Reports on Geodesy and Geoinformatics, 2024, Vol. 118, pp. 101–108

DOI: 10.2478/rgg-2024-0019
Received: 20 July 2024 / Accepted: 5 November 2024
Published online: 29 November 2024

ORIGINAL ARTICLE

Proposing a concept of least-squares-based outlier-exposing potential of Gauss-Markov models: Examples in geodesy

Witold Prószyński [1*] and Zenon Parzyński [1]

[1]Faculty of Geodesy and Cartography, Warsaw University of Technology, Pl. Politechniki 1, 00-661, Warsaw, Poland

*witold.proszynski@pw.edu.pl

Abstract

Outlier detection and identification are still important issues in the quality control of geodetic networks based on least squares estimation (LSE). In addition to existing network reliability measures, the paper proposes the LSE-based concept (together with the associated measures) of the Outlier-Exposing Potential (OEP) for Gauss-Markov models. The greater the model's redundancy, the more the configuration of its responses to gross errors exposes the location of these errors, and hence, the greater the model's OEP. The potential is given in the basic version and the extended version. The former considers only the effect of the model's redundancy, while the latter also considers the masking effect due to random observation errors at a specified magnitude of gross error. For models that have regions of unidentifiable errors, the corresponding OEP components have zero values. The reflection of OEP in the values of Minimal Identifiable Bias (MIB) is shown. It is proposed that OEP derived based on least squares adjustment be treated as a property of the model itself. The theory is illustrated on several 1D and 2D networks. The research is limited to models with uncorrelated observations and the case of a single gross error. These limitations enabled the formulation of clear properties of general character, not complicated by observation correlations and multiple-outlier combinations.

Key words: least-squares adjustment, uncorrelated observations, redundancy matrix, disturbance/response relationship, response concentration index

1 Introduction

Gauss-Markov Models (GMMs) represent theoretical models for a wide class of observation systems used in geodesy. As sensitive to outliers, the LSE is not classified as a robust estimation (Huber, 1972; Maronna et al., 2006). However, when applied in an iterative mode for the GMMs with a particularly high level of redundancy, it may, by successive removal of outliers, yield parameter estimates almost cleared from the influence of gross errors. The use of the term 'almost' is here necessary since, according to (Teunissen et al., 2017), the bias in the solution can never be removed completely.

We would like to point out that following the understanding of the notion of an outlier as in Lehmann (2013), we will treat the outlier as an observation contaminated by a gross error.

The model's redundancy causes that in parameter determination the observations become interrelated, and thanks to it, they may check themselves mutually. This enables detection and identification of outliers. It is redundancy that is an essential element in the Baarda reliability theory for quality control of geodetic networks (Baarda, 1968) and in its further developments (e.g., Teunissen 2006, 2018; Rofatto et al. 2020; Lehmann et al. 2020). Baarda's theory shows that the greater the redundancy of a GMM, the greater the concentration of its LS-induced responses to gross errors in their location or equivalently, the smaller the smearing effect of LSE (Durdag et al., 2022). This makes detecting and identifying outliers easier and more effective.

Publisher: De Gruyter

Getting to know and characterizing the abovementioned ability of GMMs of response concentration and thus the exposure of outliers, confirmed in practice, would require careful analysis. Undertaking such research would be primarily important from the point of view of quality control theory for geodetic observation systems modeled by GMMs. Some advantages for practical applications would also be expected from the research.

Following the above considerations, the objectives of the paper are:

- to characterize and quantify the redundancy-dependent property of exposing the outliers in LS adjustment in GMMs, considering the model's redundancy impact only,
- to characterize and quantify the property as above, considering both the impact of the model's redundancy and that of random observation errors,
- to investigate the possible reflection of OEP level in MIB values.

We shall assume the case of a single gross error and uncorrelated observations. The case of multiple gross errors would complicate the formulation of clear general properties. As for the second assumption, the research was originally intended to include models with correlated observations. However, the feasibility analysis of such a broadly outlined task led to confinement to models with uncorrelated observations. The main arguments, apart from the one in the first assumption, were that:

- due to observation correlations the redundancy matrix in a standardized model loses its valour of being the covariance matrix of LS residuals;
- the correlation matrices bring into models a considerable variety of information characterizing the measurement process, but being different from the original models' features.

2 Properties of the redundancy matrix for a standardized Gauss-Markov Model

The redundancy matrix for a standardized GMM is a main element of the theoretical basis of the present research and that is why we present its properties.

Let us consider a GMM with uncorrelated observations:

$$\mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{y}; \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{C}) \tag{1a}$$

where:
$\mathbf{y}$ – the $n \times 1$ vector of observations,
$\mathbf{A}$ – the $n \times u$ design matrix; rank $\mathbf{A} = u - d$ (d – defect, $d \geq 0$),
$\mathbf{x}$ – the $u \times 1$ vector of unknown parameters,
$\mathbf{e}$ – the $n \times 1$ vector of random errors,
$\mathbf{C}$ – the $n \times n$ covariance matrix for observations; diagonal, pos. definite.

With the standardization matrix $\mathbf{S} = \mathbf{C}^{-\frac{1}{2}}$, we get the equivalent standardized form of (1a):

$$\mathbf{A}_S\mathbf{x} + \mathbf{e}_S = \mathbf{y}_S; \qquad \mathbf{e}_S \sim N(\mathbf{0}, \mathbf{I}) \tag{1b}$$

where, e.g., $\mathbf{A}_S = \mathbf{C}^{-\frac{1}{2}}\mathbf{A}$.

The LS estimator of the vector $\mathbf{v}_S = -\mathbf{e}_S$, being the vector of LS residuals in the model (1b), is given by (Prószyński, 1994):

$$\hat{\mathbf{v}}_S = -\mathbf{R}_S \cdot \mathbf{y}_S \tag{2}$$

where:

$$\mathbf{R}_S = \mathbf{I} - \mathbf{A}_S(\mathbf{A}_S^T\mathbf{A}_S)^{+}\mathbf{A}_S^T \tag{3}$$

$(*)^{+}$denotes a pseudoinverse, for $d = 0$ it becomes a regular inverse.

As an idempotent and symmetric matrix, $\mathbf{R}_S$ is an orthogonal projector; rank $\mathbf{R}_S = \mathrm{Tr}\,\mathbf{R}_S = n - u + d = f$, where f is a model redundancy. $\mathbf{R}_S$ is a redundancy matrix for the standardized GMM (1b).

It readily be shown that $\mathbf{R}_S = \mathbf{C}_{\hat{\mathbf{v}}_S} = \mathbf{I} - \mathbf{C}_{\hat{\mathbf{y}}_S}$, where $\mathbf{C}_{\hat{\mathbf{y}}_S}$ and $\mathbf{C}_{\hat{\mathbf{v}}_S}$ are the covariance matrices for the vector of estimated standardized observations and the vector of residuals in a standardized model, respectively.

Let us note that the commonly used redundancy matrix $\mathbf{R} = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{+}\mathbf{A}^T\mathbf{C}^{-1}$ (for the original model (1a), e.g., Ding and Coleman 1996; Durdag et al. 2022) is not symmetric and is not a covariance matrix like $\mathbf{R}_S$. Since $\mathbf{R}_S = \mathbf{S}\mathbf{R}\mathbf{S}^{-1}$, $\mathbf{R}_S$ and $\mathbf{R}$ are similar matrices with the corresponding diagonal elements being equal.

On the basis of (2) we immediately get a *disturbance/response* relationship for the model (1b), i.e.,

$$\Delta\hat{\mathbf{v}}_S = -\mathbf{R}_S \cdot \Delta\mathbf{y}_S \tag{4}$$

where:
$\Delta\mathbf{y}_S$ – the vector of standardized gross errors (i.e., disturbances),
$\Delta\hat{\mathbf{v}}_S$ – the vector of changes in LS residuals (i.e., responses).

The value ranges of the elements of the matrix $\mathbf{R}_S$ are as follows (Chatterjee and Hadi, 2009; Prószyński, 2012):

$$r_{ii},\ i = 1, \ldots, n,\ 0 \leq r_{ii} < 1$$
$$r_{ij},\ i, j = 1, \ldots, n,\ j \neq i,\ 0 \leq \left|r_{ij}\right| \leq 0.5$$
$$r_{ii} = 0 \text{ implies } r_{ik} = 0, \text{ where } k = 1, \ldots, n, k \neq i.$$

The diagonal elements r_{ii} of the matrix $\mathbf{R}_S$, abbreviated as $r_i (i = 1, \ldots, n)$, are partial redundancies, also termed the redundancy contributions of individual observations (Ding and Coleman, 1996). We will simply call them redundancy indices.

The model's average redundancy index $\bar{r}$ is the following:

$$\bar{r} = \frac{\sum_n r_i}{n} = \frac{n - (u - d)}{n} = \frac{f}{n}. \tag{5}$$

Denoting by n_0 the number of observations for which $f = 1$ for a given model (1a), we get from the formula (5):

$$\bar{r}_0 = \frac{1}{n_0} = \frac{1}{u - d + 1}, \qquad 0 < \bar{r}_0 \leq 0.5 \tag{6}$$

where $\bar{r}_0$ is a minimal non-zero value of $\bar{r}$ for a given model. Hence, the range of non-zero values of $\bar{r}$ for models can be defined as $[\bar{r}_0, 1)$. We shall call $\bar{r}_0$ a model minimal average redundancy.

The non-diagonal elements r_{ij} are the covariances between the LS-estimated i-th and j-th standardized observation. With $r_i > 0.5$, the increase in the values of r_i implies the decrease in r_{ij} $(j = 1, \ldots, n); j \neq i)$ in terms of the absolute value.

Proof. For $\mathbf{R}_S$ as an orthogonal projector, we can write $r_i^2 + \sum_{j=1, j\neq i}^{n} r_{ij}^2 = r_i$, and so, $\sum_{j=1, j\neq i}^{n} r_{ij}^2 = r_i - r_i^2$, which finally yields $\left|r_{ij}\right|_{max} < \sqrt{r_i - r_i^2}$ $(j = 1, \ldots, n; j \neq i)$. For $r_i > 0.5$, it is a decreasing function. □

From the relationship (4) it follows immediately that with $r_{ij} = 0$ for a particular pair of the observations such that $r_i \neq 0$ and $r_j \neq 0$, the disturbance in the i-th one will not induce any response in the j-th one.

There are models whose matrix $\mathbf{R}_S$ has some or even all rows (and therefore columns) linearly dependent, which makes outlier identification impossible. The observations corresponding to these rows were called the Region of Unidentifiable Errors –

RUE (Prószyński, 2008). Since, in general, there can be several RUEs in a model, we give a definition of a single RUE being a set of q observations.

Definition:

$$\mathrm{RUE}_{(q)} = \Big\{ y_i, y_j, \ldots, y_{m_{(q)}} : \ |r_{ij}| = \sqrt{r_i \cdot r_j}, \ldots, \\ |r_{im}| = \sqrt{r_i \cdot r_m}, \ldots, |r_{jm}| = \sqrt{r_j \cdot r_m}_{(q(q-1)/2)} \Big\} \tag{7}$$

where $r_i, r_j, r_m, r_{ij}, r_{im}, r_{jm}$ are the corresponding elements of the matrix $\mathbf{R}_s$, $q(q-1)/2$ is a number of unidentifiability conditions.

The unidentifiability condition for a pair of observations can be transformed into the following equivalent form:

$$|r_{ij}| = \sqrt{r_i \cdot r_j} \equiv \frac{|r_{ij}|}{\sqrt{r_i} \cdot \sqrt{r_j}} = 1 \equiv \left| \rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right) \right| = 1 \tag{8}$$

which means that the LS residuals for the i-th and j-th observation in the standardized model should be linearly dependent.

We quote below a condition for non-existence of RUEs:

- in a model with a redundancy level such that $r_i > 0.5$ ($i = 1, \ldots, n$), there cannot exist any RUE.

The condition results directly from the maximal absolute value of non-diagonal elements of $\mathbf{R}_s$ being 0.5.

It is easy to prove that for the q observations constituting a particular $\mathrm{RUE}_{(q)}$, all the Baarda w-statistics, i.e.,

$$w_i = \frac{\hat{v}_{s,i}}{\sigma(\hat{v}_{s,i})} = -\frac{\{\mathbf{R}_s\}_{i\bullet}}{\sqrt{r_i}} \cdot \mathbf{y}_s, \qquad i = 1, \ldots, q \tag{9}$$

where $\{\mathbf{R}_s\}_{i\bullet}$ – the corresponding row of the matrix $\mathbf{R}_s$, will have identical absolute values.

The property holds for any set of observation standard deviations that can be assumed in a model (1a). This means that the existence of RUEs in a model (1a) depends only on the structure of its design matrix $\mathbf{A}$.

We can distinguish a Global RUE ($q = n$), covering all the observations in a model, and a Local RUE ($2 \leq q < n$) covering some observations in a model. There can be several Local RUEs in a model. A model with $\mathbf{R}_s$ of rank equal to 1 (i.e., $f = 1$), always forms a Global RUE, e.g., a leveling loop. In such models, outliers can be detected depending on their magnitude but cannot be identified.

The above properties of RUEs lead to the following conclusion regarding the identifiability of outliers and determining Minimal Identifiable Biases (MIBs) (Teunissen, 2018; Imparato et al., 2018) or Identifiability Indices (IDs) (Prószyński, 2015):

- an outlier y_k(with $r_k \neq 0$) is identifiable, and hence MIB_k(or ID_k) is determinable if $f \geq 2$ and $y_k \notin$ RUE.

3 Proposal of a concept of outlier-exposing potential of a GMM

The LSE based disturbance/response relationship shows that the greater the redundancy of the model, the greater the concentration of model responses in the location of observation disturbance. According to this property, it becomes easier and easier to identify observations contaminated by gross errors, i.e., outliers. By successive removal of dominating outliers one may finally determine the LS estimates of the model parameters comparable in terms of correctness to those obtained in robust estimations, provided random errors are normally distributed.

Therefore, we have both theoretical and empirical grounds to conclude that with a sufficiently high level of model redundancy, LSE demonstrates in terms of standardized residuals an advantageous behavior in the presence of outliers. So, we propose terming this specific LS-based property of a GMM *the outlier-exposing potential.*

In contrast to robust estimation procedures, LSE does not introduce observation weighting other than that being an integral element of a given GMM. Let us also note that the redundancy matrix $\mathbf{R}_s$ that generates the LS model responses depends only on the standardized design matrix $\mathbf{A}_s$ (see formula (3), which determines the model's redundancy. So, the advantageous concentration of model responses mentioned above can be regarded as associated with the original model (1a). The above statements lead to the conclusion that the outlier-exposing potential displayed by LSE can also be considered a redundancy-resulting property of the model itself.

We will finally propose the following descriptive definition: *Outlier-Exposing Potential (OEP) of a GM model* – the ability of GMM, conditioned by redundancy level, to yield LS responses to a gross error so that the response dominating in terms of size is in the contaminated observation, thereby exposing an outlier.

A precise definition will be contained in a corresponding measure given further on. The presentation of the measure will be preceded by deriving an auxiliary index based on the changes in the outlier test statistics (Δw) induced by a single gross error in a model.

The index denoted as RC_{ij} and termed *a response concentration index*, referring to the i-th observation contaminated by a gross error and the j-th not contaminated observation, is defined by:

$$\mathrm{RC}_{ij} = \frac{|\Delta \mathrm{w}_{i(i)}| - |\Delta \mathrm{w}_{j(i)}|}{|\Delta \mathrm{w}_{i(i)}|} = 1 - \frac{|\Delta \mathrm{w}_{j(i)}|}{|\Delta \mathrm{w}_{i(i)}|} = 1 - \left| \rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right) \right| \tag{10}$$

where $\rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right)$ is the coefficient of correlation between the LS residuals in the standardized model (1b), RC_{ij} takes values within the interval [0, 1].

Since $\rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right) = \rho\left(\mathrm{w}_{i(i)}, \mathrm{w}_{j(i)}\right)$, the formula (10) is consistent with a well-known property (Förstner, 1983). The property says that in addition to the probabilities of I type, and II type errors (not applicable in the proposed approach), the correlation between the outlier test statistics is decisive for identifying the contaminated i-th observation.

Derivation of the formula (10):

On the basis of (9), we get $\Delta \mathrm{w}_{i(i)} = \frac{\Delta \hat{v}_{s,i(i)}}{\sqrt{r_i}}$ and $\Delta \mathrm{w}_{j(i)} = \frac{\Delta \hat{v}_{s,j(i)}}{\sqrt{r_j}}$.

Assuming a gross error in the i-th observation ($\Delta y_{s,i}$) and using (4), we can write $\Delta \hat{v}_{s,i(i)} = -r_i \cdot \Delta y_{s,i}$ and $\Delta \hat{v}_{s,j(i)} = -r_{ij} \cdot \Delta y_{s,i}$.

Hence, we obtain:

$$1 - \frac{|\Delta \mathrm{w}_{j(i)}|}{|\Delta \mathrm{w}_{i(i)}|} = 1 - \frac{|r_{ij}| \cdot \Delta y_{s,i}}{\sqrt{r_j}\sqrt{r_i} \cdot \Delta y_{s,i}} \\ = 1 - \frac{|r_{ij}|}{\sqrt{r_i}\sqrt{r_j}} = 1 - \left| \rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right) \right|. \quad \square \tag{11}$$

From semi-positive definiteness of the matrix $\mathbf{R}_s$ it follows that $|\Delta \mathrm{w}_{i(i)}| \geq |\Delta \mathrm{w}_{j(i)}|$. The more the value of $|\Delta \mathrm{w}_{i(i)}|$ exceeds the value of $|\Delta \mathrm{w}_{j(i)}|$, the greater is RC_{ij}, i.e., the concentration of the model's response to a gross error at its location. For the i-th and j-th observation belonging to RUE (i.e., unidentifiable), we have RC_{ij}= 0.

Based on the RC_{ij} indices, we establish a measure for OEP:

$$\mathrm{OEP} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} RC_{ij} \tag{12}$$

which is an average value of the RC_{ij} index in a model.

The RC_{ij} indices are non-diagonal elements of a matrix $\mathbf{K}(n \times n)$ defined by $\mathbf{K} = \mathbf{1}_n - |\mathrm{Cor}(\mathbf{R}_S)|$, where: $\mathrm{Cor}(\mathbf{R}_S)$ – a correlation matrix formed based on the redundancy matrix $\mathbf{R}_S$, $\mathrm{Cor}(\mathbf{R}_S) = (\mathrm{diag}\,\mathbf{R}_S)^{-1/2} \cdot \mathbf{R}_S \cdot (\mathrm{diag}\,\mathbf{R}_S)^{-1/2}$, $\mathbf{1}_n$ – the $n \times n$ matrix of ones, $|\mathrm{Cor}(\mathbf{R}_S)|$ – a matrix with elements being the absolute values of the elements of $\mathrm{Cor}(\mathbf{R}_S)$.

The matrix $\mathrm{Cor}(\mathbf{R}_S)$ contains complete information about the model's response to gross error in each observation. It indicates pairs of observations for which there is no disturbance transmission, i.e., $RC_{ij} = 1$, and pairs belonging to RUE, i.e., $RC_{ij} = 0$.

With the appropriate summation order in the formula (12), one may indirectly determine the OEP component values (denoted as OEP_i) for each of the n observations, i.e.,

$$\mathrm{OEP}_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} RC_{ij}, \qquad i = 1, \ldots, n. \tag{13}$$

OEP for a given GMM could be represented as, for example, a set of the following three elements:

$$\{\mathrm{OEP}, \mathrm{OEP}_{\min}, \mathrm{OEP}_{\max}\} \tag{14}$$

where $\mathrm{OEP}_{\min}$ and $\mathrm{OEP}_{\max}$ are the extreme values of OEP_i.

For a model being a Global RUE (e.g., a leveling loop), OEP = 0. It is easy to check that for the model (1a) with $\mathbf{A}(n \times 1) = [1 1 \ldots 1]^{\mathrm{T}}$ and $\mathbf{C}(n \times n) = \sigma^2 \mathbf{I}$ $(n \geq 2)$, $\mathrm{OEP} = \mathrm{OEP}_{\min} = \mathrm{OEP}_{\max}$ will be the following:

$$\mathrm{OEP} = \frac{n-2}{n-1} \tag{15}$$

and since $\bar{r} = 1 - n^{-1}$, we get for this model:

$$\mathrm{OEP} = 2 - \frac{1}{\bar{r}}, \qquad \bar{r}_0 = 0.5. \tag{16}$$

Properties of RC_{ij} *and OEP*:

- as it is shown in the derivation of the formula (10), RC_{ij} and hence OEP do not depend on the magnitude of the gross error $\Delta y_{s,i}$;
- the smaller the correlation between the i-th and j-th LS residual in the standardized model, the greater the RC_{ij}. It is known that the correlation is the smaller, the more remote (in terms of the connections in a network) is the j-th observation from the i-th one;
- with $r_{ij} = 0$ where $r_i \neq 0$ and $r_j \neq 0$, we have $\rho\left(\hat{v}_{s,i}, \hat{v}_{s,j}\right) = 0$ and hence $RC_{ij} = 1$;
- for a pair of observations where $r_j = r_i$, we can modify (11) as shown below:

$$1 - \frac{\left|\Delta w_{j(i)}\right|}{\left|\Delta w_{i(i)}\right|} = 1 - \frac{\left|r_{ij}\right|}{\sqrt{r_i}\sqrt{r_i}} = 1 - \frac{\left|r_{ij}\right|}{r_i} = 1 - \frac{\left|\Delta \hat{v}_{s,j(i)}\right|}{\left|\Delta \hat{v}_{s,i(i)}\right|}$$

  which means that RC_{ij} for such a pair can equivalently be determined using increments in the LS residuals in the standardized model (1b);
- an increase in the model's redundancy due to adding some observations results in an increase in OEP. For $r_i > 0.5$, with the increase in n, r_i values increase while r_{ij} values decrease, and this causes an increase in OEP;
- from the formula (15) it follows that with $n \to \infty$, OEP $\to 1$. The range of OEP values is thus $[0, 1)$;
- since for $f = 1$, OEP = 0, the "0" level of the OEP for a given model falls at $\bar{r}_0$ (see (6));
- even though OEP for a given model increases with increasing $\bar{r}$, due to the different positions of the OEP's "0" on the $\bar{r}$ axis for various models (see (6)), it is not possible to formulate a general dependence of OEP on $\bar{r}$ covering all the models;
- changing standard deviations of the observations, which is a modification of the model, implies the change in OEP value;
- OEP is invariant to equal changes in all the σ values since this does not affect the redundancy matrix $\mathbf{R}_S$;
- it is difficult to establish a strictly justifiable lower acceptable limit for the OEP values. The definition of OEP shows that the higher the OEP value, the more the distribution of the model's responses to gross error (without taking into account the effect of random errors) exposes the location of the infected observation. Since at OEP = 0.5, the response at the location of the gross error is on average twice as large as the response in any other observation, this value could be considered a lower acceptable limit. Then OEP > 0.5 would be treated as a required level of the potential. With very high OEP values, e.g., OEP > 0.9, it could be expected that the response distribution may expose 2 or more gross errors (especially those located remotely from each other).

4 Extending the RC_{ij} and OEP measures to cover the impact of observation errors

The RC_{ij} definition as in (10) will be replaced by that below:

$$RC_{ij}^{(r)} = \frac{\left|w_{i(i)}\right| - \left|w_{j(i)}\right|}{\left|w_{i(i)}\right|} = 1 - \frac{\left|w_{j(i)}\right|}{\left|w_{i(i)}\right|}. \tag{17}$$

On the basis of (9), we get:

$$\begin{aligned} w_{i(i)} &= -\left(\frac{\{\mathbf{R}\}_{i\bullet}}{\sqrt{r_i}} \cdot \mathbf{e}_S + \sqrt{r_i} \cdot \Delta y_{s,i}\right), \\ w_{j(i)} &= -\left(\frac{\{\mathbf{R}\}_{j\bullet}}{\sqrt{r_j}} \cdot \mathbf{e}_S + \frac{r_{ij}}{\sqrt{r_j}} \cdot \Delta y_{s,i}\right), \end{aligned} \tag{18}$$

where $\mathbf{e}_S$ is a vector of standardized observation errors; $\mathbf{e}_S \sim N(0, 1)$, $\Delta y_{s,i}$ is a standardized gross error in the i-th observation, and hence:

$$RC_{ij}^{(r)} = 1 - \frac{\left|\frac{\{\mathbf{R}\}_{j\bullet}}{\sqrt{r_j}} \cdot \mathbf{e}_S + \frac{r_{ij}}{\sqrt{r_j}} \Delta y_{s,i}\right|}{\left|\frac{\{\mathbf{R}\}_{i\bullet}}{\sqrt{r_i}} \cdot \mathbf{e}_S + \sqrt{r_i} \cdot \Delta y_{s,i}\right|}, \tag{19}$$

where the resulting index $RC_{ij}^{(r)}$ is a random variable, with probability distribution difficult to establish.

To have an unambiguous measure for model analyses, representative for all possible vectors $\mathbf{e}_S$, we will assume the expectation of this variable, i.e., $E(RC_{ij}^{(r)})$, determined empirically.

To determine $E(RC_{ij}^{(r)})$ empirically we have to realize the formula (19) using a sufficiently large number of numerically simulated vectors $\mathbf{e}_S$ (10000 simulations were applied). As a final result, we take an arithmetic mean of all the values obtained for individual vectors $\mathbf{e}_S$. Also, empirical variance, i.e., $\mathrm{var}(RC_{ij}^{(r)})$, can be determined. We estimate the precision of indices for ± 0.01.

To avoid complicating the notation, we will further omit the expectation value symbol. Although the $\Delta y_{s,i}$ symbol does not appear in a general notation $RC_{ij}^{(r)}$, it will be visible in a specific notation $RC_{ij}^{(r)}(\Delta y_{s,i} = a)$.

We can see that with $\mathbf{e}_s = \mathbf{0}$, $RC_{ij}^{(r)}$ for any value of $\Delta y_{s,i}$ becomes RC_{ij}. Like in the case of RC_{ij} and OEP (see (12)), on the basis of $RC_{ij}^{(r)}$ one can form an outlier-exposing potential for a whole network, denoted by $OEP^{(r)}$ and being an extended version of OEP:

$$OEP^{(r)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} RC_{ij}^{(r)}, \tag{20}$$

where both all $RC_{ij}^{(r)}$ and $OEP^{(r)}$ are determined for a specified value of $\Delta y_{s,i}$.

Analogously to (13) and (14), we shall write:

$$OEP_i^{(r)} = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \left(RC_{ij}^{(r)} \right), \qquad i = 1, \ldots, n \tag{21}$$

$$\left\{ OEP^{(r)}, OEP_{min}^{(r)}, OEP_{max}^{(r)} \right\} \tag{22}$$

where all the quantities are based on the same specified value of $\Delta y_{s,i}$.

Properties of $RC_{ij}^{(r)}$ and $OEP^{(r)}$ obtained empirically:

- $RC_{ij}^{(r)}(\Delta y_{s,i} = a) = RC_{ij}^{(r)}(\Delta y_{s,i} = -a)$;
- for some size of gross error $\Delta y_{s,i}$, $RC_{ij}^{(r)}$ may take negative values;
- with increasing $\Delta y_{s,i}$, $RC_{ij}^{(r)}$ tends asymptotically to RC_{ij};
- if the i-th and j-th observations belong to RUE, then in addition to $RC_{ij} = 0$ (as was already proved) there is also $RC_{ij}^{(r)} = 0$, regardless of the size of $\Delta y_{s,i}$;
- $OEP^{(r)}(\Delta y_{s,i} = a) = OEP^{(r)}(\Delta y_{s,i} = -a)$;
- for some size of gross error $\Delta y_{s,i}$, $OEP^{(r)}$ may take negative values;
- with increasing $\Delta y_{s,i}$, $OEP^{(r)}$ tends asymptotically to OEP;
- for the i-th and j-th observation not belonging to RUE, there is a relationship $RC_{ij}^{(r)} < RC_{ij}$, regardless of the size of $\Delta y_{s,i}$. For models without RUEs this results in that $OEP^{(r)} < OEP$, regardless of the size of $\Delta y_{s,i}$.

The last three properties apply also to $OEP_i^{(r)}$, $OEP_{min}^{(r)}$ and $OEP_{max}^{(r)}$.

Some of the above properties concerning OEP and $OEP^{(r)}$ are shown in Figure 1, for the model (1a) with $\mathbf{A}(n \times 1) = [11 \ldots 1]^T$ and $\mathbf{C}(n \times n) = \sigma^2 \mathbf{I}$ for $n = 3, \ldots, 10$. It should be noted that for $n = 2$, due to a Global RUE in this model, OEP and $OEP^{(r)}$ have zero values.

5 Seeking reflection of OEP in MIB values

Due to conceptual differences, OEP and MIB (Teunissen, 2018; Imparato et al., 2018) cannot be compared directly. However, we can talk about a relation such as the reflection of OEP in the MIB values. Following the OEP definition, in models with larger OEP we can expect lower values of Minimal Identifiable

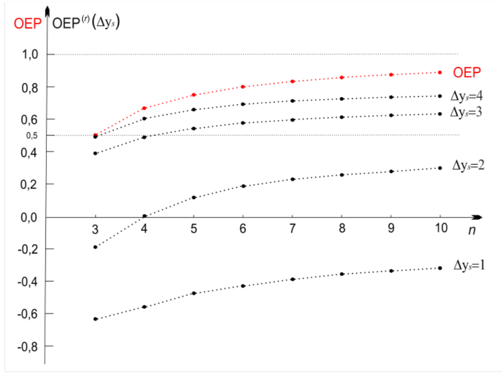


Figure 1. Variability of OEP and $OEP^{(r)}(\Delta y_s)$ with increase in n (for $n = 3$ and $\Delta y_s = 4$, $OEP = 0.50$, $OEP^{(r)} = 0.49$)

Bias (MIB). In other words, in a model with greater OEP it should be possible to identify an observation contaminated by a gross error of smaller magnitude.

Prior to the analysis of specific examples, we introduce the following auxiliary coefficient:

$$\gamma_i = \frac{OEP_{i,MIB}^{(r)}}{OEP_i}, \tag{23}$$

where: $OEP_{i,MIB}^{(r)}$ – a simplified notation for $OEP_i^{(r)}(\Delta y_{s,i} = MIB_{s,i})$, $MIB_{s,i}$ – standardized MIB for the i-th observation ($\alpha = 0.05$, $\beta = 0.20$).

The coefficient γ_i indicates the amount of reduction of the redundancy-induced potential of the i-th observation due to random observation errors and a gross error of the size of $MIB_{s,i}$, contaminating this observation.

For models where $OEP_i = OEP$ for $i = 1, \ldots, n$, the index i may be omitted. This applies also to $MIB_{s,i}$ where $MIB_{s,i} = MIB_s$ ($i = 1, \ldots, n$).

In Table 1 we show the results of computations for the observation scheme as in Ex. 5 (see Sect. 6). The results for some other schemes are given in a descriptive form.

The range of variation of γ_i for Ex.5 is [0.87; 0.94] and the average value $\overline{\gamma} = 0.91$.

The results as regards γ for other schemes are the following:

- Ex. 2-6 obs. For each observation we have $OEP_{i,MIB}^{(r)} = 0.54$; $OEP_i = 0.6$; $MIB_{s,i} = 5.21$; $\gamma_i = 0.90$;
- Ex. 4-9 obs. For observation y_8 and y_9 that do not belong to RUEs $\gamma_i = 0.91$ and 0.94 respectively;
- Ex. 4-10 obs. Except for y_3, y_4 and y_5 that belong to RUE: $\gamma_i[0.84; 0.94]$, $\overline{\gamma} = 0.88$;
- Ex. 4-12 obs. Except for y_3 and y_4 that belong to RUE: $\gamma_i[0.85; 0.91]$, $\overline{\gamma} = 0.88$;
- For Ex. 3*(GPS)-12 obs., where the matrix $\mathbf{R}_s$ has a specifically disadvantageous structure regarding the transfer of disturbance effect between the individual observations (see Sect. 6), we got $\gamma_i[0.82; 0.84]$, $\overline{\gamma} = 0.83$.

In the search for the relationship between OEP_i and $MIB_{s,i}$, the Pearson correlation coefficient for these quantities was checked for three Examples, obtaining:

- Ex. 4-12 obs. (excl. y_3 and y_4 that belong to RUE) $\rho = -0.87$ (strong correlation);
- Ex. 5-13 obs. (see Table 1); $\rho = -0.80$ (strong correlation);
- Ex. 3(GPS)-12 obs.; $\rho = -0.53$ (moderate correlation – most probably due to the specifically disadvantageous structure of $\mathbf{R}_s$).

The negative correlation is consistent with the expectation

Table 1. Relation between OEP_i, $\mathrm{OEP}_i^{(r)}$ and $\mathrm{MIB}_{s,i}$ for observation scheme as in Ex. 5-13 obs.

Obs. No.	1	2	3	4	5	6	7	8	9	10	11	12	13
$\mathrm{MIB}_{s,i}$	9.58	5.89	7.81	7.25	5.95	5.04	7.19	5.81	5.21	7.73	5.78	4.67	4.70
$\mathrm{OEP}_{i,\mathrm{MIB}}^{(r)}$	0.61	0.65	0.65	0.65	0.66	0.66	0.65	0.64	0.69	0.64	0.66	0.71	0.69
OEP_i	0.65	0.71	0.72	0.71	0.74	0.73	0.70	0.70	0.78	0.70	0.73	0.82	0.78
γ_i	0.94	0.91	0.91	0.92	0.90	0.90	0.92	0.92	0.88	0.92	0.90	0.87	0.88

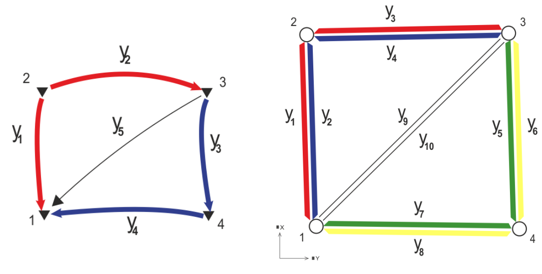
that the greater redundancy-induced potential of a model for a given observation should enable its identification (as an outlier) if infected with a gross error of smaller size.

The empirically obtained overall result $\gamma_i[0.84, 0.94]$, $\overline{\gamma} = 0.90$ (excl. Ex. 3), and especially the negative linear correlation between OEP_i and $\mathrm{MIB}_{s,i}(i = 1, \ldots, n)$, show that there exists a reflection of the OEP_i values in the $\mathrm{MIB}_{s,i}$ values. The reflection allows us to state that the concept of outlier-exposing potential proposed in this work is reasonable and correct. We may then expect that higher OEP and OEP_i values may imply greater effectiveness and efficiency in identifying outliers.

6 Observation schemes used in tests

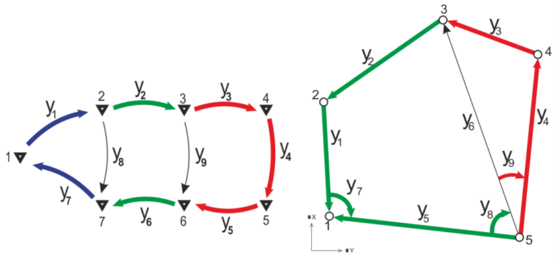
The computations were carried out for the following observation schemes and their observational variants:

- Ex. 1. Repeated measurements of one quantity; $\mathbf{A}(n \times 1) = [1 1 \ldots 1]^T$
  - 3 obs.; $\sigma_1 = \sigma_2 = \sigma_3 = 1\,\mathrm{mm}$
  - 3* obs.; $\sigma_1 = 1\,\mathrm{mm}$, $\sigma_2 = 3\,\mathrm{mm}$, $\sigma_3 = 2\,\mathrm{mm}$
  - 4 obs.; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1\,\mathrm{mm}$
- Ex. 2. 4-point leveling scheme (Figure 2a)
  - 5 obs. $\sigma_1 = \sigma_2 = \ldots = \sigma_5 = 1\mathrm{mm}$, RUE 1 = $\{y_1, y_2\}$; RUE 2 = $\{y_3, y_4\}$
  - 5* obs. $\sigma_1 \div \sigma_5$: 2, 2, 3, 3, 2 [mm], RUE as above;
  - 6 obs. $\sigma_1 = \sigma_2 = \ldots = \sigma_6 = 1\mathrm{mm}$; $r_{ij} = 0$ for (y_1,y_3), (y_2,y_4), (y_5,y_6)
- Ex. 3. 4-point 2D GPS scheme (Figure 2b)
  - 10 obs. $\sigma_1 = \sigma_2 = \ldots = \sigma_{10} = 10\,\mathrm{mm}$, RUE 1 = $\{y_1, y_3\}$; RUE 2 = $\{y_2,y_4\}$; RUE 3 = $\{y_5,y_7\}$; RUE 4 = $\{y_6,y_8\}$, 50 elements $r_{ij} = 0$ in $\mathbf{R}_S$;
  - 12 obs. $\sigma_1 = \sigma_2 = \ldots = \sigma_{12} = 10\,\mathrm{mm}$, 84 elements $r_{ij} = 0$ in $\mathbf{R}_S$;
  - 12* obs. $\sigma_1 \div \sigma_{12}$: 2, 2, 3, 3, 2, 2, 4, 4, 2.5, 2.5, 5, 5 [mm], $r_{ij} = 0$ as above;
- Ex. 4. 7-point levelling scheme (Figure 3a)
  - 9 obs. $\sigma_1 \div \sigma_9$: 2, 1, 2, 1, 2, 2, 1, 2, 2 [mm]; RUE 1 = $\{y_1, y_7\}$; RUE 2 = $\{y_2, y_6\}$; RUE 3 = $\{y_3, y_4, y_5\}$
  - 10 obs. as 9 obs. + y_{10}, 1–3, $\sigma = 1\,\mathrm{mm}$; RUE= $\{y_3, y_4, y_5\}$
  - 11 obs. as 10 obs. + y_{11}, 3–5, $\sigma = 1\,\mathrm{mm}$; RUE= $\{y_3, y_4, y_5\}$
  - 12 obs. as 11 obs. + y_{12}, 2–6, $\sigma = 1\,\mathrm{mm}$; RUE= $\{y_3, y_4\}$
- Ex. 5. 5-point angle-linear scheme (Figure 3b, Table 2)
  - 9 obs. $\sigma_1 \div \sigma_9$: 2, 3, 2, 3.5, 3, 4 [mm], 3, 3, 3 ["]; RUE 1 = $\{y_1,y_2,y_5,y_7,y_8\}$; RUE 2 = $\{y_3,y_4,y_9\}$
  - 10 obs. as 9 obs. + angle y_{10}, 2–3–1 (C–L–R code), σ = 3", RUE = $\{y_3,y_4,y_9\}$
  - 11 obs. as 10 obs. + angle y_{11}, 3–5–2 (C–L–R code), σ = 3", RUE = $\{y_3,y_4,y_9\}$
  - 13 obs. as 11 obs. + distances y_{12}, 1–4, $\sigma = 4.5\,\mathrm{mm}$, y_{13}, 2–4, $\sigma = 4\,\mathrm{mm}$
- Ex. 6. 14-point angle-linear scheme (Figure 4)
  - 48 angles (σ = 10") and 48 distances ($\sigma = 5\,\mathrm{mm}$)



(a) (b)

Figure 2. (a) 4-point levelling scheme, (b) 2D GPS scheme; Local RUEs marked with colours



(a) (b)

Figure 3. (a) 7-point leveling scheme, (b) 5-point angle-linear scheme; Local RUEs marked with colours

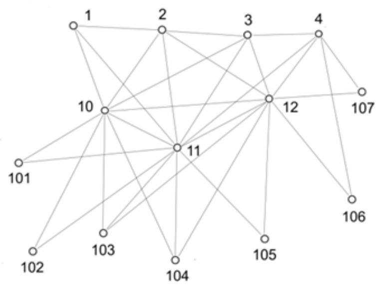


Figure 4. 14-point angle-linear network

Table 2. Approximate coordinates [m]

No.	1	2	3	4	5
X	25.0	112.5	240.0	202.0	2.0
Y	11.0	7.0	135.0	240.0	200.0

Table 3. The computed OEP values versus the redundancy levels $\bar{r}$

Ex.	u	d	$\bar{r}_0$	n	OEP_{min}	OEP_{max}	OEP	$\bar{r}$
Ex. 1	2	1	0.50	3	0.50	0.50	0.50	0.67
				3*	0.36	0.68	0.52	0.67
				4	0.67	0.67	0.67	0.75
Ex. 2	4	1	0.25	5	0.42	0.44	0.43	0.40
				5*	0.33	0.54	0.49	0.40
				6	0.60	0.60	0.60	0.50
Ex. 3 GPS	8	2	0.14	10	0.74	0.75	0.75	0.40
				12	0.82	0.82	0.82	0.50
				12*	0.78	0.85	0.81	0.50
Ex. 4	7	1	0.14	9	0.52	0.63	0.57	0.33
				10	0.61	0.79	0.66	0.40
				11	0.66	0.81	0.74	0.46
				12	0.70	0.82	0.77	0.50
Ex. 5	10	3	0.13	9	0.37	0.53	0.43	0.22
				10	0.51	0.62	0.57	0.30
				11	0.58	0.69	0.64	0.36
				13	0.65	0.82	0.73	0.46
Ex. 6	28	3	0.04	96	0.93	0.98	0.96	0.74

Table 4. Values of OEP and $OEP^{(r)}(\Delta y_s)$ for Ex. 2

Ex. 2. 6 obs.			Δy_s 1	2	3	4	5	6
OEP	0.60	$OEP^{(r)}$	−0.44	−0.07	0.30	0.48	0.53	0.55
OEP_{min}	0.60	$OEP^{(r)}_{min}$	−0.45	−0.08	0.30	0.48	0.53	0.55
OEP_{max}	0.60	$OEP^{(r)}_{max}$	−0.43	−0.06	0.30	0.48	0.53	0.55

7 Results of the OEP and $OEP^{(r)}$ computations and their discussion

In the case of models containing RUEs, the extended version of the potential is determined only for individual observations that do not belong to RUE, i.e., with non-zero $OEP_i^{(r)}$.

i. Empirical study of OEP

The results shown in Table 3 confirm the properties formulated based on theoretical considerations (see Sect. 3). Here are some additional conclusions:

- differentiation of observation accuracy results in the differentiation of the OEP_i values (see Ex. 1-3*obs., Ex. 2-5*obs. and Ex. 3-12*obs. as compared to Ex. 1-3obs., Ex. 2-5obs. and Ex. 3-12obs., respectively);
- the residuals' correlation structure in Ex. 3(GPS), especially for 12 observations, is specific due to many observation pairs with $r_{ij} = 0$, and hence $RC_{ij} = 1$. This contributes to high OEP values at a not particularly high level of $\bar{r}$;
- very high OEP values in Ex. 6 have been caused not only by a high redundancy level of a model but also by the influence of a great many mutually remote observations (i.e., with RC_{ij} close to 1).

ii. Empirical study of $OEP^{(r)}$

The relation between OEP and $OEP^{(r)}$ for some levels of $\Delta y_{s,i}$ is shown in Figure 1 (Sect. 4) for the observation scheme as in Ex. 1 with $n = 3, \ldots, 10$. Similar information, but in a tabular form, is presented for the observation systems as in Ex. 2-6obs., Ex. 5-13obs. and Ex. 6 (Tables 4, 5, and 6).

The above results (together with the intermediate ones) constitute the information that was used for the formulation of the properties of $RC_{ij}^{(r)}$, OEP, and $OEP^{(r)}$, presented in Section 4.

Table 5. Values of OEP and $OEP^{(r)}(\Delta y_s)$ for Ex. 5

Ex. 5. 13 obs.			Δy_s 1	2	3	4	5	6
OEP	0.73	$OEP^{(r)}$	−0.43	−0.04	0.30	0.50	0.60	0.65
OEP_{min}	0.65	$OEP^{(r)}_{min}$	−0.52	−0.32	−0.07	0.16	0.35	0.48
OEP_{max}	0.82	$OEP^{(r)}_{max}$	−0.35	0.17	0.52	0.66	0.72	0.75

Table 6. Values of OEP and $OEP^{(r)}(\Delta y_s)$ for Ex. 6

Ex. 6. 96 obs.			Δy_s 1	2	3	4	5	6	7
OEP	0.96	$OEP^{(r)}$	−0.30	0.25	0.58	0.72	0.79	0.83	0.85
OEP_{min}	0.93	$OEP^{(r)}_{min}$	−0.43	−0.02	0.35	0.57	0.69	0.75	0.79
OEP_{max}	0.98	$OEP^{(r)}_{max}$	−0.22	0.38	0.67	0.78	0.83	0.86	0.88

Verifying these properties on a larger set of observation systems would also help develop the optimal number of simulations of random error vectors. The weakness of the measures adopted for the outlier-exposing potential of a model is the averaging of the values of RC_{ij} indices, which does not reflect their actual variation in a model.

Although the OEP and $OEP^{(r)}$ measures do not have a probabilistic basis, such as in statistical hypotheses testing, they describe the model's properties related to the process of identifying outliers. That is why we believe they can be considered auxiliary reliability measures in addition to the existing ones such as MDB and MIB.

8 Concluding remarks

The basic version of OEP indicates the outlier-exposing ability of a given model induced by its redundancy, independent of the magnitude of gross error. The extended version $OEP^{(r)}$ shows the reduction of this ability due to the masking effect of observation errors, for the specified magnitude of gross error. Both versions enable comparing different models despite their different minimal average redundancy.

Although they do not have a probabilistic basis, they provide information about the model's outlier-exposing ability level, making outlier detection and identification easier. In this respect, they can supplement the existing reliability measures such as MDB and MIB.

Besides the GMMs for network adjustment, the proposed concept also applies to such GMMs as, e.g., linear regression, similarity transformation, or affine transformation.

A desired direction of future research would be to investigate the degree of correlation of the OEP level of the model with the level of effectiveness and efficiency of identifying outliers (as a separate or inherent diagnostic procedure) in parameter estimation using robust and non-robust methods. The results of the research presented in this paper allow us to expect a positive correlation. It is also planned to work on specifying the value of the acceptable lower limit of OEP, adopted here based on approximate reasoning.

Acknowledgments

The authors would like to express their thanks to Dr. Krzysztof Nowel from the University of Warmia and Mazury in Olsztyn for sharing his original MDB_MIB computation program.

Author Contributions

W.P. theory and analyses, figure design; Z.P. computations, preparation of figures.

References

Baarda, W. (1968). A testing procedure for use in geodetic networks. Technical report. *Netherlands Geodetic Commission. Publications on Geodesy, New Series*, 2(5).

Chatterjee, S. and Hadi, A. S. (2009). *Sensitivity analysis in linear regression*. John Wiley & Sons.

Ding, X. and Coleman, R. (1996). Multiple outlier detection by evaluating redundancy contributions of observations. *Journal of geodesy*, 70:489–498, doi:10.1007/BF00863621.

Durdag, U. M., Hekimoglu, S., and Erdogan, B. (2022). What is the relation between smearing effect of least squares estimation and its influence function? *Survey Review*, 54(385):320–331, doi:10.1080/00396265.2021.1939590.

Förstner, W. (1983). Reliability and discernability of extended Gauss-Markov models. In *Deut. Geodact. Komm. Seminar on Math. Models of Geodetic Photogrammetric Point Determination with Regard to Outliers and Systematic Errors, Munchen, Germany*, number 98, pages 79–104.

Huber, P. J. (1972). Robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, doi:10.1214/aoms/1177692459.

Imparato, D., Teunissen, P., and Tiberius, C. (2018). Minimal detectable and identifiable biases for quality control. *Survey review*, 51(367):289–299, doi:10.1080/00396265.2018.1437947.

Lehmann, R. (2013). On the formulation of the alternative hypothesis for geodetic outlier detection. *Journal of geodesy*, 87:373–386, doi:10.1007/s00190-012-0607-y.

Lehmann, R., Lösler, M., and Neitzel, F. (2020). Mean shift versus variance inflation approach for outlier detection – A comparative study. *Mathematics*, 8(6):991, doi:10.3390/math8060991.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: theory and methods*. John Wiley & Sons.

Prószyński, W. (1994). Criteria for internal reliability of linear least squares models. *Bulletin géodésique*, 68:162–167, doi:10.1007/BF00808289.

Prószyński, W. (2008). The vector space of imperceptible observation errors: a supplement to the theory of network reliability. *Geodesy and Cartography*, 57(1):3–19.

Prószyński, W. (2012). *Internal robustness of linear models to disturbances in data – uncorrelated and correlated observations (in Polish)*, volume Centre for Advanced Studies, Lecture Notes No.6. Warsaw University of Technology.

Prószyński, W. (2015). Revisiting Baarda's concept of minimal detectable bias with regard to outlier identifiability. *Journal of geodesy*, 89:993–1003, doi:10.1007/s00190-015-0828-y.

Rofatto, V. F., Matsuoka, M., Klein, I., Veronez, M., Bonimani, M., and Lehmann, R. (2020). A half-century of Baarda's concept of reliability: a review, new perspectives, and applications. *Survey review*, 52(372):261–277, doi:10.1080/00396265.2018.1548118.

Teunissen, P. (2006). *Testing theory: an introduction*, volume Series on Mathematical Geodesy and Positoning. VSSD.

Teunissen, P. J. (2018). Distributional theory for the DIA method. *Journal of geodesy*, 92(1):59–80, doi:10.1007/s00190-017-1045-7.

Teunissen, P. J., Imparato, D., and Tiberius, C. C. (2017). Does RAIM with correct exclusion produce unbiased positions? *Sensors*, 17(7):1508, doi:10.3390/s17071508.