent of t. For the BLUP this already happens in case* $y_0$ *and t are uncorrelated.*

(ix) Gaussian case*: The BEP takes the form of the BLUP in case* $y_0$ *and y have a joint normal distribution.*

*Proof* see Appendix. □

Corollary 4 shows that the BEP and the BLUP are both unbiased, just like the BP and the BLP. There is, however, one marked difference between these four predictors, namely that in the case of the BLUP, unbiasedness is enforced a priori, this in contrast to the other three predictors. Thus, although the BLP and the BLUP are both minimum error variance linear unbiased predictors, the minimum error variance of the BLUP has been achieved in a more restrictive class. Corollary 4 also shows that the BEP outperforms the BLUP, in terms of both the MSE and the error variance. Similarly, the BEE outperforms the BLUE. Both are unbiased, but the variance of the BEE is smaller or, at most, equal to that of the BLUE.

Finally note that there exist analogies between the BP–BLP pair and the BEP–BLUP pair (compare Corollaries 1 and 4). In the Gaussian case, the BP takes the form of the BLP and the BEP takes the form of the BLUP. This implies that, in the Gaussian case, the

minimum error variance of the BLUP holds true for a larger class than the LUPs. As another analogy, we note that the prediction errors of the first pair are uncorrelated with any function of $y$ and any linear function of $y$, respectively, whereas the prediction errors of the second pair are uncorrelated with any function of $t$ and any linear function of $t$, respectively. Thus, in the case of the BEP–BLUP pair, the redundancy vector of misclosures $t$ takes over the role of $y$. This is also clear if one considers the representations of the four different classes of predictors. In the case of the classes of arbitrary predictors and LPs, the predictors are represented by arbitrary functions and linear functions of $y$, respectively, whereas in case of the classes of EPs and LUPs, the predictors are represented by $A_0\hat{x}$ plus arbitrary functions and linear functions of $t$, respectively.

## 5 Best integer equivariant prediction

### 5.1 Integer equivariant prediction

So far, the unknown parameter vector of our linear model was considered to be real-valued, $x \in R^n$. Now we will assume it to be integer-valued, $x \in Z^n$. Although the BLUP and the BEP are still applicable to this case, these 'best' predictors do not make use of the fact that $x$ is now an integer vector. We therefore introduce a new class of predictors, which does take this information into account. The approach used is similar to that of equivariant prediction, but the equivariance is now assumed to hold only for integer pertubations. This leads to the class of integer equivariant predictors.

**Definition 3** (Integer equivariant prediction) The predictor $G(y)$ is said to be an *integer equivariant predictor* (IEP) of $y_0$, with respect to the linear model as defined in Eqs. (14) and (15), if

$$G(y + Az) = G(y) + A_0 z \quad \forall y \in R^m, z \in Z^n \tag{29}$$

Comparing this definition with Definition 2 [cf. Eq. 23], it will be clear that the class of IEPs is larger than the class of EPs, which, according to Lemma 7, is again larger than the class of LUPs. Hence, we have the following ordering for these three classes of predictors: LUP $\subset$ EP $\subset$ IEP. The MSE of the best IEP will therefore not be larger (and, in fact, in most cases smaller) than the MSEs of the BEP and the BLUP. Before we determine the MSE of the best IEP, we first give a useful representation of the IEPs.

**Lemma 10** (IEP representation) *Let $G(y)$ be an IEP of $y_0$. Then a function $H : R^n \times R^{m-n} \mapsto R^{m_0}$ exists such that*

$$G(y) = A_0\hat{x} + H(\hat{x}, t) \tag{30}$$

*where $H(\hat{x} + z, t) = H(\hat{x}, t)$, $\forall z \in Z^n$.*

*Proof* We make use of the reparametrization $y = A\hat{x} + Ct$, where $C = Q_{yy}B(B^{\mathrm{T}}Q_{yy}B)^{-1}$. First we prove that any $G$ of the form given in Eq. (30) is an IEP. With $y = A\hat{x} + Ct$ and $y' = y + Az$, we have $y' = A(\hat{x} + z) + Ct = A\hat{x}' + Ct$. Therefore, $G(y + Az) = G(y') = A_0\hat{x}' + H(\hat{x}', t) = A_0(\hat{x} + z) + H(\hat{x}, t) = G(y) + A_0 z$. We now prove the converse. If we subtract $A_0(\hat{x} + z)$ from both sides of $G(y + Az) = G(y) + A_0 z$, we obtain $G(y + Az) - A_0(\hat{x} + z) = G(y) - A_0\hat{x}$. Now let $H(\hat{x}, t) = G(y) - A_0\hat{x}$, then $H(\hat{x} + z, t) = H(\hat{x}, t)$. □

Comparing Eq. (30) with Eqs. (24) and (16) shows the differences among the three classes of predictors. In case of an IEP, the function $H$ depends on both $\hat{x}$ and $t$, but is invariant for an integer pertubation in its first slot. In case of an EP, the dependence on $\hat{x}$ is absent and $H$ is only a function of $t$, whereas in case of a LUP, the dependence on $t$ is reduced to a linear one. Also note that, if redundancy is absent and thus $t$ is identically zero, different IEPs still exist. This in contrast with the LUPs and unbiased EPs.

Since LUPs and EPs are also IEPs, the class of IEPs is richer. The following example gives an IEP which is not a LUP nor an EP.

*Example 8* (Prediction based on rounding) We assume that all entries of the parameter vector $x$ in the linear model of Eq. (14) are integer-valued. Let $\lceil\hat{x}\rfloor$ denote the integer vector that is obtained by rounding all the entries of $\hat{x}$ to their nearest integer. Then

$$\hat{y}_{0\mathrm{IEP}} = A_0\lceil\hat{x}\rfloor + Q_{y_0y}Q_{yy}^{-1}(y - A\lceil\hat{x}\rfloor)$$

is an IEP of $y_0$. This predictor has the same structure as the BLUP. In fact, it has been obtained from the expression of the BLUP, by replacing $\hat{x}$ by the integer vector $\lceil\hat{x}\rfloor$. Note that $\hat{y}_{0\mathrm{IEP}}$ can be written as $\hat{y}_{0\mathrm{IEP}} = A_0\hat{x} + H(\hat{x}, t)$, with $H(\hat{x}, t) = Q_{y_0t}Q_{tt}^{-1}t - A_{0|y}(\hat{x} - \lceil\hat{x}\rfloor)$ and $A_{0|y} = A_0 - Q_{y_0y}Q_{yy}^{-1}A$. Thus, since $H(\hat{x}, t)$ is invariant for integer pertubations in its first slot, the predictor $\hat{y}_{0\mathrm{IEP}}$ is indeed an IEP. Other IEPs can be obtained in a similar way. If one replaces $\lceil\hat{x}\rfloor$ by any other integer estimator, e.g. the integer bootstrapped estimator or the integer least-squares estimator (Teunissen 1999), then again an IEP of $y_0$ is obtained.

### 5.2 Best integer equivariant predictor

Now that we have defined the class of IEPs, we are in the position to determine the best predictor of this class.

**Theorem 6 (Best integer equivariant predictor)** *Given the linear model as defined in Eqs. (14) and (15), a predictor $\hat{y}_0 = \hat{G}(y)$ is said to be the* best integer equivariant predictor *(BIEP) of $y_0$, if $E||y_0-\hat{y}_0||^2_W \leq E||y_0-G(y)||^2_W$ for any $G$, which is of the form given in Eq. (30). The BIEP of $y_0$ is given as*

$$\hat{y}_0 = A_0\hat{x} + \sum_{v\in\hat{x}+Z^n} E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v,t)\frac{f_{\hat{x}|t}(v|t)}{\sum_{v\in\hat{x}+Z^n} f_{\hat{x}|t}(v|t)} \tag{31}$$

*Proof* see Appendix. □

If one considers how EPs and IEPs are defined, one can expect the BEP and the BIEP to be closely related [cf. Eqs. (25) and (31)]. This shows that the BIEP follows from the BEP, if the averaging operation $\int_{R^n}[\bullet]f_{\hat{x}|t}(v|t)\mathrm{d}v$ of the BEP is replaced by the discretized version $\sum_{v\in\hat{x}+Z^n}[\bullet]f_{\hat{x}|t}(v|t)/(\sum_{v\in\hat{x}+Z^n} f_{\hat{x}|t}(v|t))$. The difference between the two predictors will therefore become less, the finer the integer grid of $Z^n$ becomes in comparison with the PDF $f_{\hat{x}|t}(v|t)$.

Just like the BLUE and the BEE may be considered special cases of the BLUP and the BEP, respectively, the best integer equivariant estimator may be considered a special case of the BIEP. This follows from Theorem 6 by assuming $e_0$ to be identically zero.

**Corollary 5** (Best integer equivariant estimator) *The* best integer equivariant estimator *(BIEE) of $E(y_0) = A_0x$ is given as*

$$\hat{y}_0 = A_0 \sum_{z\in Z^n} z\frac{f_{\hat{x}|t}(\hat{x}+x-z|t)}{\sum_{z\in Z^n} f_{\hat{x}|t}(\hat{x}+x-z|t)} \tag{32}$$

*Proof* If $e_0$ is identically zero, then $E_{y_0|\hat{x}t}(y_0-A_0\hat{x}|v,t) = A_0(x-v)$. Substitution of this result into Eq. (31) gives Eq. (32). □

We now list the properties of the BIEP.

**Corollary 6** (BIEP properties)

(i) Zero-mean error*: The BIEP is unbiased and therefore has a zero-mean prediction error.*

(ii) Error covariance*: The BIEP prediction error is uncorrelated with any function of $\hat{x}$ and $t$ which is invariant for integer pertubations of $\hat{x}$. Thus*

$$Q_{\hat{e}_0 H(\hat{x},t)} = 0 \tag{33}$$

*for any $H$ that satisfies $H(\hat{x}+z,t) = H(\hat{x},t)$, $\forall z\in Z^n$.*

(iii) Error variance*: The error variance matrix of the BIEP is equal to the difference of the variance matrices of $y_0 - A_0\hat{x}$ and $\hat{y}_0 - A_0\hat{x}$, respectively,*

$$Q_{\hat{e}_0\hat{e}_0} = Q_{(y_0-A_0\hat{x})(y_0-A_0\hat{x})} - Q_{(\hat{y}_0-A_0\hat{x})(\hat{y}_0-A_0\hat{x})} \tag{34}$$

(iv) Minimum error variance*: The BIEP is a minimum error variance integer equivariant unbiased predictor.*

(v) Mean squared error*: The MSEs of the BIEP, BEP and BLUP are equal to the traces of their error variance matrices. Their MSEs are related as*

$$\begin{aligned}\mathrm{MSE(BEP)} &= \mathrm{MSE(BIEP)} + E||\hat{y}_{0\mathrm{BEP}} - \hat{y}_{0\mathrm{BIEP}}||^2\\ \mathrm{MSE(BLUP)} &= \mathrm{MSE(BIEP)} + E||\hat{y}_{0\mathrm{BLUP}} - \hat{y}_{0\mathrm{BIEP}}||^2\end{aligned} \tag{35}$$

(vi) Predicting a function*: The BIEP of a linear function of $y_0$ is the linear function of the BIEP of $y_0$.*

(vii) Predicting the observable*: An observable is its own BIEP.*

(viii) Independence*: The BIEP reduces to the BIEE if $y_0$ and $y$ are independent.*

*Proof* see Appendix. □

The Gaussian case will be treated separately in the next section. As was pointed out in relation to the BP and the BLP, one can—in a similar fashion—also take the above two properties, $E(\hat{e}_0) = 0$ and $Q_{\hat{e}_0 H(\hat{x},t)} = 0$, as the defining principle for best integer equivariant prediction.

## 6 Best mixed equivariant prediction

### 6.1 Best mixed equivariant predictor

So far, we have considered the all-integer and all-real cases. In most applications, however, e.g. GNSS and InSAR, a part of the unknown parameters will be integer-valued (i.e., ambiguities), while the other part will be real-valued (e.g., baseline or troposphere).

To treat this mixed case, we again consider the linear model of Eqs. (14) and (15), but now with $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})^{\mathrm{T}}$, $x_1 \in Z^p$, $x_2 \in R^{n-p}$ and a likewise partitioning of the matrices, $A = (A_1, A_2)$, $A_0 = (A_{01}, A_{02})$. Thus, the first $p$ entries of $x$ are assumed to be integer-valued, while the last $n-p$ entries are assumed to be real-valued. The results of Sects. 4 and 5 can now be used as building blocks for studying the mixed case. The definition of mixed equivariant prediction follows then quite

naturally from combining the definitions of equivariant and integer equivariant prediction.

**Definition 4** (Mixed equivariant prediction) The predictor $G(y)$ is said to be a *mixed equivariant predictor* (MEP) of $y_0$, with respect to the linear model as defined in Eqs. (14) and (15), if

$$\begin{aligned} &G(y + A_1 z_1 + A_2 \alpha_2) \\ &\quad = G(y) + A_{01} z_1 + A_{02} \alpha_2 \\ &\qquad \forall y \in R^m,\ z_1 \in Z^p, \alpha_2 \in R^{n-p} \end{aligned} \tag{36}$$

Also the MEP-representation follows then quite naturally.

**Lemma 11** (MEP representation) *Let $G(y)$ be a MEP of $y_0$. Then a function $H : R^p \times R^{m-n} \mapsto R^{m_0}$ exists such that*

$$G(y) = A_0 \hat{x} + H(\hat{x}_1, t) \tag{37}$$

*where $H(\hat{x}_1 + z_1, t) = H(\hat{x}_1, t)$, $\forall z_1 \in Z^p$.*

*Proof* The proof goes along the same lines as the proof of Lemma 10, so will not be presented here. □

This representation is also very useful for deriving the best mixed equivariant predictor.

**Theorem 7** (Best mixed equivariant predictor) *Given the linear model as defined in Eqs.* (14) *and* (15)*, a predictor $\hat{y}_0 = \hat{G}(y)$ is said to be the* best mixed equivariant predictor *(BMEP) of $y_0$, if $E||y_0 - \hat{y}_0||^2_W \leq E||y_0 - G(y)||^2_W$ for any $G$, which is of the form given in Eq.* (37)*. The BMEP is given as*

$$\begin{aligned} \hat{y}_0 &= \frac{\sum_{z_1 \in Z^p} \int_{R^{n-p}} \int_{R^{m_0}} [A_{01} z_1 + A_{02} \beta_2 + e_0] f_{e_0 y}(e_0, y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2)) \mathrm{d}e_0 \mathrm{d}\beta_2}{\sum_{z_1 \in Z^p} \int_{R^{n-p}} f_y(y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2)) \mathrm{d}\beta_2} \\ &= \frac{\sum_{z_1 \in Z^p} \int_{R^{n-p}} \int_{R^{m_0}} [A_{01} z_1 + A_{02} \beta_2 + e_0] f_{e_0 \hat{x}_1 \hat{x}_2 | t}(e_0, \hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2 | t) \mathrm{d}e_0 \mathrm{d}\beta_2}{\sum_{z_1 \in Z^p} \int_{R^{n-p}} f_{\hat{x}_1 \hat{x}_2 | t}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2 | t) \mathrm{d}\beta_2} \end{aligned} \tag{38}$$

*Proof* The proof goes along the same lines as the proof of Theorem 6. The second expression in Eq. (38) follows from the first by noting that $f_{e_0 y}(e_0, y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2)) \propto f_{e_0 \hat{x}_1 \hat{x}_2 | t}(e_0, \hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2 | t)$. □

Note that the difference between the two expressions of Eq. (38) lies in the way the input is specified. The first expression requires the original data vector $y$, whereas the second expression requires the BLUE $\hat{x}$ and the misclosure vector $t$. Thus, to draw a parallel with GNSS ambiguity resolution, one can base the mixed equivariant prediction and estimation on the 'float' solution $\hat{x}, t$. We have also given the two expressions to make an easy comparison possible with our earlier results. Note that, if $p = n$, $A_2 = 0$, and $A_{02} = 0$, the BMEP takes the form of the BIEP (see Theorem 6). Similarly, the BMEP takes the form of the BEP if no integer-valued parameters are present in the linear model, that is, if $p = 0$, $A_1 = 0$, and $A_{01} = 0$ (see Theorem 5).

The BMEP inherits its properties quite naturally from the BEP and the BIEP (see Corollaries 4 and 6). As to its MSE performance, we have, since EP $\subset$ MEP $\subset$ IEP, that MSE(BIEP) $\leq$ MSE(BMEP) $\leq$ MSE(BEP). The same ordering holds true for their error variance matrices. Thus, the BMEP also outperforms the BLUP.

We have seen that the BLUP, BEP and BIEP of $y_0 = A_0 x + e_0$, can be written as the sum of the corresponding estimator of $A_0 x$ and predictor of $e_0$. An almost similar decomposition also holds true for the BMEP. From the structure of the two expressions given for the BMEP in Eq. (38), one can easily identify the best mixed equivariant estimator (BMEE) of $x$ and the BMEP of $e_0$. Note, however, that the BMEE of $x_1$ is identical to the BIEE of $x_1$, while the BMEE of $x_2$ is *not* identical to the BEE of $x_2$. This shows that knowing that $x_2$ is real-valued does not help us to improve the BIEE of $x_1$, but the knowledge that $x_1$ is integer-valued does allow us to improve the BEE of $x_2$ to the BMEE of $x_2$.

With the estimators of $x_1$ and $x_2$, and the predictor of $e_0$ identified, we can thus decompose the BMEP of $y_0$ to

$$\hat{y}_{0\mathrm{BMEP}} = A_{01} \hat{x}_{1\mathrm{BIEE}} + A_{02} \hat{x}_{2\mathrm{BMEE}} + \hat{e}_{0\mathrm{BMEP}} \tag{39}$$

with

$$\begin{cases} \hat{x}_{1\mathrm{BIEE}} = \sum_{z_1 \in Z^p} z_1 \omega_{z_1}(y), & \sum_{z_1 \in Z^p} \omega_{z_1}(y) = 1 \\ \hat{x}_{2\mathrm{BMEE}} = \int_{\beta_2 \in R^{n-p}} \beta_2 \omega_{\beta_2}(y) \mathrm{d}\beta_2, & \int_{\beta_2 \in R^{n-p}} \omega_{\beta_2}(y) \mathrm{d}\beta_2 = 1 \\ \hat{e}_{0\mathrm{BMEP}} = \int_{e_0 \in R^{m_0}} e_0 \omega_{e_0}(y) \mathrm{d}e_0, & \int_{e_0 \in R^{m_0}} \omega_{e_0}(y) \mathrm{d}e_0 = 1 \end{cases} \tag{40}$$

and

$$\begin{aligned} \omega_{z_1}(y) &= \int_{\beta_2} \omega_{z_1 \beta_2}(y) \mathrm{d}\beta_2, & \omega_{\beta_2}(y) &= \textstyle\sum_{z_1} \omega_{z_1 \beta_2}(y) \\ \omega_{z_1 \beta_2}(y) &= \int_{e_0} \omega_{e_0 z_1 \beta_2}(y) \mathrm{d}e_0, & \omega_{e_0}(y) &= \textstyle\sum_{z_1} \int_{\beta_2} \omega_{e_0 z_1 \beta_2}(y) \mathrm{d}\beta_2 \end{aligned} \tag{41}$$

where

$$\omega_{e_0 z_1 \beta_2}(y) = \frac{f_{e_0 y}\,(e_0, y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2))}{\sum_{z_1} \int f_y\,(y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2))\,\mathrm{d}\beta_2} \quad (42)$$

### 6.2 The Gaussian case

In our derivation of the BMEP, we have not yet made a particular choice for the joint PDF of $y_0$ and $y$. Hence, the results obtained up to now hold true for any PDF that $y_0$ and $y$ might have. In many applications, however, it is assumed that the joint PDF is Gaussian. The following corollary shows how the BMEP is derived for the Gaussian case.

**Corollary 7** (BMEP in Gaussian case) *If $y_0$ and $y$ have a joint normal distribution, then*

$$\hat{y}_{0\mathrm{BMEP}} = A_0 \hat{x}_{\mathrm{BMEE}} + Q_{y_0 y} Q_{yy}^{-1} (y - A\hat{x}_{\mathrm{BMEE}}) \quad (43)$$

*with* $\hat{x}_{\mathrm{BMEE}} = (\hat{x}_{1\mathrm{BIEE}}^{\mathrm{T}}, \hat{x}_{2\mathrm{BMEE}}^{\mathrm{T}})^{\mathrm{T}}$, *where*

$$\hat{x}_{1\mathrm{BIEE}} = \sum_{z_1 \in Z^p} z_1 \frac{\exp\left\{-\frac{1}{2}||\hat{x}_1 - z_1||^2_{Q^{-1}_{\hat{x}_1\hat{x}_1}}\right\}}{\sum_{z_1 \in Z^p} \exp\left\{-\frac{1}{2}||\hat{x}_1 - z_1||^2_{Q^{-1}_{\hat{x}_1\hat{x}_1}}\right\}} \quad (44)$$

*and*

$$\hat{x}_{2\mathrm{BMEE}} = \hat{x}_2 - Q_{\hat{x}_2\hat{x}_1} Q^{-1}_{\hat{x}_1\hat{x}_1} (\hat{x}_1 - \hat{x}_{1\mathrm{BIEE}}) \quad (45)$$

*Proof* see Appendix. □

Thus, in the Gaussian case, the BMEP has the same structure as the BLUP, the difference being that, in the expression of the BLUP, the BLUE $\hat{x}$ gets replaced by the BMEE of $x$. Note that, since $\hat{x}$ and $t$ are independent in the Gaussian case, the BIEE (cf. Eq. 44) is now not dependent on $t$. This in contrast to the general case (see Eq. 32).

We now give some examples of the BMEP and the BMEE.

*Example 9* (GNSS ambiguity resolution) Any linear(ized) GNSS model can be cast in the following system of linear equations: $E(y) = Aa + Bb$, with $a \in Z^p$, $b \in R^{n-p}$. The data vector $y$, which is usually assumed to be normally distributed, will then consist of the 'observed minus computed' single- or multi-frequency double-difference carrier phase and/or pseudorange (code) observables accumulated over all observation epochs. The entries of the integer vector $a$ are the double differences of the carrier phase ambiguities, expressed in units of cycles, while the entries of the real-valued vector $b$ will consist of the remaining unknown parameters, such as increments of the baseline components (coordinates) and atmospheric delay parameters (troposphere, ionosphere).

It is the goal of GNSS ambiguity resolution to exploit the integerness of the ambiguity vector $a$ when estimating the parameters of interest, which are usually the components of $b$. The GNSS model is a prime example of a linear model of the mixed type. Hence, if the MSE or the variance is used as the criterion for estimation, the optimal estimator of $b$ will be given as $\hat{b}_{\mathrm{BMEE}} = \hat{b} - Q_{\hat{b}\hat{a}} Q^{-1}_{\hat{a}\hat{a}} (\hat{a} - \hat{a}_{\mathrm{BIEE}})$. The precision of this unbiased estimator will be better than the precision of the BLUE $\hat{b}$.

As was pointed out in Teunissen (2003), the expression for the $\hat{a}_{\mathrm{BIEE}}$ is identical to its Bayesian counterpart as given in Betti et al. (1993) and Gundlich and Koch (2002); also see Teunissen (2001) and Gundlich and Teunissen (2004). This equivalence nicely bridges the gap between the present nonBayesian approach and the Bayesian approach. Despite this similarity, however, there are important differences in the probabilistic evaluation of the solution, as described in Teunissen (2003).

*Example 10* (Prediction of a random vector with unknown mixed real/integer mean) Let $y = Ax' + e$, where $x'$ is a random vector with unknown mixed real/integer mean $x$, and $e$ is a zero-mean random vector that is independent of $x'$. Both $x'$ and $e$ are assumed to have a joint normal distribution. The goal is to predict $x'$ on the basis of $y$.

If we use the settings $e \mapsto A(x' - x) + e$, $y_0 \mapsto x'$, $A_0 \mapsto I$ and $e_0 \mapsto x' - x$ in the linear model defined in Eqs. (14) and (15), we have $Q_{y_0 y} = Q_{x'x'} A^{\mathrm{T}}$ and $Q_{yy} = AQ_{x'x'}A^{\mathrm{T}} + Q_{ee}$. With the use of Eq. (43), the BMEP of $x'$ then follows as $\hat{x}'_{\mathrm{BMEP}} = \hat{x}_{\mathrm{BMEE}} + Q_{x'x'} A^{\mathrm{T}} (AQ_{x'x'}A^{\mathrm{T}} + Q_{ee})^{-1} (y - A\hat{x}_{\mathrm{BMEE}})$. This representation of the BMEP is referred to as the *variance form*. Using the well-known matrix inversion lemma (e.g., Teunissen et al. 2005), the corresponding *information form* follows as $\hat{x}'_{\mathrm{BMEP}} = \hat{x}_{\mathrm{BMEE}} + (Q^{-1}_{x'x'} + A^{\mathrm{T}} Q^{-1}_{ee} A)^{-1} A^{\mathrm{T}} Q^{-1}_{ee} (y - A\hat{x}_{\mathrm{BMEE}})$. □

*Example 11* (The linear model with derived observables) In some applications, the original data vector $y$ is not used to set up the observation equations, but rather linear functions of $y$ (e.g., in the case of GNSS, the double-difference carrier phase observations rather than the undifferenced phase observations, or in the case of levelling, the observed height difference of a levelling line rather than the individual readings). This is often done to reduce the number of unknowns by eliminating so-called nuisance parameters. The linear model with derived observables has the form $B^{\mathrm{T}} y = Ax + B^{\mathrm{T}} e$, with $e$ the error component of $y$ and where $B^{\mathrm{T}} y$ is the vector of derived observables. Although one works in this set

up with $B^{\mathrm{T}}y$, one could still have the need to recover the error component of $y$ itself.

If we use the settings $y \mapsto B^{\mathrm{T}}y$, $e \mapsto B^{\mathrm{T}}e$, $y_0 \mapsto e$, $A_0 \mapsto 0$, $e_0 \mapsto e$ in the linear model defined in Eqs. (14) and (15), we have $Q_{y_0y} = Q_{ee}B$ and $Q_{yy} = B^{\mathrm{T}}Q_{ee}B$. Hence, if $y$ is normally distributed and $x$ is a mixed real/integer vector of unknown parameters, application of Eq. (43) gives the BMEP of $e$ as $\hat{e}_{\mathrm{BMEP}} = Q_{ee}B(B^{\mathrm{T}}Q_{ee}B)^{-1}(B^{\mathrm{T}}y - A\hat{x}_{\mathrm{BMEE}})$. Note that the linear model reduces to the model of condition equations when $A = 0$. In that case, $\hat{e}_{\mathrm{BMEP}}$ reduces to the BLUP of $e$.

*Example 12* (Ionospheric prediction) In Example 6 no use was made of the fact that $x_1$ is integer-valued. Hence, one can improve the predictor by using the principle of equivariance. The improved predictor, being the BMEP of $s_0$, is given as

$$\hat{s}_{0\mathrm{BMEP}} = \frac{\sigma_{s_0s}/\sigma_1^2}{1+\sigma_s^2/\sigma_1^2+\sigma_s^2/\sigma_2^2} \times \left[(y_1 - \lambda\hat{x}_{1\mathrm{BIEE}} - \hat{x}_{2\mathrm{BMEE}}) - \frac{\sigma_1^2}{\sigma_2^2}(y_2 - \hat{x}_{2\mathrm{BMEE}})\right]$$

with $\hat{x}_{2\mathrm{BMEE}} = \hat{x}_2 + \lambda\frac{2\sigma_s^2+\sigma_2^2}{4\sigma_s^2+\sigma_1^2+\sigma_2^2}(\hat{x}_1 - \hat{x}_{1\mathrm{BIEE}})$.

In the Gaussian case, we can give an explicit expression for the difference of the error variance matrices of the BLUP and the BMEP.

**Corollary 8** (BLUP and BMEP compared)

(i) Error variance*: If $y_0$ and $y$ have a joint normal distribution, then the error variance matrices of the BLUP and the BMEP are related by*

$$Q^{\mathrm{BLUP}}_{\hat{e}_0\hat{e}_0} = Q^{\mathrm{BMEP}}_{\hat{e}_0\hat{e}_0} + B_{01|y}Q_{\epsilon\epsilon}B^{\mathrm{T}}_{01|y} \tag{46}$$

*where $\epsilon = \hat{x}_1 - \hat{x}_{1\mathrm{BIEE}}$, $B_{01|y} = A_{01|y} + A_{02|y}Q_{\hat{x}_2\hat{x}_1}Q^{-1}_{\hat{x}_1\hat{x}_1}$, $A_{0|y} = (A_{01|y}, A_{02|y})$, and $A_{0|y} = A_0 - Q_{y_0y}Q^{-1}_{yy}A$.*

(ii) Mean squared error*: The MSEs of the BLUP and the BMEP are related by*

$$\mathrm{MSE(BLUP)} = \mathrm{MSE(BMEP)} + E||\hat{y}_{0\mathrm{BLUP}} - \hat{y}_{0\mathrm{BMEP}}||^2 \tag{47}$$

*Proof* see Appendix. □

Compare Corollaries 8 and 2. Equation (46) shows that the difference of the error variance matrices is driven by the difference of the BLUE and the BIEE of $x_1$. These two estimators will differ less, the less peaked the PDF of $\hat{x}_1$ is in relation to the integer grid size of $Z^n$ (see Lemma 12).

### 6.3 Weighted integer least-squares prediction

We have seen that the WLSP is a LUP and that it becomes identical to the BLUP if the weight matrix is taken as the inverse of the joint variance matrix of $y_0$ and $y$. We will now introduce the weighted integer least-squares predictor and show how it relates to the BLUP and the BMEP, respectively. We start from the same objective function $F(y, y_0, x)$ considered before (cf. Eq. 21), but now with the stipulation that $x \in Z^p \times R^{n-p}$.

**Theorem 8** (Weighted integer least-squares prediction) *The unique pair $\hat{x}_{\mathrm{WILSE}}$, $\hat{y}_{0\mathrm{WILSP}}$ satisfying $F(y, \hat{y}_{0\mathrm{WILSP}}, \hat{x}_{\mathrm{WILSE}}) \le F(y, y_0, x)$, for all $x \in Z^p \times R^{n-p}$, $y_0 \in R^{m_0}$, is said to be the* weighted integer least-squares estimator–predictor *(WILSE–WILSP) pair of $x$, $y_0$. The WILSP of $y_0$ is given as*

$$\hat{y}_{0\mathrm{WILSP}} = A_0\hat{x}_{\mathrm{WILSE}} - W^{-1}_{y_0y_0}W_{y_0y}(y - A\hat{x}_{\mathrm{WILSE}}) \tag{48}$$

*and the WILSE of $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})^{\mathrm{T}}$ is given as*

$$\hat{x}_{1\mathrm{WILSE}} = \arg\min_{z_1 \in Z^p} ||\hat{x}_{1\mathrm{WLSE}} - z_1||^2_{W_{11|2}} \tag{49}$$

*and*

$$\hat{x}_{2\mathrm{WILSE}} = \hat{x}_{2\mathrm{WLSE}} + W^{-1}_{22}W_{21}(\hat{x}_{1\mathrm{WLSE}} - \hat{x}_{1\mathrm{WILSE}}) \tag{50}$$

*respectively, where $W_{11|2} = (\bar{A}_1^{\mathrm{T}}W_{yy|y_0}\bar{A}_1)$, $\bar{A}_1 = (I - P_{A_2})A_1$, $P_{A_1} = A_1(A_1^{\mathrm{T}}W_{yy|y_0}A_1)^{-1}A_1^{\mathrm{T}}W_{yy|y_0}$, $W_{yy|y_0} = W_{yy} - W_{yy_0}W^{-1}_{y_0y_0}W_{y_0y}$, $W_{22} = A_2^{\mathrm{T}}W_{yy|y_0}A_2$ and $W_{21} = A_2^{\mathrm{T}}W_{yy|y_0}A_1$.*

*Proof* see Appendix. □

Compare Theorem 8 with Corollary 7. Note that the WILSP is a member of the class of MEPs, just like the WLSP is a member of the class of LUPs. However, unlike the WLSP, which becomes identical to the BLUP if the weight matrix is set equal to the inverse of the joint variance matrix of $y_0, y$, the WILSP does *not* become identical to the BMEP in this case. Thus, the WILSP will then still have a poorer MSE-performance than the BMEP.

For the WILSE, however, it can be shown that if $W_{11|2} = Q^{-1}_{\hat{x}_1\hat{x}_1}$, then $\hat{x}_{1\mathrm{WILSE}}$ has the highest possible probability of correct integer estimation (Teunissen 1999). The current GNSS standard for computing the integer least-squares estimator is provided by the LAMBDA-method (Teunissen 1995).

Although the WILSP differs from the BMEP in case the weight matrix is chosen as the inverse of the variance matrix, one can expect that the difference between these two predictors will get less, the more peaked the PDF of $\hat{x}$ becomes in relation to the integer grid size. Similarly, if the integer grid size gets smaller in relation to the

size and extent of the PDF of $\hat{x}$, then one can expect the difference between the BMEP and the BLUP to become smaller. This is made precise in the following lemma.

**Lemma 12** (WILSP and BLUP as limits of the BMEP) *Let $y_0, y$ have a joint normal distribution, let the weight matrix of the WILSP be equal to the inverse of their joint variance matrix, and let the variance matrix of $\hat{x}$ be factored as $Q_{\hat{x}\hat{x}} = \sigma^2 G_{\hat{x}\hat{x}}$. Then*

$$\lim_{\sigma \mapsto \infty} \hat{y}_{0\mathrm{BMEP}} = \hat{y}_{0\mathrm{BLUP}} \quad \text{and}$$
$$\lim_{\sigma \mapsto 0} \hat{y}_{0\mathrm{BMEP}} = \hat{y}_{0\mathrm{WILSP}} \tag{51}$$

*Proof* see Appendix. □

## 7 Summary and conclusions

In this contribution, we developed the theory of best prediction for linear models with mixed real-integer unknowns and showed how it compares to the existing theory of best prediction. We spoke of prediction if a function of an observable random vector $y$, say $G(y)$, is used to guess the outcome of another random, but unobservable, vector $y_0$. Prediction was called 'best' if it minimizes the mean squared error (MSE). Since one can minimize the MSE within different classes of predictors, there are different predictors that one can call 'best'.

The first three classes of predictors that were considered, are P, LP and LUP, respectively,

| | | |
|---|---|---|
| P : | $G(y)$ | (arbitrary) |
| LP : | $G(y) = L_0 y + l_0$ | (linear) |
| LUP : | $G(y) = L_0 y + l_0, L_0 A = A_0, l_0 = 0$ | (constrained linear) |

The LUP-class was defined with respect to the linear model given in Eqs. (14) and (15). The P-class is the largest and the LUP-class the smallest. The best predictors of these three classes are given as

| | |
|---|---|
| BP : | $\hat{y}_0 = E(y_0 \vert y)$ |
| BLP : | $\hat{y}_0 = E(y_0) + Q_{y_0 y} Q_{yy}^{-1}(y - E(y))$ |
| BLUP : | $\hat{y}_0 = A_0 \hat{x} + Q_{y_0 y} Q_{yy}^{-1}(y - A\hat{x})$ |

The BP requires the conditional PDF $f_{y_0|y}(y_0|y)$ and is generally a nonlinear predictor. The BP takes the form of the BLP in case $y_0$ and $y$ have a joint normal distribution. The BLP requires the first- and (central) second-order moments of $y_0, y$, whereas the BLUP only requires their (central) second-order moments. For the BLUP to be applicable, however, the unknown means of $y_0$ and $y$ need to be linked by a known linear relationship [cf. Eqs. (14) and (15)]. This is true for many applications in the spatial and Earth science disciplines. The BLUP follows from the BLP by replacing the unknown means $E(y_0)$ and $E(y)$, by their BLUEs, $A_0\hat{x}$ and $A\hat{x}$, respectively. The BLUP of $y_0 = A_0 x + e_0$ is equal to the sum of the BLUE of $E(y_0)$ and the BLUP of $e_0$.

Since LUP $\subset$ LP $\subset$ P, the minimum MSEs of their best predictors are ordered as

$$\mathrm{MSE(BP)} \leq \mathrm{MSE(BLP)} \leq \mathrm{MSE(BLUP)}$$

Since all three best predictors are unbiased, the same ordering holds true for their error variance matrices. It was pointed out that the BLUP-theorem (Theorem 3) generalizes the classical Gauss–Markov theorem of best linear unbiased estimation. If $e_0$ in $y_0 = A_0 x + e_0$ is assumed identically zero, the BLUP of $y_0$ reduces to the BLUE of $E(y_0)$ and the minimum error variance of the BLUP becomes a minimum variance of the BLUE.

Apart from the minimum MSE property and the minimum error variance property, the above three best predictors (BP, BLP, BLUP) can also be characterized by the class of data-functions that are uncorrelated with the best prediction errors. For the covariance matrix of the best prediction error and functions of the data, we have for the three cases

| | |
|---|---|
| BP : | $Q_{\hat{e}_0 H_1(\hat{x},t)} = 0,$ |
| BLP : | $Q_{\hat{e}_0 H_2(\hat{x},t)} = 0,$ |
| BLUP : | $Q_{\hat{e}_0 H_3(t)} = 0$ |

in which $H_1$ is any function of $\hat{x}, t$, $H_2$ is any linear function of $\hat{x}, t$, and $H_3$ is any linear function of only $t$. Thus, since $\hat{x}, t$ stands in a one-to-one linear relationship with the data vector $y$, the BP prediction error is uncorrelated with any function of $y$, the BLP prediction error is uncorrelated with any linear function of $y$, and the BLUP prediction error is uncorrelated with any linear function of only the vector of misclosures $t$. This also shows that the more restrictions are put on the class of predictors, the smaller the class of functions the best prediction error is uncorrelated with.

We also gave a deterministic least-squares-based interpretation of the BLUP. The weighted least-squares estimator–predictor pair of $x, y_0$ is given as

$$\hat{x}_{\mathrm{WLSE}} = (A^{\mathrm{T}} W_{yy|y_0} A)^{-1} A^{\mathrm{T}} W_{yy|y_0} y \quad \text{and}$$
$$\hat{y}_{0\mathrm{WLSP}} = A_0 \hat{x}_{\mathrm{WLSE}} - W_{y_0 y_0}^{-1} W_{y_0 y}(y - A\hat{x}_{\mathrm{WLSE}})$$

respectively, with $W$ being an arbitrary positive-definite weight matrix. The WLSP is a LUP for any choice of $W$. The WLSP becomes identical to the BLUP, if $W$ is chosen as the inverse of the joint variance matrix of $y_0$ and $y$.

For the linear model defined by Eqs. (14) and (15), the class of LUPs is a natural follow-up if one starts

from the class of LPs. That is, when one starts from the class of LPs, one has to enforce the unbiasedness condition in order to ensure that predictors are obtained that are independent of the unknown parameter vector $x$. As it was shown, however, this is not needed, and moreover, it does not do justice to the linear model in case all parameters are integer-valued, instead of real-valued. It is not needed, since instead of starting from the class of linear predictors, one can start from a larger class of predictors and one that still does justice to the equivariance that is present in the linear model.

For real-valued parameters, this has led us to the introduction of a new class of predictors, the class of equivariant predictors (EP). Predictors from this class satisfy the property $G(y + A\alpha) = G(y) + A_0\alpha$, $\forall y \in R^m, \alpha \in R^n$. Similarly, we introduced for the case of integer-valued parameters, the new class of integer equivariant predictors (IEP). Predictors from this class satisfy the property $G(y + Az) = G(y) + A_0 z$, $\forall y \in R^m, z \in Z^n$. This class is larger than the class of EPs, which in turn is larger than the class of LUPs. It was shown that the predictors from these three classes can be represented as

| | | |
|---|---|---|
| LUP : | $G(y) = A_0\hat{x} + Ht$ | (linear in $t$) |
| EP : | $G(y) = A_0\hat{x} + H(t)$ | (possibly nonlinear in $t$) |
| IEP : | $G(y) = A_0\hat{x} + H(\hat{x}, t)$ | (possibly nonlinear in $\hat{x}, t$; periodic in $\hat{x}$) |

The best predictors of these three classes are given as

$$\begin{array}{ll} \text{BLUP :} & \hat{y}_0 = A_0\hat{x} + Q_{y_0 t}Q_{tt}^{-1}t \\ \text{BEP :} & \hat{y}_0 = A_0\hat{x} + E\left(E\left(y_0 - A_0\hat{x}|\hat{x}, t\right)|t\right) \\ \text{BIEP :} & \hat{y}_0 = A_0\hat{x} + E'\left(E\left(y_0 - A_0\hat{x}|\hat{x}, t\right)|t\right) \end{array}$$

in which $E'(.|t)$ is a discretized version of $E(.|t)$. Just like the BLUE is a special case of the BLUP, the best equivariant estimator (BEE) and the best integer equivariant estimator (BIEE) are special cases of the BEP and the BIEP, respectively. Similarly to the BLUP's decomposition, the BEP resp. BIEP of $y_0 = A_0 x + e_0$ is equal to the sum of the BEE resp. the BIEE of $E(y_0)$ and the BEP resp. the BIEP of $e_0$. In case $y_0$ and $y$ have a joint normal distribution, the BEP takes the form of the BLUP and the BIEP takes a form which is similar to the BLUP, namely $\hat{y}_0 = A_0\hat{x}_{\text{BIEE}} + Q_{y_0 y}Q_{yy}^{-1}(y - A\hat{x}_{\text{BIEE}})$, with

$$\hat{x}_{\text{BIEE}} = \sum_{z \in Z^n} z \frac{\exp\left\{\frac{1}{2}||\hat{x} - z||^2_{Q^{-1}_{\hat{x}\hat{x}}}\right\}}{\sum_{z \in Z^n} \exp\left\{\frac{1}{2}||\hat{x} - z||^2_{Q^{-1}_{\hat{x}\hat{x}}}\right\}} \tag{52}$$

Since LUP $\subset$ EP $\subset$ IEP, the minimum MSEs of their best predictors are ordered as

$$\text{MSE(BIEP)} \leq \text{MSE(BEP)} \leq \text{MSE(BLUP)}$$

Since all three best predictors are unbiased, the same ordering holds true for their error variance matrices. Hence, the error variance of the BIEP is smaller than that of the BLUP and the variance of the BIEE is smaller than that of the BLUE. Apart from the minimum MSE property and the minimum error variance property, the above three best predictors can also be characterized by the class of data-functions that are uncorrelated with the best prediction errors. For the covariance matrix of the best prediction error and functions of the data, we have for the three cases,

$$\begin{array}{ll} \text{BIEP :} & Q_{\hat{e}_0 H_5(\hat{x},t)} = 0, \\ \text{BEP :} & Q_{\hat{e}_0 H_4(t)} = 0, \\ \text{BLUP :} & Q_{\hat{e}_0 H_3(t)} = 0 \end{array}$$

in which $H_3$ is any linear function of $t$, $H_4$ is any function of $t$ and $H_5$ is any function of $\hat{x}, t$, that is invariant for integer pertubations in its first slot.

The above treatment is based on the all-integer case. In most applications, however, e.g. those of GNSS and InSAR, a part of the unknown parameters will be integer-valued, while the other part will be real-valued. For this mixed integer/real parameter case, with $x = (x_1^{\text{T}}, x_2^{\text{T}})^{\text{T}}, x_1 \in Z^p, x_2 \in R^{n-p}$, the class of mixed equivariant predictors is characterized by $G(y + A_1 z + A_2\alpha) = G(y) + A_{01}z + A_{02}\alpha$, $\forall y \in R^m, z \in Z^p, \alpha \in R^{n-p}$. Such predictors can be represented as $G(y) = A_0\hat{x} + H(\hat{x}_1, t)$, for some $H$ that is invariant for integer pertubations in its first slot. The class of MEPs is smaller than the class of IEPs, but it is still larger than the class of EPs and the class of LUPs, respectively. Hence, the above given ordering of the MSEs and error variance matrices still holds true when the BIEP is replaced by the BMEP. The class of data-functions that are uncorrelated with the prediction error gets reduced, however, to functions of the form $H(\hat{x}_1, t)$.

We presented the BMEP solution for the general case and showed that, if $y_0$ and $y$ have a joint normal distribution, it reduces to

$$\hat{y}_{0\text{BMEP}} = A_0\hat{x}_{\text{BMEE}} + Q_{y_0 y}Q_{yy}^{-1}(y - A\hat{x}_{\text{BMEE}})$$

with $\hat{x}_{2\text{BMEE}} = \hat{x}_2 - Q_{\hat{x}_2\hat{x}_1}Q^{-1}_{\hat{x}_1\hat{x}_1}(\hat{x}_1 - \hat{x}_{1\text{BIEE}})$ and $\hat{x}_{1\text{BIEE}}$ given by Eq. (52), with $\hat{x}$, $z$ and $Q_{\hat{x}\hat{x}}$ replaced by $\hat{x}_1$, $z_1$ and $Q_{\hat{x}_1\hat{x}_1}$, respectively. Since integer estimators are members of the class of integer equivariant estimators, we also studied the relation between the BMEP and the integer-based WLSP. For the mixed integer/real parameter case, the WILSP is given as

$$\hat{y}_{0\text{WILSP}} = A_0\hat{x}_{\text{WILSE}} - W^{-1}_{y_0 y_0}W_{y_0 y}(y - A\hat{x}_{\text{WILSE}})$$

with $\hat{x}_{2\text{WILSE}} = \hat{x}_2 + W_{22}^{-1}W_{21}(\hat{x}_1 - \hat{x}_{1\text{WILSE}})$ and $\hat{x}_{1\text{WILSE}} = \arg\min_{z_1 \in Z^p} ||\hat{x}_1 - z||^2_{W_{11|2}}$. Thus, although the

WILSP is a member of the class of MEPs for any choice of the weight matrix, the WILSP is not identical to the BMEP in case the weight matrix is chosen equal to the inverse of the joint variance matrix of $y_0$ and $y$. This is opposed to the WLSP, which does become identical to the BLUP if the proper weight matrix is chosen. It was shown that the WILSP (with a properly chosen weight matrix) and the BLUP could be seen as two limiting cases of the BMEP. If the integer grid size gets smaller in relation to the size and extent of the PDF of $\hat{x}_1$, then the difference between the BMEP and the BLUP also gets smaller. As the other extreme, if the PDF of $\hat{x}_1$ gets more peaked (i.e. improved precision), then the difference between the BMEP and the WILSP becomes smaller.

## Appendix

*Proof of Lemma 3* (MSE decomposition) Since Eq. (6) is a direct consequence of Eq. (5), we only prove Eq. (5). With $\hat{G}(y)$ being the best predictor of class $\Omega$ and $G'(y) = \hat{G}(y) + \lambda\left(G(y) - \hat{G}(y)\right) \in \Omega$ for any $\lambda \in R$, we have $E||y_0 - G'(y)||^2_W \geq E||y_0 - \hat{G}(y)||^2_W$ and thus $E||y_0 - G'(y)||^2_W - E||y_0 - \hat{G}(y)||^2_W = -2\lambda E\big(\hat{e}_0^{\mathrm{T}} W[G(y) - \hat{G}(y)]\big) + \lambda^2 E||G(y) - \hat{G}(y)||^2_W \geq 0$ for any $\lambda \in R$. Hence, as function of $\lambda$, the function must be nonnegative for any $\lambda$. Since this is possible if and only if $E\left(\hat{e}_0^{\mathrm{T}} W[G(y) - \hat{G}(y)]\right) = 0$, the result follows. □

*Proof of Theorem 1* (Best predictor) Since both functions in the integral of $E||y_0 - G(y)||^2_W = \int E\big(||y_0 - G(y)||^2_W|y\big) f_y(y)\mathrm{d}y$ are nonnegative, the integral is minimized if $E\big(||y_0 - G(y)||^2_W|y\big)$ is minimized for every $y$. This conditional mean square can be written, with the use of the 'variance-plus-squared-bias' decomposition, as $E\big(||y_0 - G(y)||^2_W|y\big) = E\big(||y_0 - E(y_0|y)||^2_W|y\big) + ||G(y) - E(y_0|y)||^2_W$. This shows that the conditional mean square is minimized for every $y$, when $G(y)$ is chosen equal to the conditional mean $E(y_0|y)$. □

*Proof of Theorem 2* (Best linear predictor) We first write the MSE in a more convenient form. With $G(y) = L_0 y + l_0$, $\bar{y}_0 = E(y_0)$ and $\bar{y} = E(y)$, we have $E||y_0 - G(y)||^2_W = E||(y_0 - \bar{y}_0) - L_0(y - \bar{y}) + (\bar{y}_0 - L_0\bar{y} - l_0)||^2_W$, from which it follows that

$$E||y_0 - G(y)||^2_W = E||(y_0 - \bar{y}_0) - L_0(y - \bar{y})||^2_W + ||\bar{y}_0 - L_0\bar{y} - l_0||^2_W \tag{53}$$

This objective function needs to be minimized as a function of the matrix $L_0$ and the vector $l_0$. Note that the second square on the right-hand side of Eq. (53) can be made zero for any $L_0$. Hence, the optimal $l_0$ is related to the optimal $L_0$ as

$$\hat{l}_0 = \bar{y}_0 - \hat{L}_0\bar{y} \tag{54}$$

To minimize the first square on the right-hand side of Eq. (53), we recognize that $E||(y_0 - \bar{y}_0) - L_0(y - \bar{y})||^2_W = \mathrm{trace}\big([Q_{y_0y_0} - 2L_0Q_{yy_0} + L_0Q_{yy}L_0^{\mathrm{T}}]W\big)$, which can be written as

$$\begin{aligned} E||(y_0 - \bar{y}_0) - L_0(y - \bar{y})||^2_W &= \mathrm{trace}\left([Q_{y_0y_0} - Q_{y_0y}Q_{yy}^{-1}Q_{yy_0}]W\right) \\ &\quad + \mathrm{trace}\Big(\Big[L_0 - Q_{y_0y}Q_{yy}^{-1}\Big]Q_{yy} \\ &\qquad \times \Big[L_0 - Q_{y_0y}Q_{yy}^{-1}\Big]^{\mathrm{T}} W\Big) \end{aligned} \tag{55}$$

Hence, the optimal $L_0$ follows as the minimizer of the second term,

$$\hat{L}_0 = Q_{y_0y}Q_{yy}^{-1} \tag{56}$$

Substitution of Eqs. (54) and (56) into $\hat{y}_0 = \hat{L}_0 y + \hat{l}_0$ gives the result of Eq. (10). □

*Proof of Corollary 1* (BP and BLP properties)

(i) The unbiasedness of the BP follows from $E\left(E(y_0|y)\right) = E(y_0)$ and the unbiasedness of the BLP follows from applying the mean propagation law to Eq. (10).

(ii) It is not difficult to verify that the conditions of Lemma 3 are satisfied for the class of arbitrary predictors, as well as for the class of linear predictors. Hence, for the BP it follows from Eq. (5) that for any $W \geq 0$, $E(\hat{e}_{0\mathrm{BP}}^{\mathrm{T}} WH(y)) = 0$ for any $H$ and therefore $Q_{\hat{e}_{0\mathrm{BP}}H(y)} = 0$ for any function $H$. Similarly, for the BLP it follows from Eq. (5) that for any $W \geq 0$, $E(\hat{e}_{0\mathrm{BLP}}^{\mathrm{T}} WH(y)) = 0$ for any linear function $H$ and therefore $Q_{\hat{e}_{0\mathrm{BLP}}H(y)} = 0$ for any linear function $H$.

(iii) Since $\hat{y}_0$ is a nonlinear function of $y$ in case of the BP and a linear function of $y$ in case of the BLP, it follows from Eq. (11) that in both cases $Q_{\hat{e}_0\hat{y}_0} = 0$ and thus $Q_{y_0\hat{y}_0} = Q_{\hat{y}_0\hat{y}_0}$. Substitution into $Q_{\hat{e}_0\hat{e}_0} = Q_{y_0y_0} - Q_{y_0\hat{y}_0} - Q_{\hat{y}_0y_0} + Q_{\hat{y}_0\hat{y}_0}$ gives $Q_{\hat{e}_0\hat{e}_0} = Q_{y_0y_0} - Q_{\hat{y}_0\hat{y}_0}$.

(iv) Follows from a direct application of Lemma 5.

(v) Follows from an application of Lemma 3, cf Eq. (6), with $G$ as the BLP and $\hat{G}$ as the BP.

(vi) Follows from an application of Lemma 2.

(vii) Let $y_0$ be the $i$th entry of $y$. Then $y_i = E(y_i|y)$, which shows that an observable is its own BP. For the BLP, we have in that case, $Q_{y_0y}Q_{yy}^{-1} = c_i^{\mathrm{T}}$,

where $c_i$ is the canonical unit vector having a 1 as its $i$th entry and zeros otherwise. When substituted into Eq. (10), the result follows.

(viii) If $y_0$ and $y$ are independent, then $f_{y_0y}(y_0, y) = f_{y_0}(y_0)f_y(y)$ and therefore $E(y_0|y) = E(y_0)$. If $y_0$ and $y$ are uncorrelated, then $Q_{y_0y} = 0$ and the result follows from Eq. (10).

(ix) If $y_0$ and $y$ have a joint normal distribution, with means $\bar{y}_0$ and $\bar{y}$, then the conditional distribution of $y_0$ given $y$ is known to be given as $y_0|y \sim \mathrm{N}\left(E(y_0|y), Q_{y_0y_0|y}\right)$, with $E(y_0|y) = E(y_0) + Q_{y_0y} Q_{yy}^{-1}(y - E(y))$. □

*Proof of Theorem 3* (Best linear unbiased predictor) With Eq. (16), we may write the MSE as $E||y_0 - G(y)||_W^2 = E||(y_0 - A_0\hat{x}) - Ht||_W^2$. This objective function needs to be minimized as function of $H$. Note that since $E(y_0 - A_0\hat{x}) = 0$ and $E(t) = 0$, this objective function has the same structure as the function of $L_0$ in Eq. (53). Hence, in parallel with Eq. (54) and since $\hat{x}$ and $t$ are uncorrelated, the optimal $H$ follows as $\hat{H} = Q_{y_0t}Q_{tt}^{-1}$. Substitution into Eq. (16) gives the first expression of Eq. (18). To determine the second expression of Eq. (18) from its first, we note that $Q_{y_0t}Q_{tt}^{-1}t = Q_{y_0y}B(B^{\mathrm{T}}Q_{yy}B)^{-1}B^{\mathrm{T}}y$. With the use of the projector identity $Q_{yy}B(B^{\mathrm{T}}Q_{yy}B)^{-1}B^{\mathrm{T}} = I_m - A(A^{\mathrm{T}}Q_{yy}^{-1}A)^{-1}A^{\mathrm{T}}Q_{yy}^{-1}$, we obtain $Q_{y_0t}Q_{tt}^{-1}t = Q_{y_0y}Q_{yy}^{-1}(y - A\hat{x})$, which proves the second expression of Eq. (18). □

*Proof of Theorem 4* (Weighted least-squares predictor) First note that the weight matrix can be given the block-triangular decomposition

$$\begin{bmatrix} W_{yy} & W_{yy_0} \\ W_{y_0y} & W_{y_0y_0} \end{bmatrix} = \begin{bmatrix} I & W_{yy_0}W_{y_0y_0}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} W_{yy|y_0} & 0 \\ 0 & W_{y_0y_0} \end{bmatrix} \begin{bmatrix} I & 0 \\ W_{y_0y_0}^{-1}W_{y_0y} & I \end{bmatrix} \tag{57}$$

where $W_{yy|y_0} = W_{yy} - W_{yy_0}W_{y_0y_0}^{-1}W_{y_0y}$. Hence, the quadratic form in Eq. (21) can be written as a sum of two squares, $F(y, y_0, x) = ||y - Ax||_{W_{yy|y_0}}^2 + ||y_0 - A_0x + W_{y_0y_0}^{-1}W_{y_0y}(y - Ax)||_{W_{y_0y_0}}^2$. Since $\hat{x}_{\mathrm{WLSE}}, \hat{y}_{0\mathrm{WLSP}}$ set the second positive term on the right-hand side equal to zero, while $\hat{x}_{\mathrm{WLSE}}$ minimizes the first positive term, it follows that Eq. (22) is indeed the solution sought. □

*Proof of Theorem 5* (Best equivariant predictor) With Eq. (24), we have $E||y_0 - G(y)||_W^2 = E||y_0 - A_0\hat{x} - H(t)||_W^2 = E||y_0 - A_0\hat{x}||_W^2 + \Phi(H(t))$, where $\Phi(H(t)) = -2E\left((y_0 - A_0\hat{x})^{\mathrm{T}}WH(t)\right) + E||H(t)||_W^2$. The task is to find the function $H$ that minimizes $\Phi$. If we write $\Phi$ in integral form, we obtain

$$\begin{aligned}
&\Phi(H(t)) \\
&= \int_{R^{m-n}} \int_{R^n} \int_{R^{m_0}} \Big[-2(y_0 - A_0\hat{x})^{\mathrm{T}}WH(t) \\
&\qquad + ||H(t)||_W^2\Big] f_{y_0\hat{x}t}(y_0, \hat{x}, t)\mathrm{d}y_0\mathrm{d}\hat{x}\mathrm{d}t \\
&= \int_{R^{m-n}} \Bigg[ -2\Bigg( \int_{R^n} \int_{R^{m_0}} (y_0 - A_0\hat{x}) f_{y_0\hat{x}|t}(y_0, \hat{x}|t)\mathrm{d}y_0\mathrm{d}\hat{x}\Bigg)^{\mathrm{T}} \\
&\qquad \times WH(t) + ||H(t)||_W^2\Big] f_t(t)\mathrm{d}t \\
&= \int_{R^{m-n}} \Big[||E(y_0 - A_0\hat{x}|t) - H(t)||_W^2 - ||E(y_0 \\
&\qquad - A_0\hat{x}|t)||_W^2\Big] f_t(t)\mathrm{d}t
\end{aligned}$$

from which the optimal $H$ follows as $\hat{H}(t) = E(y_0 - A_0\hat{x}|t)$. Hence, the best equivariant predictor is given as $\hat{y}_0 = A_0\hat{x} + \hat{H}(t) = A_0\hat{x} + E(y_0 - A_0\hat{x}|t)$. The second expression of Eq. (25) follows from noting that $E(y_0 - A_0\hat{x}|t) = \int_{R^n}\int_{R^{m_0}}(y_0 - A_0\hat{x})f_{y_0|\hat{x}t}(y_0|\hat{x}, t)f_{\hat{x}|t}(\hat{x}|t)\mathrm{d}y_0\mathrm{d}\hat{x} = \int_{R^n} E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v, t)f_{\hat{x}|t}(v|t)\mathrm{d}v$, since $E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v, t) = \int_{R^{m_0}}(y_0 - A_0v)f_{y_0|\hat{x},t}(y_0|v, t)\mathrm{d}y_0$. □

*Proof of Corollary 4* (BEP and BLUP properties)

(i) The unbiasedness of the BEP follows from an application of the mean propagation law to Eq. (25), noting that $E(A_0\hat{x}) = A_0x = E(y_0)$ and $E\left(E(y_0 - A_0\hat{x}|t)\right) = 0$. Similarly, the unbiasedness of the BLUP follows from an application of the mean propagation law to Eq. (18), noting that $E(t) = 0$.

(ii) It is not difficult to verify that the conditions of Lemma 3 are satisfied for the class of EPs and the class of LUPs. Hence, for the BEP it follows from Eq. (5), since the difference between the BEP and any EP is an arbitrary function of $t$, that for any $W \geq 0$, $E(\hat{e}_{0\mathrm{BEP}}^{\mathrm{T}}WH(t)) = 0$ for any $H$ and therefore $Q_{\hat{e}_0H(t)}^{\mathrm{BEP}} = 0$ for any function $H$. Similarly, for the BLUP it follows from Eq. (5), since the difference between the BLUP and any LUP is an arbitrary linear function of $t$, that for any $W \geq 0$, $E(\hat{e}_{0\mathrm{BLUP}}^{\mathrm{T}}WH(t)) = 0$ for any linear function $H$ and therefore $Q_{\hat{e}_0H(t)}^{\mathrm{BLUP}} = 0$ for any linear function $H$.

(iii) The proof of Eq. (27) goes along similar lines as the one given in (iii) of Corollary 1.

(iv) Follows from a direct application of Lemma 5.

(v) Follows from an application of Lemma 3, cf Eq. (6), with $G$ as the BLUP and $\hat{G}$ as the BEP.

(vi) Follows from an application of Lemma 2.

(vii) If we replace in Eq. (18), $A_0$ by $A$ and $Q_{y_0y}$ by $Q_{yy}$, we obtain $\hat{y}_{0\mathrm{BLUP}} = y$. Similarly, if we replace $A_0$ by $A$ and $y_0$ by $y$ in Eq. (25), and recall that $y - A\hat{x} = Ct$, with $C = Q_{yy}B(B^{\mathrm{T}}Q_{yy}B)^{-1}$, we obtain, with the use of $t = E(t|t)$, that $\hat{y}_{0\mathrm{BEP}} = A\hat{x} + Ct = y$.

(viii) If $y_0$ and $\hat{x}$ are both independent of $t$, then $E(y_0 - A_0\hat{x}|t) = 0$ and therefore $\hat{y}_{0\mathrm{BEP}} = A_0\hat{x}$. If $y_0$ and $t$ are uncorrelated, then $Q_{y_0t} = 0$ and thus $\hat{y}_{0\mathrm{BLUP}} = A_0\hat{x}$.

(ix) Makes use of the fact that if two random vectors $a$ and $b$ have a joint normal distribution, then the mean of the conditional distribution of $b$ given $a$, is given as $E(b|a) = E(b) + Q_{ba}Q_{aa}^{-1}(a - E(a))$.

□

*Proof of Theorem 6* (Best integer equivariant predictor) With Eq. (30), we have $E||y_0 - G(y)||_W^2 = E||y_0 - A_0\hat{x} - H(\hat{x},t)||_W^2 = E||y_0 - A_0\hat{x}||_W^2 + \Psi(H(\hat{x},t))$, where $\Psi(H(\hat{x},t)) = -2E\left((y_0 - A_0\hat{x})^{\mathrm{T}}WH(\hat{x},t)\right) + E||H(\hat{x},t)||_W^2$. The task is to find the function $H$ that minimizes $\Psi$. If we write $\Psi$ in integral form, we obtain

$$\begin{aligned}
&\Psi(H(\hat{x},t)) \\
&= \int_{R^{m-n}} \int_{R^n} \int_{R^{m_0}} \Big[ -2(y_0 - A_0\hat{x})^{\mathrm{T}}WH(\hat{x},t) \\
&\qquad + ||H(\hat{x},t)||_W^2 \Big] f_{y_0\hat{x}t}(y_0,\hat{x},t)\mathrm{d}y_0\mathrm{d}\hat{x}\mathrm{d}t \\
&= \int_{R^{m-n}} \int_{R^n} \Big[ -2\Big( \int_{R^{m_0}} (y_0 - A_0\hat{x}) f_{y_0\hat{x}t}(y_0,\hat{x},t)\mathrm{d}y_0 \Big)^{\mathrm{T}} \\
&\qquad \times WH(\hat{x},t) + ||H(\hat{x},t)||_W^2 f_{\hat{x}t}(\hat{x},t) \Big] \mathrm{d}\hat{x}\mathrm{d}t
\end{aligned}$$

If we replace $\int_{\hat{x}\in R^n}$ by $\sum_{z\in Z^n}\int_{\hat{x}\in S_z}$, where the $S_z$'s form a partition of $R^n$, apply the change of variables $\alpha = \hat{x} - z \in S_0$ and make use of te property $H(\alpha + z, t) = H(\alpha, t)$, we can write

$$\begin{aligned}
&\Psi(H(\hat{x},t)) \\
&= \int_{R^{m-n}} \sum_{z\in Z^n} \int_{\alpha\in S_0} \Big[ -2 \Big( \int_{R^{m_0}} (y_0 - A_0(\alpha + z)) f_{y_0\hat{x}t}(y_0,\alpha+z,t)\mathrm{d}y_0 \Big)^{\mathrm{T}} \\
&\quad \times WH(\alpha,t) + ||H(\alpha,t)||_W^2 f_{\hat{x}t}(\alpha + z,t) \Big] \mathrm{d}\alpha\mathrm{d}t \\
&= \int_{R^{m-n}} \int_{\alpha\in S_0} \Big[ -2 \Big( \frac{\sum_{z\in Z^n} \int_{R^{m_0}} (y_0 - A_0(\alpha+z)) f_{y_0\hat{x}t}(y_0,\alpha+z,t)\mathrm{d}y_0}{\sum_{z\in Z^n} f_{\hat{x}t}(\alpha+z,t)} \Big)^{\mathrm{T}} \\
&\quad \times WH(\alpha,t) + ||H(\alpha,t)||_W^2 \Big] \sum_{z\in Z^n} f_{\hat{x}t}(\alpha+z,t)\mathrm{d}\alpha\mathrm{d}t \\
&= \int_{R^{m-n}} \int_{\alpha\in S_0} \Big[ ||\hat{H}(\alpha,t) - H(\alpha,t)||_W^2 - ||\hat{H}(\alpha,t)||_W^2 \Big] \\
&\quad \times \sum_{z\in Z^n} f_{\hat{x}t}(\alpha+z,t)\mathrm{d}\alpha\mathrm{d}t
\end{aligned}$$

where

$$\begin{aligned}
&\hat{H}(\alpha,t) \\
&= \frac{\sum_{z\in Z^n} \int_{R^{m_0}} (y_0 - A_0(\alpha+z)) f_{y_0\hat{x}t}(y_0,\alpha+z,t)\mathrm{d}y_0}{\sum_{z\in Z^n} f_{\hat{x}t}(\alpha+z,t)} \\
&= \frac{\sum_{v\in\alpha+Z^n} \int_{R^{m_0}} (y_0 - A_0 v) f_{y_0|\hat{x}t}(y_0|v,t)\mathrm{d}y_0 f_{\hat{x}|t}(v|t)}{\sum_{v\in\alpha+Z^n} f_{\hat{x}|t}(v|t)} \\
&= \frac{\sum_{v\in\alpha+Z^n} E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v,t) f_{\hat{x}|t}(v|t)}{\sum_{v\in\alpha+Z^n} f_{\hat{x}|t}(v|t)} \qquad (58)
\end{aligned}$$

This shows that $\Psi$ is minimized if $H$ is chosen equal to $\hat{H}$. With $\hat{y}_{0\mathrm{BIEP}} = A_0\hat{x} + \hat{H}(\hat{x},t)$, the result follows. □

*Proof of Corollary 6* (BIEP properties)

(i) The BIEP is given as $\hat{y}_0 = A_0\hat{x} + \hat{H}(\hat{x},t)$, with $\hat{H}$ given by Eq. (58). Hence, the BIEP is unbiased if we can prove that $E\left(\hat{H}(\hat{x},t)\right) = 0$. We have $E\left(\hat{H}(\hat{x},t)\right) = \int_{R^{m-n}}\int_{R^n} \hat{H}(\hat{x},t) f_{\hat{x}t}(\hat{x},t)\mathrm{d}\hat{x}\mathrm{d}t = \int_{R^{m-n}} \Big[ \sum_{z\in Z^n} \int_{\hat{x}\in S_z} \hat{H}(\hat{x},t) f_{\hat{x}t}(\hat{x},t)\mathrm{d}\hat{x} \Big]\mathrm{d}t = \int_{R^{m-n}} \Big[ \int_{\alpha\in S_0} \sum_{z\in Z^n} \hat{H}(\alpha,t) f_{\hat{x}t}(\alpha,t)\mathrm{d}\alpha \Big]\mathrm{d}t$. From Eq. (58) follows that $\sum_{z\in Z^n} \hat{H}(\alpha,t) f_{\hat{x}t}(\alpha+z,t) = \sum_{z\in Z^n} \int_{R^{m_0}} \big[y_0 - A_0(\alpha+z)\big] f_{y_0\hat{x}t}(y_0,\alpha+z,t)\mathrm{d}y_0$. Hence, we have $E\left(\hat{H}(\hat{x},t)\right) = \int_{R^{m-n}} \sum_{z\in Z^n} \int_{\alpha\in S_0} \int_{R^{m_0}} \big[y_0 - A_0(\alpha+z)\big] f_{y_0\hat{x}t}(y_0,\alpha + z,t)\mathrm{d}y_0\mathrm{d}\alpha\mathrm{d}t$ and therefore $E\left(\hat{H}(\hat{x},t)\right) = \int_{R^n}\int_{R^{m_0}} \big[y_0 - A_0\hat{x}\big] f_{y_0\hat{x}}(y_0,\hat{x})\mathrm{d}y_0\,\mathrm{d}\hat{x} = E(y_0 - A_0\hat{x}) = 0$.

(ii) It is not difficult to verify that the conditions of Lemma 3 are satisfied for the class of IEPs. Hence, for the BIEP it follows from Eq. (5), since the difference between the BIEP and any IEP is any function $H(\hat{x},t)$ for which $H(\hat{x}+z,t) = H(\hat{x},t)$ for all $z \in Z^n$, that for any $W \geq 0$, $E(\hat{e}_{0\mathrm{BIEP}}^{\mathrm{T}} WH(\hat{x},t)) = 0$ for any such $H$ and therefore $Q^{\mathrm{BIEP}}_{\hat{e}_0 H(\hat{x},t)} = 0$ for any such $H$.

(iii) The proof of Eq. (34) goes along similar lines as the one given in (iii) of Corollary 1.

(iv) Follows from a direct application of Lemma 5.

(v) Follows from an application of Lemma 3, cf Eq. (6), with $G$ as the BEP and $\hat{G}$ as the BIEP.

(vi) Follows from an application of Lemma 2.

(vii) If $y_0 = y$, then $A_0 = A$ and $f_{y_0|y}(y_0|y) = \delta(y_0 - y)$, and therefore $E_{y_0|\hat{x}t}(y_0 - A_0\hat{x}|v,t) = y - A\hat{x}$. Substitution into Eq. (31), gives $\hat{y}_0 = y$.

(viii) If $y_0$ and $y$ are independent, then $E_{y_0|\hat{x},t}(y_0 - A_0\hat{x}|v,t) = A_0(x - v)$. Substitution into Eq. (31) gives Eq. (32). □

*Proof of Corollary 7* (BMEP in Gaussian case) We will first prove the first expression of (44). According to the transformation rule for PDFs we have $f_y(y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2)) \propto f_{\hat{x}_1\hat{x}_2 t}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2, t) \propto f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2) f_t(t)$, where use has been made of the fact that $\hat{x}$ and $t$ are independent in the Gaussian case. We therefore have

$$\omega_{z_1}(y) = \frac{\int f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2)\mathrm{d}\beta_2}{\sum_{z_1}\int f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2)\mathrm{d}\beta_2} = \frac{f_{\hat{x}_1}(\hat{x}_1 + x_1 - z_1)}{\sum_{z_1} f_{\hat{x}_1}(\hat{x}_1 + x_1 - z_1)} = \frac{\exp\left\{-\frac{1}{2}||\hat{x}_1 - z_1||^2_{Q^{-1}_{\hat{x}_1\hat{x}_1}}\right\}}{\sum_{z_1}\exp\left\{-\frac{1}{2}||\hat{x}_1 - z_1||^2_{Q^{-1}_{\hat{x}_1\hat{x}_1}}\right\}}$$

since $\hat{x}_1 \sim \mathrm{N}(x_1, Q_{\hat{x}_1\hat{x}_1})$. With $\hat{x}_{1\mathrm{BIEE}} = \sum_{z_1} z_1\omega_{z_1}(y)$, the first expression of Eq. (44) follows. We now prove the second expression of Eq. (44). We have

$$\begin{aligned}\omega_{\beta_2}(y) &= \frac{\sum_{z_1} f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2)}{\sum_{z_1}\int f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2)\mathrm{d}\beta_2} \\ &= \frac{\sum_{z_1} f_{\hat{x}_1\hat{x}_2}(\hat{x}_1 + x_1 - z_1, \hat{x}_2 + x_2 - \beta_2)}{\sum_{z_1} f_{\hat{x}_1}(\hat{x}_1 + x_1 - z_1)} \\ &= \sum_{z_1} f_{\hat{x}_2|\hat{x}_1}(\hat{x}_2 + x_2 - \beta_2|\hat{x}_1 + x_1 - z_1)\omega_{z_1}(y)\end{aligned}$$

Therefore $\hat{x}_{2\mathrm{BMEE}} = \int \beta_2\omega_{\beta_2}(y)\mathrm{d}\beta_2 = \int(\hat{x}_2 + x_2 - \gamma_2)\sum_{z_1} f_{\hat{x}_2|\hat{x}_1}(\gamma_2|\hat{x}_1 + x_1 - z_1)\omega_{z_1}(y)\mathrm{d}\gamma_2 = \hat{x}_2 + x_2 - \sum_{z_1} E(\hat{x}_2|\hat{x}_1 + x_1 - z_1)\omega_{z_1}(y)$. Since $E(\hat{x}_2|\hat{x}_1) = x_2 + Q_{\hat{x}_2\hat{x}_1}Q^{-1}_{\hat{x}_1\hat{x}_1}(\hat{x}_1 - x_1)$ is the conditional mean in the Gaussian case, we have $E(\hat{x}_2|\hat{x}_1 + x_1 - z_1) = x_2 + Q_{\hat{x}_2\hat{x}_1}Q^{-1}_{\hat{x}_1\hat{x}_1}(\hat{x}_1 - z_1)$ and therefore $\hat{x}_{2\mathrm{BMEE}} = \hat{x}_2 - Q_{\hat{x}_2\hat{x}_1}Q^{-1}_{\hat{x}_1\hat{x}_1}(\hat{x}_1 - \sum_{z_1} z_1\omega_{z_1}(y))$, which proves the second expression of Eq. (44).

To prove the last expression of Eq. (44), we first write $\hat{e}_{0\mathrm{BMEP}}$ in terms of a conditional mean, $\hat{e}_{0\mathrm{BMEP}} = \sum_{z_1}\int E(e_0|y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2))\omega_{z_1\beta_2}(y)\mathrm{d}\beta_2$. Since $E(e_0|y) = Q_{y_0y}Q^{-1}_{yy}(y - A_1x_1 - A_2x_2)$ is the conditional mean in the Gaussian case, we have $E(e_0|y + A_1(x_1 - z_1) + A_2(x_2 - \beta_2)) = Q_{y_0y}Q^{-1}_{yy}(y - A_1z_1 - A_2\beta_2)$ and therefore $\hat{e}_{0\mathrm{BMEP}} = \sum_{z_1}\int Q_{y_0y}Q^{-1}_{yy}(y - A_1z_1 - A_2\beta_2)\omega_{z_1\beta_2}(y)\mathrm{d}\beta_2 = Q_{y_0y}Q^{-1}_{yy}(y - A_1\hat{x}_{1\mathrm{BIEE}} - A_2\hat{x}_{2\mathrm{BMEE}})$, which proves the last expression of Eq. (44). □

*Proof of Corollary 8* (BLUP and BMEP compared)

(i) To prove Eq. (46), we first write the prediction error of the BMEP in terms of the prediction error of the BLUP. This gives $\hat{e}_{0\mathrm{BMEP}} = \hat{e}_{0\mathrm{BLUP}} - B_{01|y}(\hat{x}_1 - \hat{x}_{1\mathrm{BIEE}})$. Since $\hat{e}_{0\mathrm{BMEP}}$ is uncorrelated with any function $H(\hat{x}_1, t)$ which is invariant for an integer pertubation in its first slot, $\hat{e}_{0\mathrm{BMEP}}$ is also uncorrelated with $\hat{x}_1 - \hat{x}_{1\mathrm{BIEE}}$. Application of the variance propagation law gives therefore $Q^{\mathrm{BMEP}}_{\hat{e}_0\hat{e}_0} = Q^{\mathrm{BLUP}}_{\hat{e}_0\hat{e}_0} - B_{01|y}Q_{\epsilon\epsilon}B^{\mathrm{T}}_{01|y}$.

(ii) Follows from an application of Lemma 3 (cf. Eq. 6), with $G$ as the BLUP and $\hat{G}$ as the BMEP. □

*Proof of Theorem 8* (Weighted integer least-squares prediction) We start by decomposing the objective function $F(y, y_0, x)$ of Eq. (21) into a sum of squares. It will be decomposed into a constant term and three variable terms. We have

$$\begin{aligned}F(y, y_0, x) &= ||y - Ax||^2_{W_{yy|y_0}} + ||y_0 - A_0x \\ &\quad + W^{-1}_{y_0y_0}W_{y_0y}(y - Ax)||^2_{W_{y_0y_0}} \\ &= ||y - A\hat{x}_{\mathrm{WLSE}}||^2_{W_{yy|y_0}} + ||\hat{x}_{\mathrm{WLSE}} - x||^2_{W_{\hat{x}\hat{x}}} \\ &\quad + ||y_0 - A_0x + W^{-1}_{y_0y_0}W_{y_0y}(y - Ax)||^2_{W_{y_0y_0}} \\ &= ||y - A\hat{x}_{\mathrm{WLSE}}||^2_{W_{yy|y_0}} + ||\hat{x}_{1\mathrm{WLSE}} - x_1||^2_{W_{11|2}} \\ &\quad + ||\hat{x}_{2\mathrm{WLSE}} - x_2 + \\ &\quad - W^{-1}_{21}W_{11}(\hat{x}_{1\mathrm{WLSE}} - x_1)||^2_{W_{22}} + ||y_0 \\ &\quad - A_0x + W^{-1}_{y_0y_0}W_{y_0y}(y - Ax)||^2_{W_{y_0y_0}} \qquad (59)\end{aligned}$$

with $W_{\hat{x}\hat{x}} = (A^{\mathrm{T}}W_{yy|y_0}A)$. Note that the last term in the third equality can be made zero for any $x \in Z^p \times R^{n-p}$ and that the before last term can be made zero for any $x_1 \in Z^p$. Hence, the sought-for minimizer is indeed given by Eqs. (50) and (48). □

*Proof of Lemma 12* (WILSP and BLUP as limits of the BMEP) Since, in the Gaussian case, the BLUP, WILSP and BMEP have the same structure, it suffices to show that the BLUE $\hat{x}_1$ and the WILSE $\hat{x}_{1\mathrm{WILSE}}$ are the corresponding limits of $\hat{x}_{1\mathrm{BIEE}}$, the proof of which is given in Teunissen (2003, p. 410). □

## References


Baarda W (1968) A testing procedure for use in geodetic networks. Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 2, No. 5, Delft, The Netherlands

Betti B, Crespi M, Sanso F (1993) A geometric illustration of ambiguity resolution in GPS theory and a Bayesian approach. Manuscr Geod 18:317–330

Bibby JM, Toutenburg H (1977) Prediction and improved estimation in linear models. Wiley, New York

Blais JAR (1982) Synthesis of Kriging estimation methods. Manuscr Geod 7:325–352

Cressie N (1991) Statistics for spatial data. Wiley, New York

Dermanis A (1980) Adjustment of geodetic observations in the presence of signals. Bollettino di Geodesia e Scienze Affini 38:419–445

Eeg J, Krarup T (1973) Integrated geodesy. Dan. Geod. Inst., int. rep. 7, Copenhagen
Gandin LS (1963) Objective analysis of meteorological fields. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad
Grafarend EW (1976) Geodetic applications of stochastic processes. Phys Earth Planet Interiors 12:151–179
Grafarend EW, Rapp RH (eds)(1980) Advances in geodesy. Selected papers from Rev Geophys Space Phys, Richmond, Virg., Am Geophys Union, Washington
Gundlich B, Koch KR (2002) Confidence regions for GPS baselines by Bayesian statistics. J Geod 76:55–62
Gundlich B, Teunissen PJG (2004) Multiple models: fixed, switching and interacting. In: Proceedings V Hotine-Marussi Symposium 2002, Matera, Italy, Band 127, Reihe: International Association of Geodesy Symposia, 2004
Hein GW (1986) Integrated geodesy. In: Suenkel H (ed) Mathematical and numerical techniques in physical geodesy. Lecture Notes in Earth Sciences, Vol 7. Springer, Berlin, pp 505–548
Journel AG, Huijbregts ChJ (1991) Mining geostatistics. Academic, New York
Koch KR (1980) Parameterschaetzung und Hypothesetests in linearen Modellen. Dummler, Bonn
Kolmogorov AN (1941) Interpolation and extrapolation of stationary random sequences. Izvestiia Akademii Nauk SSSR, Seriia Matematicheskiia 5:3–14
Krarup T (1969) A contribution to the mathematical foundation of physical geodesy. Publ. Danish Geod. Inst. 44, Copenhagen
Krarup T (1980) Integrated geodesy. Bollettino di Geodesia e Scienze Affini 38:480–496
Krige D (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. J Chem Metall Min Soc S Afr 52:119–139
Matheron G (1970) The theory of regionalized variables and its applications. Fascicule 5, Les Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau, 211p
Moritz H (1973) Least-squares collocation. Deutsche Geodaetische Kommission, Reihe A, No. 59, Muenchen
Moritz H (1980) Advanced physical geodesy. Wichmann, Karlsruhe
Moritz H, Suenkel H (Eds) (1978) Approximation methods in geodesy. Sammlung Wichmann Neue Folge, Band 10, Wichmann, Karlsruhe
Rao CR, Toutenburg H (1999) Linear models: least squares and alternatives. Springer Series in Statistics, Berlin
Reguzzoni M, Sanso F, Venuti G (2005) The theory of general kriging, with applications to the determination of a local geoid. Geophys J Int 162:303–314
Rummel R (1976) A model comparison in least-squares collocation. Bull Geod 50:181–192
Sanso F (1980) The minimum mean square estimation principle in physical geodesy. Bollettino di Geodesia e Scienze Affini 39(2):111–129
Sanso F (1986) Statistical methods in physical geodesy. In: Suenkel H (ed) Mathematical and numerical techniques in physical geodesy, Lecture Notes in Earth Sciences, Vol. 7. Springer, Berlin, pp 49–156
Stark H (Ed.) (1987) Image recovery: theory and application. Academic, New York
Teunissen PJG (1995) The least-squares ambiguity decorrelation adjustment: a method for fast GPS integer ambiguity estimation. J Geod 70:65–82
Teunissen PJG (1999) An optimality property of the integer least-squares estimator. J Geod 73:587–593
Teunissen PJG (2001) Statistical GNSS carrier phase ambiguity resolution: a review. In: Proceedings IEEE Symposium Statistical Signal Processing, pp 4–12
Teunissen PJG (2003) Theory of integer equivariant estimation with application to GNSS. J Geod 77:402–410
Teunissen PJG (2006) Least-squares prediction in linear models with integer unknowns. J Geodesy (in press). DOI 10.1007/s00190-007-0138-0
Teunissen PJG, Simons DG, Tiberius CCJM (2005) Probability and Observation Theory. Lectures Notes Delft University of Technology, Delft, The Netherlands
Tscherning CC (1978) Collocation and least-squares methods as a tool for handling gravity field dependent data obtained through space research techniques. Bull Geod 52:199–212
Wackernagel H (1995) Multivariate geostatistics. Springer, Berlin
Wiener N (1949) Extrapolation, interpolation, and smoothing of stationary time series. MIT, Cambridge