J Geod (2015) 89:993–1003
DOI 10.1007/s00190-015-0828-y

ORIGINAL ARTICLE

Revisiting Baarda's concept of minimal detectable bias with regard to outlier identifiability

W. Prószyński[1]

Received: 3 October 2014 / Accepted: 25 May 2015 / Published online: 2 July 2015

Abstract The concept of minimal detectable bias (MDB) as initiated by Baarda (Publ Geod New Ser 2(5), 1968) and later developed by Wang and Chen (Acta Geodaet et Cartograph Sin Engl Edn 42–51, 1994), Schaffrin (J Eng Surv 123:126–137, 1997), Teunissen (IEEE Aerosp Electron Syst Mag 5(7):35–41, 1990, J Geod 72:236–244 1998, Testing theory: an introduction. Delft University Press, Delft, 2000) and others, refers to the issue of outlier detectability. A supplementation of the concept is proposed for the case of correlated observations contaminated with a single gross error. The supplementation consists mainly of an outlier identifiability index assigned to each individual observation in a network and a mis-identifiability index being the maximum probability of identifying a wrong observation. To those indices there can also be added the MDB multiplying factor to increase the identifiability index to a satisfactory level. As auxiliary measures there are indices of partial identifiability concerning pairs of observations. The indices were derived assuming the generalized outlier identification procedure as in Knight et al. (J Geod. doi:10.1007/s00190-010-0392-4, 2010), which with one outlier case being assumed is similar to Baarda's *w*-test (Baarda in Publ Geod New Ser 2(5), 1968). The following two options of identifiability indices and partial identifiability indices are distinguished: I. the indices related to identification of a contaminated observation within a set of observations suspected of containing a gross error (identifiability), II. the indices related to identification of a contaminated observation within a whole set of observations (pseudo-identifiability). To characterize the proposed approach in the context of the existing solutions of similar topic being the separability testing, the properties of both types of identifiability indices are discussed with reference to the concept of Minimal Separable Bias (Wang and Knight in J Glob Position Syst 11(1):46–57, 2012) and a general approach in Yang et al. (J Geod 87(6):591–604, 2013). Numerical examples are provided to verify the proposed approach.

Keywords Outlier-exposing responses · Outlier identifiability index · Mis-identifiability index · Partial identifiability index · Folded normal distribution

1 Introduction

The concept of minimal detectable error (Baarda 1968), later termed minimal detectable bias (MDB), was a pioneering tool for the analysis of behaviour of a network in the presence of an outlier. Being assumed as a measure of network internal reliability it was meant to span the a priori analysis of network sensitivity to an outlier with the chances to detect it. The original formula for MDB covering the case of correlated observations, was later analyzed by Wang and Chen (1994), Schaffrin (1997), Teunissen (1990, 1998, 2000) and was further extended upon the case of multiple outliers (Teunissen 2000; Knight et al. 2010). It was noticed in numerical tests that the gross errors of MDB magnitudes are often not identified, but identification can be successful at greater magnitudes (e.g. Hekimoglu and Erenoglu 2005). The concept of III-type error was introduced (Hawkins 1980; Förstner 1983) to cover the situations when the error-free observation can be identified mistakenly as the one contaminated by a gross error.

The MDB concept itself does not cover the issue of outlier identifiability. It only determines the minimal magnitude of a

✉ W. Prószyński
wpr@gik.pw.edu.pl

[1] Warsaw University of Technology, Pl. Politechniki 1, 00 661 Warsaw, Poland

gross error in a particular observation, the presence of which in a system can be disclosed through excessive non-centrality effect in a global test. Hence, extending the MDB concept upon the issue of outlier identifiability would be a desirable research task.

Also the "response-based" measures of network internal reliability (Prószyński 2010) that provide reliability criteria clearly interpretable in terms of network responses to outliers, are not associated with the chances for outlier identification.

Taking into account the above description of the problem, the objective of the research was assumed to be the following:

i. to work out a method of evaluating the chances to identify a gross error of the MDB magnitude (assumed to be a single gross error in a system), and together with some other related characteristics to create supplementation of the MDB concept with regard to outlier identifiability,
ii. to propose a method for a priori evaluation of increase of MDB necessary to ensure that the thus obtained gross error can be reliably identified in practice,
iii. to provide probabilistic support for response-based reliability criteria with regard to outlier identifiability.

Since the identifiability issue has much in common with the concept of outlier separability, some common elements are discussed of the proposed approach and the chosen existing methods of outlier separability analysis (Wang and Knight 2012; Yang et al. 2013).

2 Preliminaries

The main part of the paper will be preceded with some preliminary statements and auxiliary concepts describing the approach and presenting the notation applied in the analyses.

2.1 Specifying the terms "detectable gross error" and "identifiable gross error"

Since the distinction between "outlier detection" and "outlier identification" is clearly defined (Teunissen 2000), we give some details that specify the approach to a priori analysis of outlier identifiability proposed in the present paper.

We confine the explanations to the case when a network is contaminated with a single gross error (i.e. one outlier case).

Detectable gross error—an observation error of the magnitude such that its presence in a network is signalized by the global model test statistic exceeding its critical value.

Identifiable gross error—a detectable gross error the exact location of which in a network, i.e. in a particular observation, can be identified among the suspected observations in the first adjustment run, i.e. without subsequent diagnostic operations such as removal or re-weighting of observations. It is when the outlier test statistic of maximum absolute value of all the outlier test statistics that exceed the critical value, corresponds to the contaminated observation.

In the above definition "outlier identification" is clearly separated from "outlier detection", since it is meant as a subsequent process of forming the set of suspected observations and finding among them the contaminated observation.

Unidentifiable gross error—a detectable gross error located in such a specific region of a network (consisting of at least two observations), where all the observations obtain equal values of outlier test statistics. The error is unidentifiable within the region (Cen et al. 2003; Prószyński 2008).

The conditions concerning the existence of the Regions of Unidentifiable Errors (RUE) for correlated observations are derived in Appendix A.

2.2 GM model and the disturbance/response relationship

Let us consider a GM model, written in an original form

$$\mathbf{Ax} + \mathbf{e} = \mathbf{y}; \quad \mathbf{e} \sim (\mathbf{0}, \mathbf{C}) \tag{1}$$

and in the equivalent modified form that exposes the correlation matrix (Prószyński 2010)

$$\mathbf{A}_S\mathbf{x} + \mathbf{e}_S = \mathbf{y}_S; \quad \mathbf{e}_s \sim (\mathbf{0}, \mathbf{C}_s) \tag{2}$$

where $\mathbf{y}$ the $n\times 1$ vector of observations; $\mathbf{A}$ the $n\times u$ design matrix; rank $\mathbf{A} = u-d$ (d—system defect, $d \geq 0$); $\mathbf{x}$ the $u\times 1$ vector of unknown parameters; $\mathbf{e}$ the $n\times 1$ vector of random errors; we shall also use $\mathbf{v} = -\mathbf{e}$; $\mathbf{C}$ the $n\times n$ covariance matrix of $\mathbf{e}$ (positive definite), $\mathbf{C} = \sigma_o^2\mathbf{P}^{-1} = \sigma_o^2\mathbf{Q}$; $\boldsymbol{\sigma} = (\text{diag }\mathbf{C})^{1/2}$, $\mathbf{A}_s = \boldsymbol{\sigma}^{-1}\mathbf{A}$, $\mathbf{e}_s = \boldsymbol{\sigma}^{-1}\mathbf{e}$, $\mathbf{y}_s = \boldsymbol{\sigma}^{-1}\mathbf{y}$, $\mathbf{C}_s = \boldsymbol{\sigma}^{-1}\mathbf{C}\boldsymbol{\sigma}^{-1}$, $\mathbf{C}_s$ a correlation matrix; for uncorrelated observations $\mathbf{C}_s = \mathbf{I}$.

The LS estimator of the vector $\mathbf{v}_s$, where $\mathbf{v}_s = -\mathbf{e}_s$, is given by

$$\hat{\mathbf{v}}_s = -\mathbf{H}\mathbf{y}_s \tag{3}$$

where
$\mathbf{H} = \mathbf{I} - \mathbf{A}_s(\mathbf{A}_s^T\mathbf{C}_s^{-1}\mathbf{A}_s)^+\mathbf{A}_s^T\mathbf{C}_s^{-1}$ is the modified reliability matrix (Prószyński 2010), i.e. the reliability matrix for the modified GM model as in (2), $(*)^+$ denotes the pseudo-inverse.

Decomposing the vector $\mathbf{y}_s$, so that $\mathbf{y}_s = \mathbf{y}_s^{\text{true}} - \mathbf{v}_s + \Delta\mathbf{y}_s$, where $\Delta\mathbf{y}_S$ is the vector of standardized observation gross errors (i.e. standardized disturbances), and realizing that $\mathbf{H}\cdot\mathbf{y}_S^{\text{true}} = \mathbf{0}$, we obtain (3) in the form

$$\hat{\mathbf{v}}_{\mathbf{s}} = \mathbf{H}\mathbf{v}_{\mathbf{s}} - \mathbf{H} \cdot \Delta\mathbf{y}_{\mathbf{s}} \tag{4}$$

Denoting the second term in (4) by $\Delta\hat{\mathbf{v}}_{\mathbf{s}}$, being the vector of standardized increments in LS residuals (i.e. standardized responses), we get on its basis the well known *disturbance/response* relationship for the system (2), i.e.

$$\Delta\hat{\mathbf{v}}_{\mathrm{S}} = -\mathbf{H} \cdot \mathbf{\Delta y}_{\mathrm{S}} \tag{5}$$

where $\Delta\hat{\mathbf{v}}_{\mathrm{S}} = -\Delta\hat{\mathbf{e}}_{\mathrm{S}}$.

2.3 A short note on minimal detectable error (MDB) and response-based reliability measures

Below, we present the formula for MDB as given in Wang and Chen (1994), Teunissen (1990, 1996), using the notation as in Sect. 2.2

$$\mathrm{MDB}_i = \boldsymbol{\sigma}_i \cdot \sqrt{\frac{\boldsymbol{\lambda}}{r_i}} \quad r_i = \left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}; \quad r_i\,[0,\infty) \tag{6}$$

where MDB_i minimal detectable bias in the i-th observation; its standardized form i.e. $\mathrm{MDB}_{\mathrm{S},i} = \mathrm{MDB}_i/\sigma_i$ is termed as controllability of the i-th observation, σ_i the standard deviation of the i-th observation, λ the non-centrality parameter (as in a global model test), r_i a generalized reliability number for the i-th observation, $\mathbf{C}_{\mathrm{S}}$, $\mathbf{H}$ the matrices as in (2) and (3), respectively; $\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H} = \mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$.

The generalized reliability number r_i alone can also be considered as internal reliability measure (Caspary 1988).

The behaviour of a system in the presence of a single gross error can also be characterized by the so called response-based internal reliability measures (Prószyński 2010), derived on the basis of disturbance/response relationship (5), i.e. disregarding the random-error environment. For correlated observations the measures are the following pairs of indices

$h_{ii},\ w_{ii}$, or equivalently $h_{ii},\ k_i$

where h_{ii} the i-th diagonal element of the matrix $\mathbf{H}$, w_{ii} the asymmetry index for the i-th row and the i-th column of the matrix $\mathbf{H}$, k_i the ratio of the squared quasi-global response $\mathrm{Q}_{(i)}$ and the squared local response h_{ii} to an outlier in the i-th observation [see formula (28) in Appendix A].

The reliability criteria are the following

$$0.5 < h_{ii} \le 1 \quad \wedge \quad h_{ii} - 2h_{ii}^2 < w_{ii} < h_{ii} - h_{ii}^2 \quad i = 1, \ldots, n \tag{7}$$

or, equivalently

$$0.5 < h_{ii} \le 1 \quad \wedge \quad 0 < k_i < 1 \quad i = 1, \ldots, n$$

They are derived from the postulate that the maximum system response should be located in the observation in which the gross error resides, and that the responses in other observations should possibly be the smallest (Prószyński 2010). Hence, there are then the chances for effective identification of a single gross error residing in any of the observations. We can then state that the criteria determine the area of outlier-exposing responses. The set of values (h_{ii}, w_{ii}) which form this area (see Figs. 2, 3) will be denoted by $\mathrm{S_O}$.

It is not possible to interrelate the above two types of measures, i.e. r_i and (h_{ii}, w_{ii}) or (h_{ii}, k_i) on the grounds of rigorous matrix operations due to different generation principles. So, instead of direct interrelations we can establish indirect correspondence between these measures by finding their values on basis of the model (1) or model (2) components, as shown on a scheme below

$$\mathbf{A},\ \mathbf{C} \to \mathbf{A}_{\mathrm{s}},\ \mathbf{C}_{\mathrm{s}} \to \mathbf{H}, \mathbf{C}_{\mathrm{s}} \to \begin{cases} r_i \\ h_{ii};\ w_{ii} \end{cases} \quad i = 1, \ldots, n \tag{8}$$

3 A study on outlier identifiability evaluation in terms of probability

The a priori analysis of outlier identifiability presented in this paper, refers in principle to outlier identification procedure as in Knight et al. (2010). In that procedure the global model test is followed by the local outlier tests resulting in a set of suspected outliers. The final outcome of the procedure is the observation with a maximum absolute value of the test statistic. The outlier test statistics (i.e w^2) are obtained from mean-shift model. The global model test and the local outlier tests are coordinated by equalizing the non-central parameters and selecting the probabilities according to the β-Method (Baarda 1968). In the present paper, for finding the suspected outliers and identifying the contaminated observation instead of w^2 the $|w|$ values are used, assuming $|w|_{\mathrm{crit}} = \sqrt{w_{\mathrm{crit}}^2}$.

The w-variables, being the standardized random variables, are defined by

$$w_{i(i)} = \frac{\hat{z}_{i(i)}}{\sigma_{\hat{z}_{i(i)}}}; \quad w_{j(i)} = \frac{\hat{z}_{j(i)}}{\sigma_{\hat{z}_{j(i)}}} \quad i, j = 1, \ldots, n \quad j \neq i \tag{9}$$

where "i" denotes the observation contaminated with a gross error, "j" denotes any other observation; $\hat{z}$ is the LS estimator of a gross error, obtained on basis of "mean-shift" model (Knight et al. 2010)

With one outlier case being assumed as in the present research, the above testing procedure is similar to Baarda's w-test (Baarda 1968).

3.1 Parameters of outlier test statistics for the needs of identifiability analysis

In the notation of the present paper the w-variables as in (9) in a network contaminated with a single gross error $\Delta y_{\mathrm{S},i}$, have the following detailed form

$$
\begin{aligned}
w_{i(i)} &= \frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{i*}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}}\cdot(\mathbf{e}_{\mathrm{S}}+\Delta\mathbf{y}_{\mathrm{S}(i)})\\
&= \frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{i*}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}}\cdot\mathbf{e}_{\mathrm{S}}+\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}\,\Delta y_{\mathrm{S},i}\\
w_{j(i)} &= \frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{j*}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}}(\mathbf{e}_{\mathrm{S}}+\Delta\mathbf{y}_{\mathrm{S}(i)})\\
&= \frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{j*}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}}\mathbf{e}_{\mathrm{S}}+\frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ji}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}}\Delta y_{\mathrm{S},i}\\
&\quad i,j=1,\ldots,n\quad j\neq i
\end{aligned}
\tag{10}
$$

where $\{\cdot\}_{i*}$ and$\{\cdot\}_{j*}$ denote the i-th and the j-th row of $\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$.

With $\mathbf{e}\sim N(\mathbf{0},\mathbf{C})$, and consequently $\mathbf{e}_{\mathrm{s}}\sim N(\mathbf{0},\mathbf{C}_{\mathbf{s}})$, we get after simple operations

$$
w_{i(i)}\sim N(\mu_i,1)\quad \mu_i=\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}\cdot\Delta \mathrm{y}_{\mathrm{S},i}
\tag{11}
$$

$$
w_{j(i)}\sim N(\mu_j,1)\quad \mu_j=\frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ji}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}}\cdot\Delta \mathrm{y}_{\mathrm{S},i}
\tag{12}
$$

$$
\rho_{ij}=\mathrm{cor}(w_{i(i)},w_{j(i)})=\frac{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ij}}{\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}\sqrt{\left\{\mathbf{H}^T\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}}
\tag{13}
$$

as in (Förstner 1983).

To analyze outlier identifiability, it is most reasonable to consider detectable gross errors, i.e. $\Delta \mathrm{y}_{\mathrm{S},i}\geq \mathrm{MDB}_{\mathrm{S},i}$, where $\mathrm{MDB}_{\mathrm{S},i}$ as in (6). Substituting $\Delta \mathrm{y}_{\mathrm{S},i}=\mathrm{MDB}_{\mathrm{S},i}$ into (11) and (12), we obtain

$$
\mu_i=\sqrt{\lambda};\quad \mu_j=\rho_{ij}\cdot\sqrt{\lambda};\quad \rho_{ij}
\tag{14}
$$

The formula (14) reflects a well known property (Förstner 1983) that with the defined I type and II type errors the correlation between the outlier test statistics is decisive for identification of the contaminated i-th observation.

3.2 Identifiability indices and their properties

To identify within a set of suspected observations the i-th observation in which a gross error of MDB magnitude resides, we need that $|w_{i(i)}|$ is dominating over each of the corresponding absolute values for the remaining observations within this set. For each of the suspected observations we have $|w|>|w|_{\mathrm{crit}}$, or for short $|w|>c$ (we assume $|w|_{\mathrm{crit}}=\sqrt{w_{\mathrm{crit}}^2}$).

Using $\frac{|w_{j(i)}|}{|w_{i(i)}|}<1$ as an equivalent condition to $|w_{i(i)}|>|w_{j(i)}|$, we may form for the i-th observation an identifiability index denoted as ID_i, defined in terms of conditional probability

$$
\mathrm{ID}_i=\mathrm{P}(\mathrm{Q}_i\,|\bar{\mathrm{R}}_i\,)
\tag{15}
$$

where

$$
\mathrm{Q}_i=\frac{|w_{1(i)}|}{|w_{i(i)}|}<1\ \cap\cdots\cap\frac{|w_{j(i)}|}{|w_{i(i)}|}<1\ \cap\cdots\cap\frac{|w_{n-1(i)}|}{|w_{i(i)}|}<1;
$$

$$
j\neq i
$$

$$
\mathrm{R}=\left|w_{1(i)}\right|<c\cup\ldots\cup\left|w_{i(i)}\right|<c\cup\cdots\cup\left|w_{n(i)}\right|<c
$$

where non-centralities μ of the w-variables are determined for $\mathrm{MDB}_{s,i}$ as in (14); $\bar{\mathrm{R}}_i$ being an event opposite to R_i, contains all possible sets of suspected observations, each corresponding to a particular distribution of random errors in a single measurement of a network.

In formulating Q_i, we take into account the fact that domination of $|w_{i(i)}|$ within the set of all the observations implies its domination within any set of suspected observations containing $w_{i(i)}$.

Using for each component in Q_i a symbol Z as for a ratio of two folded normal variables (see Appendix B), we may write (15) in the form

$$
\begin{aligned}
\mathrm{ID}_i=\mathrm{P}\Big(&\mathrm{Z}_{1(i)}<1\ \cap\cdots\cap\mathrm{Z}_{j(i)}\\
&<1\ \cap\cdots\cap\mathrm{Z}_{n-1(i)}<1\,\Big|\bar{\mathrm{R}}_i\Big)\quad j\neq i
\end{aligned}
\tag{16}
$$

Due to a high complexity of the definition (15), increasing with the number (n) of observations in a network, an empirical method based on numerical simulation of random observation errors was applied in the research. The method consists in:

- simulating numerically a certain number (e.g. 1000) of n-dimensional vectors of correlated random errors (according to a given $\mathbf{C}$);

- computing w-variables for each vector of random errors using the formulas (10), the systematic components being as in (14);
- after elimination of the sets of w-variables where the critical values are not exceeded, computing sample frequency for the sets where $|w|$ for a contaminated observation (such that $|w| > c$) is dominating, the sample frequency being empirical approximation of ID. As a check on correctness of simulation procedure a sample frequency for the eliminated sets of w-variables (i.e. with $|w| < c$) was used as being empirical approximation of II type error probability β.

To extend the scope of identifiability analysis, the computer program written for the method contains the formulas (10) in a modified form introducing a multiplying factor, such that the systematic components are as follows

$$\mu_i = g_i \cdot \sqrt{\lambda}; \quad \mu_j = \boldsymbol{\rho}_{ij} \cdot g_i \cdot \sqrt{\lambda}; \quad g_i > 0 \tag{17}$$

which corresponds to the use of $\Delta \mathbf{y}_{\mathrm{S},i} = g_i \cdot \mathrm{MDB}_{\mathrm{S},i}$.

This modification can be used in case of unsatisfactory values of ID_i obtained with $\Delta \mathbf{y}_{\mathrm{S},i} = \mathrm{MDB}_{\mathrm{S},i}$.

We do not have exact theoretical reference for evaluating the accuracy of the simulation method. Therefore, we may only analyze the degree of dispersion of the ID values for different sets of simulated data and different observations in a network. The estimates obtained in that way for networks in Examples 1 and 2 (see Sect. 6) with 1000 simulations used are within ± 1 or ± 2 % (standard deviations).

For the purpose of this study we consider also identification of the contaminated observation without setting restrictions onto the values of w-variables. Such a procedure that covers also the outlier detection is a departure from the assumed definition of outlier identification (see Sect. 2.1) and will be termed pseudo-identification. Consequently, we shall operate with *a pseudo-identifiability index*, denoted by ID_i^*, and having the form

$$\mathrm{ID}_i^* = \mathrm{P}(\mathrm{Q}_i) \tag{18}$$

where Q_i as in (15).

Although to a smaller degree than in the case of $\mathrm{P}(\mathrm{Q}_i \,|\bar{\mathrm{R}}_i)$ (15), finding $\mathrm{P}(\mathrm{Q}_i)$ is still a complex computation task. However, we may get empirical approximation of this index $(\overline{\mathrm{ID}}_i^*)$ by means of slightly modified simulation method.

On the grounds of probability theory some relations can be established between ID_i^* and ID_i

$$\mathrm{P}\left\{\mathrm{Q}_i \,|\bar{\mathrm{R}}_i\right\} = \frac{\mathrm{P}\left\{\mathrm{Q}_i \cap \bar{\mathrm{R}}_i\right\}}{\mathrm{P}\left\{\bar{\mathrm{R}}_i\right\}} = \frac{\mathrm{P}\left\{\mathrm{Q}_i\right\} - \mathrm{P}\left\{\mathrm{Q}_i \cap \mathrm{R}_i\right\}}{\mathrm{P}\left\{\bar{\mathrm{R}}_i\right\}}$$

where $\mathrm{P}(\bar{\mathrm{R}}_i) = 1 - \beta$,

and hence

$$\mathrm{ID}_i^* = (1-\beta) \cdot \mathrm{ID}_i + \mathrm{P}\left\{\mathrm{Q}_i \cap \mathrm{R}_i\right\} \tag{19}$$

Assuming that $\mathrm{P}(\{\mathrm{Q}_i \cap \mathrm{R}_i\}) > 0$, we get

$$\mathrm{ID}_i^* > (1-\beta) \cdot \mathrm{ID}_i \tag{20}$$

Hypothetically, the case that $\mathrm{ID}\,_i^* = \mathrm{ID}\,_i$ might occur when $\mathrm{P}(\mathrm{Q}_i \cap \bar{\mathrm{R}}_i) = \mathrm{P}(\mathrm{Q}_i) \cdot \mathrm{P}(\bar{\mathrm{R}}_i)$, i.e. when Q_i and $\bar{\mathrm{R}}_i$ were independent events. Then with $\mathrm{ID}_i = 1$, we would also have $\mathrm{ID}_i^* = 1$, which would imply domination of $|w_{i(i)}|$ in each possible set in R. Since the above independency is only a detached theoretical assumption, we can only state that ID_i is an unattainable upper limit for $\mathrm{ID}\,_i^*$.

The above relations have been confirmed by the results obtained from the simulation method.

3.3 Partial identifiability indices and their properties

As an auxiliary tool for network analysis, the *partial identifiability indices* for pairs of observations were introduced, i.e. for the i-th observation contaminated by a gross error and the j-th observation being error-free. Similarly to two options of identifiability indices (Sect. 3.2) we distinguish

- partial identifiability index $\mathrm{ID}_{i/j}$

$$\mathrm{ID}_{i/j} = \mathrm{P}\left(\frac{|w_{j(i)}|}{|w_{i(i)}|} < 1 \middle| \bar{\mathrm{R}}_{ij}\right) \tag{21}$$

where $\mathrm{R}_{ij} = |w_{i(i)}| < c \cup |w_{j(i)}| < c$
or in notation of (16)

$$\mathrm{ID}_{i/j} = \mathrm{P}(\mathrm{Z}_{j(i)} < 1 \,|\bar{\mathrm{R}}_{ij})$$

- partial pseudo-identifiability index $\mathrm{ID}_{i/j}^*$.

$$\mathrm{ID}_{i/j}^* = \mathrm{P}\left(\frac{|w_{j(i)}|}{|w_{i(i)}|} < 1\right), \tag{22}$$

or in notation of (16),

$$\mathrm{ID}_{i/j} = \mathrm{P}\{\mathrm{Z}_{j(i)} < 1\}$$

The indices are the values of distribution function of ratio of two folded normal variables. In the case of $\mathrm{ID}_{i/j}$ the space of the values of w-variables is reduced in terms of absolute values, assuming that both the i-th and the j-th observation are the elements of a set of suspected observations.

For finding the values of $\mathrm{ID}_{i/j}^*$, a MATLAB-based software has been developed (Appendix B) for computing

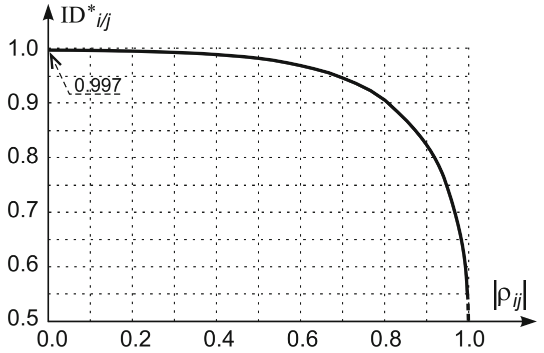

Fig. 1 Variability of the index $\mathrm{ID}^*_{i/j}$ as a function of correlation $\rho(w_i, w_j)$

the values of the distribution function of Z. We can also find empirical approximation of each type of index (i.e. $\mathrm{ID}_{i/j}$ and $\mathrm{ID}^*_{i/j}$) by means of the simulation method presented in Sect. 3.2, by computing sample frequencies for chosen pairs of observations.

The following properties of $\mathrm{ID}^*_{i/j}$ indices can be formulated:

- from the formula (14), where $\mu_i > |\mu_j|$, and the property (32) in Appendix B, it follows that for any pair of observations in a network we shall have $\mathrm{ID}^*_{i/j} > 0.5$. Figure 1 shows dependence of $\mathrm{ID}^*_{i/j}$ on magnitude of correlation $|\rho_{ij}|$ ($|\rho_{ij}| < 1$), obtained with $\mu_i = \sqrt{\lambda} = 4.13$ and $\mu_j = \rho_{ij}\sqrt{\lambda} = 4.13 \cdot \rho_{ij}$ (as in formula (14). We can see that the smaller $|\rho_{ij}|$, the greater is $\mathrm{ID}^*_{i/j}$.
- due to $\rho(w_i, w_j) = \rho(w_j, w_i)$, the $\mathrm{ID}^*_{i/j}$ indices are symmetrical within pairs of observations, i.e. $\mathrm{ID}^*_{i/j} = \mathrm{ID}^*_{j/i}$.
- for all the observations forming a RUE region in a network, we shall have $\mathrm{ID}_i = \mathrm{ID}_j = \mathrm{ID}_k \ldots = \mathrm{ID}_{\mathrm{Rue}}$, where $\mathrm{ID}_{\mathrm{Rue}}$ could be termed the identifiability index for a RUE region containing an outlier. For all pairs of observations within RUE we shall have $\mathrm{ID}^*_{i/j} = 0$. The index $\mathrm{ID}_{\mathrm{Rue}}$ does not apply to networks being a RUE as a whole. In such networks $\mathrm{ID}_i = \mathrm{ID}_j = \mathrm{ID}_k \ldots = 0$ and $\mathrm{ID}^*_{i/j} = 0$ for all the observations.

3.4 Mis-identifiability indices and probabilities of III type errors

To cover in a priori analysis the possibility of identifying the j-th error-free observation instead of the contaminated i-th observation, defined as III type error (Hawkins 1980; Förstner 1983), we introduce mis-identifiability indices as shown below

$$\mathrm{MID}_{ij} = \mathrm{P}(\mathrm{Q}_j \,|\bar{\mathrm{R}}_i)$$

$$\mathrm{Q}_j = \frac{|w_{1(i)}|}{|w_{j(i)}|} < 1 \;\cap \cdots \cap\; \frac{|w_{i(i)}|}{|w_{j(i)}|} < 1 \;\cap \cdots \cap\; \frac{|w_{n-1(i)}|}{|w_{j(i)}|} < 1 \quad i \neq j \tag{23}$$

where $\bar{\mathrm{R}}_i$ as in (19).

The indices MID_{ij} correspond to probabilities of committing III type errors, denoted by γ_{ij} (Förstner 1983). The indices are determined for gross errors of the MDB magnitudes.

Using the simulation method we can get empirical approximation of MID_{ij} by computing sample frequency for the sets where $|w_{j(i)}|$ (such that $|w_{j(i)}| > \mathrm{c}$) is dominating

Taking into account the MID_{ij} indices for all the j-th observations, we may find the observation with maximum value of MID_{ij}, i.e. $\mathrm{MID}_{ij,\max}$.

Realizing that all MID_{ij} indices together with ID_i indices refer to disjoint events that form a complete event, we may formulate on basis of (15) and (23) the following relationship

$$\overline{\mathrm{ID}}_i = 1 - \sum_{j=1, j \neq i}^{n-1} \mathrm{MID}_{ij} \tag{24}$$

where n as in (15) is the number of all the observations in a network.

According to (24), with the ID_i values being greater than 0.5 there can be no observation with $\mathrm{MID}_{ij} > 0.5$.

This confirms the well known property, that the greater the probability of finding the contaminated observation (e.g. Wang and Knight 2012), the smaller is the probability of committing the III-type error.

4 Proposed supplementation of the MDB concept for a priori analysis of network reliability

By definition the MDB concept is not associated with outlier identifiability. Based on the study of outlier identifiability evaluation (Sect. 3), we propose supplementation of the MDB concept as in formula (6) with identifiability index ID_i, as in formula (15). The pair $(\mathrm{MDB}_i, \mathrm{ID}_i)$ would characterize the minimal detectable error in a particular observation together with the chances for its identification in a network.

In case of unsatisfactory value of $\mathrm{ID}_{i,}$ we may find the multiplying factor g_i as in (17) that shows the degree of magnification of MDB_i necessary to obtain a required level of outlier identifiability. We may also find a particular j-th ($j \neq i$) observation corresponding to maximum probability of III type error, i.e. γ_{ij} (see mis-identifiability indices $\mathrm{MID}_{ij,\max}$ in Sect. 3.4).

For more detailed analysis of outlier identifiability, we may compute the index ID^*_i and the indices $\mathrm{ID}_{i/j}$ and $\mathrm{ID}^*_{i/j}$ for some chosen pairs of observations.

SUPPLEMENTATION of MDB for the i-th observation can thus be formed in the following two levels:

BASIC—ID_i, g_i, $\mathrm{MID}_{ij,\max}$; *AUXILIARY*—ID^*_i, $\mathrm{ID}_{i/j}$, $\mathrm{ID}_{i/k}$,…, $\mathrm{ID}^*_{i/j}$, $\mathrm{ID}^*_{i/k}$, …

Additionally, by finding the response-based reliability measures (h_{ii}, w_{ii}) or (h_{ii}, k_i) for the analyzed i-th observation, we obtain in an indirect way a link between the network response and the indices ID_i and $\mathrm{MID}_{ij,\max}$.

5 Common elements of the proposed approach with some chosen solutions in outlier separability testing

Although in the present paper the term "separability" is not used explicitly, the proposed identifiability indices can be considered to some extent as outlier separability measures. A direct link of the proposed approach with outlier separability analysis are mis-identifiability indices being the maximum probabilities of III type errors. Analogy can be found between the proposed approach and that in (Wang and Knight 2012). In the letter approach the concept of minimal separable bias (MSB) is presented, being the magnitude of MDB increased by the multiplying factor so as to ensure identifying of an outlier at a satisfactory confidence level (denoted there as $1 - \alpha_s$). This corresponds in the present paper to the use of partial pseudo-identifiability index $\mathrm{ID}^*_{i/j}$. Due to possibility of increasing MDB by the iteratively determined multiplying factor to reach a corresponding level of partial pseudo-identifiability, we may obtain a bias equivalent to MSB. One can also notice that the two options of the standardized separability test statistic (Wang and Knight 2012) contain exactly the arguments h'_1 and h'_2 of a distribution function $\mathrm{P}(\mathrm{Z} < 1)$ as in formula (31) in the present paper. The ratio itself can be a proposal of test statistic for the above mentioned separability test.

Analogies can also be expected between the proposed approach and multiple alternative hypotheses testing (Yang et al. 2013) with respect to definitions of probabilities of committing different types of errors as well as in the relations between these probabilities.

6 Numerical examples

To illustrate the proposed approach we use a levelling network analyzed in Knight et al. (2010) (Fig. 2a) and a GPS network (Fig. 2b). Referring to the first publication gives opportunity to expand the conclusions reached there.

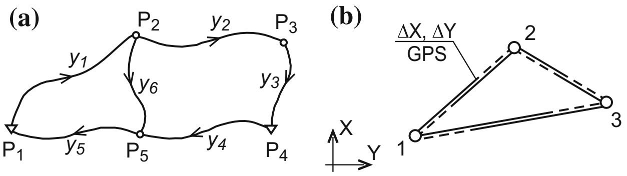

Fig. 2 Networks used in numerical examples

Example 1 For a network in Fig. 2a, we have

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \\ -1 & 0 & 1 \end{bmatrix};$$

$$\mathbf{C}_{\mathrm{S}} = \begin{bmatrix} 1.00 & 0.80 & 0.14 & -0.59 & -0.48 & 0.04 \\ 0.80 & 1.00 & 0.00 & -0.17 & -0.68 & -0.30 \\ 0.14 & 0.00 & 1.00 & -0.67 & 0.25 & 0.76 \\ -0.59 & -0.17 & -0.67 & 1.00 & -0.29 & -0.76 \\ -0.48 & -0.68 & 0.25 & -0.29 & 1.00 & 0.57 \\ 0.04 & -0.30 & 0.76 & -0.76 & 0.57 & 1.00 \end{bmatrix}$$

To save space we show only the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$ and a correlation submatrix for the variables w_2 and w_3

$$\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H} = \begin{bmatrix} 10.58 & \mathbf{-1.06} & \mathbf{-0.48} & 11.54 & 4.48 & 5.97 \\ \mathbf{-1.06} & \mathbf{0.62} & \mathbf{0.28} & \mathbf{-1.06} & \mathbf{-0.55} & \mathbf{-0.91} \\ \mathbf{-0.48} & \mathbf{0.28} & \mathbf{0.13} & \mathbf{-0.48} & \mathbf{-0.25} & \mathbf{-0.41} \\ 11.54 & \mathbf{-1.06} & \mathbf{-0.48} & 13.68 & 5.07 & 6.46 \\ 4.48 & \mathbf{-0.55} & \mathbf{-0.25} & 5.07 & 1.95 & 2.59 \\ 5.97 & \mathbf{-0.91} & \mathbf{-0.41} & 6.46 & 2.59 & 3.56 \end{bmatrix}$$

$$\boldsymbol{\rho}\begin{bmatrix} w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

The mutually parallel column-vectors (and row-vectors) in $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$ are marked in bold. Further results of analysis are given in Table 1 and Fig. 3.

None of the observations satisfy the reliability criteria required for outlier-exposing responses. The network contains RUE formed by the observations 2 and 3. Hence, based on the results of the simulation method we can write $\overline{\mathrm{ID}}_2 = \overline{\mathrm{ID}}_3 = \overline{\mathrm{ID}}_{\mathrm{RUE}} = 0.49$. The equal MDB values for these observations, represent minimal detectable gross errors that are identifiable as located in the RUE region of a network, but are unidentifiable within this region, i.e. $\mathrm{ID}^*_{2/3} = \mathrm{ID}^*_{3/2} = 0$.

It is difficult to find out a relationship between the indices ID_i and internal reliability measures r_i or (h_{ii}, w_{ii}). Except for observation 5, the indices ID_i are not specially differentiated and they represent a low level, slightly exceeding 0.5 for the observation 4.

This level does not ensure a sufficiently reliable identification of gross errors. This is reflected in the values of $\overline{\mathrm{MID}}_{ij,\max}$ indices.

Below, we show the effect upon ID_i of increasing the magnitude of a gross error by applying the multiplying factor $g_i > 1$ [see formula (17) for the observation 1, 5 and 6], i.e.

- obs. 1; $g_1 = 2$, $\overline{\mathrm{ID}}_1 = 0.63$; $g_1 = 3$, $\overline{\mathrm{ID}}_1 = 0.79$;
- obs. 5; $g_5 = 2$, $\overline{\mathrm{ID}}_5 = 0.36$; $g_5 = 3$, $\overline{\mathrm{ID}}_5 = 0.61$;

Table 1 Results of internal reliability and identifiability analysis for the network 1

Obs	σ_i	h_{ii}	w_{ii}	k_i	Crit.	r_i	MDB_i	$\text{MDB}_{S,i}$	$\overline{\text{ID}}_i$	$\overline{\text{MID}}_{ij,\max}$
1	2.35	0.96	−1.49	1.64	–	10.58	$0.72 \cdot \sqrt{\lambda}$	$0.31 \cdot \sqrt{\lambda}$	**0.43**	**0.24** obs. 6
2	1.97	0.60	−0.42	1.82	–	0.622	$2.50 \cdot \sqrt{\lambda}$	$1.27 \cdot \sqrt{\lambda}$	**0.49**	**0.50** obs. 3[a]
3	0.89	0.01	−0.20	2251	–	0.128	$2.50 \cdot \sqrt{\lambda}$	$2.80 \cdot \sqrt{\lambda}$	**0.49**	**0.50** obs. 2[a]
4	2.32	1.02	−4.50	4.29	–	13.68	$0.63 \cdot \sqrt{\lambda}$	$0.27 \cdot \sqrt{\lambda}$	**0.57**	**0.18** obs. 1
5	0.45	0.13	−0.29	23.02	–	1.954	$0.32 \cdot \sqrt{\lambda}$	$0.72 \cdot \sqrt{\lambda}$	**0.19**	**0.31** obs. 4
6	1.18	0.27	−0.39	8.12	–	3.558	$0.63 \cdot \sqrt{\lambda}$	$0.53 \cdot \sqrt{\lambda}$	**0.50**	**0.23** obs. 1
		$\Sigma = 3.00$								

[a] Theoretical values

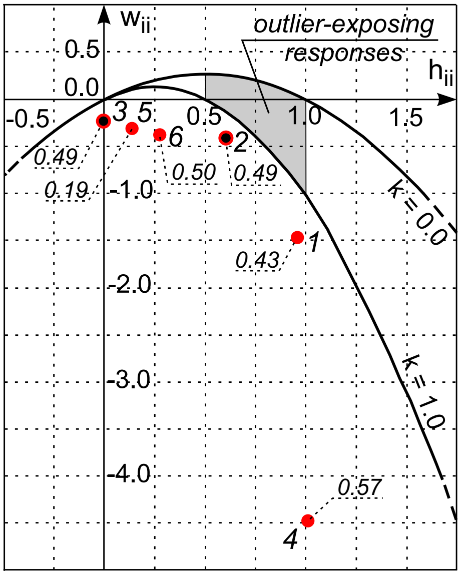

Fig. 3 Identifiability indices (ID_i) shown in a (h_{ii},w_{ii}) system for Network 1

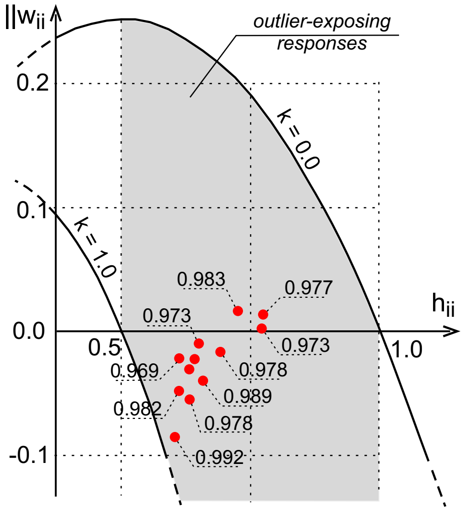

Fig. 4 Identifiability indices (ID_i) shown in a (h_{ii},w_{ii}) system for Network 2

– obs. 6; $g_6 = 2$, $\overline{\text{ID}}_6 = 0.67$; $g_6 = 3$, $\overline{\text{ID}}_6 = 0.82$.

The correlations ρ_{ij} between the w-variables for the observations not forming a RUE are in absolute values within the interval [0.36, 0.98], and hence, the values of $\text{ID}^*_{i/j}$ indices are within [0.64, 0.99] (see Fig. 1).

Example 2 We omit showing the design matrix **A** (12 × 4) (with elements 1 and −1), and confine presentation of the covariance matrix **C** to the range of values of standard deviations and correlation coefficients, i.e. $\sigma[2.5, 3.2]$, $\rho[-0.24, 0.21]$.

As we can see in Fig. 4 all the observations satisfy the reliability criteria required for outlier-exposing responses. The indices $\overline{\text{ID}}_i$ represent high values as they all lay in the interval [0.969, 0.992]. The values of $\overline{\text{MID}}_{ij,\max}$ indices are within the interval [0.002, 0.023]. This means that identification of a gross error of MDB magnitude in each observation is highly reliable.

The correlations ρ_{ij} between the w-variables being in absolute values within the interval [0.002, 0.563] are much smaller than in Network 1. Consequently, the values of $\text{ID}^*_{i/j}$ indices are much greater (see Fig. 1), i.e. [0.973, 0.996].

The values of $\overline{\text{ID}}^*_i$ indices are only slightly smaller than those of $\overline{\text{ID}}_i$, i.e. [0.943, 0.972], which confirms the case discussed in Sect. 3.2.

7 Concluding remarks

For networks satisfying the response-based reliability criteria we have a high level of outlier identifiability, which is due to small correlations between w-variables.

By supplementing the MDB concept with the identifiability index one may evaluate at an a priori analysis the probability of identifying a gross error of MDB magnitude in the first adjustment run. For the observations with small identifiability indices one may find the magnitude of gross error, greater than MDB, necessary to ensure a satisfactory level of identifiability. While setting a certain requirement for the probability level, e.g. $\mathrm{ID} \geq 0.95$, as well as for the level of mis-identification, e.g. $\overline{\mathrm{MID}}_{ij,\max} < 0.02$, the proposed approach can be used in optimizing networks with respect to internal reliability. The identifiability index can also be useful in explaining discrepancies between the MDB values and the actual results of outlier identification. The significant discrepancies reported in some papers do not indicate weaknesses of the MDB concept but are a result of incorrect treating the magnitudes of actually identified gross errors as the quantities equivalent to the corresponding MDBs.

The proposal requires further clarification in terms of the theoretical basis and testing on a wider range of observation systems. That would allow one to determine the optimal sample size for the simulation method and the actual accuracy of empirical estimates. Also the working program used for this purpose needs optimization to reduce the operation time. The relationship between the indices ID_i^* and ID_i deserves a more in-depth analysis.

The similar issue of reliable identification of outliers, termed as outlier separability (Wang et al. 2012) and (Yang et al. 2013), slightly touched in this paper, can serve as future reference for more detailed comparative analysis of the approach proposed herein.

The question of analogous approach for the case of multiple outliers is a much more complicated problem and is planned to be a topic of the next research. In seeking solution an interesting concept of maximum MDB (Knight et al. 2010) will be taken into account. Also the approach to correlation between multiple outlier detection statistics (i.e. the use of maximum correlation and global correlation coefficients) as in (Wang et al. 2012), will play an important role in shaping the strategy of further research.

Acknowledgments The author wishes to thank Prof. Kim Hao of Dongguk University for some valuable guidelines concerning dependence case in the theory of ratio of two folded normal variables. The guidelines were helpful in constructing a probabilistic measure for partial pseudo-identifiability. The author feels greatly indebted to Prof. Mieczysław Kwaśniak and Dr. Artur Wilkowski of Warsaw University of Technology for their assistance in performing the computations and creating the graphs.

Appendix A

1. Condition for existence of RUE

Let us consider the following w-variables as in (Knight et al. 2010),

- with a gross error $\Delta y_{s,i}$ in the i-th observation:

$$w_{i(i)} = \frac{\hat{z}_{i(i)}}{\sigma_{\hat{z}_{i(i)}}};\ w_{j(i)} = \frac{\hat{z}_{j(i)}}{\sigma_{\hat{z}_{j(i)}}}$$

- with a gross error $\Delta y_{s,j}$ in the j-th observation:

$$w_{i(j)} = \frac{\hat{z}_{i(j)}}{\sigma_{\hat{z}_{i(j)}}};\ w_{j(j)} = \frac{\hat{z}_{j(j)}}{\sigma_{\hat{z}_{j(j)}}}$$

Applying the formula (10), we get the corresponding pairs of relationships

$$w_{i(i)} = \frac{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{i*}}{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}} \cdot (\mathbf{e}_{\mathrm{S}} + \Delta y_{\mathrm{S}(i)})$$

$$w_{j(i)} = \frac{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{j*}}{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}} \cdot (\mathbf{e}_{\mathrm{S}} + \Delta y_{\mathrm{S}(i)})$$

$$w_{i(j)} = \frac{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{i*}}{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}}} \cdot (\mathbf{e}_{\mathrm{S}} + \Delta y_{\mathrm{S}(j)})$$

$$w_{j(j)} = \frac{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{j*}}{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj}}} \cdot (\mathbf{e}_{\mathrm{S}} + \Delta y_{\mathrm{S}(j)})$$

where $\{\cdot\}_{i*}$, $\{\cdot\}_{j*}$ are the i-th and the j-th row of $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$.

Requiring that $|w_{i(i)}| = |w_{j(i)}|$ and $|w_{i(j)}| = |w_{j(j)}|$, and taking into account that the w-variables can be of the same or opposite signs, we get after simple operations the condition

$$\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{i*} = \pm\frac{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{ii}}}{\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{jj}}}\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{j*} \quad (25)$$

i.e. the i-th and the j-th row in $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$ are linearly dependent vectors with positive or negative coefficient.

Applying (25) to pairs of the corresponding elements in the i-th and the j-th row, we get the following condition for the i-th and the j-th observation to be a RUE region

$$\sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{ii}} \cdot \sqrt{\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{jj}} = \left|\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{ij}\right| \quad (26)$$

or, equivalently

$$\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{ii}\cdot\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{jj}=\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{s}}^{-1}\mathbf{H}\right\}_{ij}^{2} \tag{27}$$

The above reasoning can be extended upon several observations in a network.

It follows immediately from (25), that the networks with rank $(\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}) = 1$, i.e. where all the rows and columns are linearly dependent, are as a whole RUE regions, irrespective of the correlation matrix used.

Correlation of w-variables within RUE is represented by a submatrix (or a matrix) with non-diagonal elements being $|\rho_{ij}| = 1$.

2. Condition excluding the existence of RUE

For $(h_{ii}, w_{ii}) \in \mathrm{S}_O, i = 1, \ldots, n$, we have, $0.5 < h_{ii} \leq 1 \wedge 0 < k_i < 1$.

Since

$$k_i = \frac{Q_{(i)}^2}{h_{ii}^2} \quad (\text{for } h_{ii} \neq 0), \quad \text{where} \quad Q_{(i)}^2 = \sum_{q=1, q\neq i}^{n} h_{qi}^2 \tag{28}$$

we obtain $|h_{qi}| \leq 0.5$ for $q \neq i$, $i = 1, \ldots, n$. Hence, for any pair of observations, e.g. $\mathrm{y}_i, \mathrm{y}_j$, we get

$$h_{ii}\cdot h_{jj} > h_{ij}\cdot h_{ji}. \tag{29}$$

Since $\mathbf{H}\cdot\mathbf{A}_{\mathrm{S}} = \mathbf{0}$, and hence $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\cdot\mathbf{A}_{\mathrm{S}} = \mathbf{0}$, the inequality (29) implies that $\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}$, being of the same rank as $\mathbf{H}$, has all determinants of the 1st and 2nd order positive, so

$$\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ii}\cdot\left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{jj} > \left\{\mathbf{H}^{\mathrm{T}}\mathbf{C}_{\mathrm{S}}^{-1}\mathbf{H}\right\}_{ij}^{2}$$
$$i, j = 1, \ldots, n, i \neq j \tag{30}$$

which contradicts the condition (27) for the existence of RUE in a network.

Appendix B

Probabilistic tool for a priori analysis of outlier partial pseudo-identifiability

Let us consider two independent normal variables $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ and the corresponding folded normal variables $|X_1| \sim \mathrm{FN}(\mu_1, \sigma_1^2)$, $|X_2| \sim \mathrm{FN}(\mu_2, \sigma_2^2)$. The ratio $\mathrm{Z} = \frac{|X_1|}{|X_2|}$ has distribution $\mathrm{Z} \sim \mathrm{RFN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ with distribution function $F(z)$, $z > 0$ (Kim 2006). The generalization of the approach for dependency case is presented in (Kim 2014), where the distribution function of $\mathrm{Z} \sim \mathrm{RFN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ is determined, valid for $|\rho| < 1$.

Assuming $\sigma_1 = 1$, $\sigma_2 = 1$ and $z = 1$, as needed for the analysis in the present paper, we obtain on the basis of the above mentioned generalized distribution function a formula for $\mathrm{P(Z} < 1)$ as a function of μ_1, μ_2, ρ, i.e.

$$\mathrm{P(Z} < 1) = 2L(\mathrm{h}_1', -\delta', \rho_1') + 2L(\mathrm{h}_2', \overset{'}{\delta}, \rho_2')$$
$$+\Phi(\mathrm{h}_1') + \Phi(\mathrm{h}_2') - 2 \tag{31}$$

where

$$\mathrm{h}_1' = \frac{\mu_1 - \mu_2}{\sqrt{2(1-\rho)}}; \quad \mathrm{h}_2' = \frac{\mu_1 + \mu_2}{\sqrt{2(1+\rho)}}; \quad \delta' = \mu_2;$$

$$\rho_1' = \frac{\sqrt{(1-\rho)}}{\sqrt{2}}; \; \rho_2' = \frac{\sqrt{(1+\rho)}}{\sqrt{2}}$$

$$L(a, b, \rho) = \mathrm{P}(X > a, Y > b);$$
$$X, Y \sim N(0,\ 1); \quad \rho \quad (|\rho| < 1)$$

$\Phi(a) = \mathrm{P}(X < a)$; $X \sim N(0, 1)$; $L(a,\ b,\ \rho)$ can be equivalently replaced by $\Phi(-a, -b, \rho)$.

Several properties of the function $\mathrm{P(Z} < 1) = f(\mu_1,\ \mu_2,\ \rho)$ as in (31) concerning the signs of its arguments, can be readily proved, e.g.

$$f(\mu_1 > 0,\ \mu_2 > 0,\ \rho > 0) = f(\mu_1 < 0,\ \mu_2 < 0,\ \rho > 0)$$
$$f(\mu_1 > 0,\ \mu_2 > 0,\ \rho < 0) = f(\mu_1 < 0,\ \mu_2 < 0,\ \rho > 0)$$

The properties can be helpful in simplifying tables or diagrams constructed for the function.

We present a property corresponding to basic formulas used in identifiability analysis (14), Sect. 3.1, where $\mu_1 > 0$ and μ_2 is of the same sign as ρ

$$f(\mu_1 > 0, \mu_2 > 0, \rho > 0) = f(\mu_1 > 0, \mu_2 < 0, \rho < 0) \tag{32}$$

We can prove this property by substituting in right-hand side function of the above equality $\mu_2^* = -\mu_2$ and $\rho^* = -\rho$ instead of μ_2 and ρ respectively, and finally finding that

$$(\mathrm{h}_1')^* = \mathrm{h}_2'; \quad (\mathrm{h}_2')^* = \mathrm{h}_1'; \quad (\delta')^* = -\delta'; \quad (\rho_1')^* = \rho_2';$$
$$(\rho_2')^* = \rho_1'$$

which are exactly the components of the formula (31) determining the value of the left-hand side function.

For computing the values of the function $\mathrm{P(Z} < 1) = f(\mu_1, \mu_2, \rho)$, a software based on the MATLAB package was developed.

On the basis of computations we may list some important properties useful for interpretation of the results of identifiability analysis. To visualize the properties we show a graph of the function $\mathrm{P}(Z < 1) = f(\mu_1,\ \mu_2,\ \rho)$ for $0 \leq \mu_1 \leq 6$,

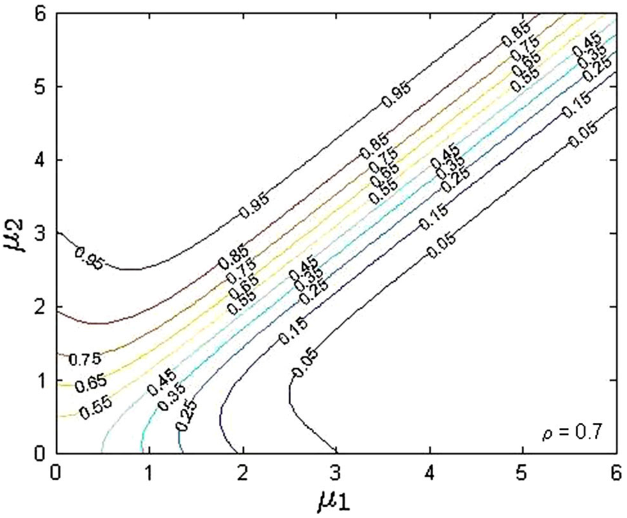

Fig. 5 Graph of the function $P(Z < 1)$ for chosen values of parameters

$0 \le \mu_2 \le 6$, $\rho = 0.7$ (Fig. 5):

- for $|\mu_2| > |\mu_1|$, we have $\mathrm{P(Z < 1)} > 0.5$; $|\rho| < 1$

The greater the difference $|\mu_2| - |\mu_1|$, the greater is $\mathrm{P(Z < 1)}$.

- for $|\mu_2| = |\mu_1|$, we have $\mathrm{P(Z < 1)} = 0.5$; $|\rho| < 1$
- for $|\mu_2| < |\mu_1|$, we have $\mathrm{P(Z < 1)} < 0.5$; $|\rho| < 1$

The greater the difference $|\mu_1| - |\mu_2|$, the smaller is $\mathrm{P(Z < 1)}$.

A specific case, when $|\mu_1| = |\mu_2|$, $|\rho| = 1$, cannot be analyzed with the use of formula (31), since the distribution function of $\mathrm{Z} \sim \mathrm{RFN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ is not valid for $|\rho| = 1$. In this case we have $|X_1| = |X_2|$ with probability $\mathrm{P} = 1$. and hence $\mathrm{Z} = \frac{|X_1|}{|X_2|} = \frac{|X_1|}{|X_1|} = 1$ (Z becomes a constant), so $\mathrm{P(Z < 1)} = 0$

References

Baarda W (1968) A testing procedure for use in geodetic networks. Publ Geod New Ser 2(5). Netherlands Geodetic Commission, Delft

Caspary WF (1988) Concepts of network and deformation analysis. Monograph 11, School of Surveying, The University of New South Wales, Kensington

Cen M, Li Z, Ding X, Zhuo J (2003) Gross error diagnostics before least squares adjustment of observations. J Geod 77:503–513

Förstner W (1983) Reliability and discernability of extended Gauss–Markov models. Deutsche Geodätische Kommission, Reihe A, No. 98, Munchen

Hawkins DM (1980) Identification of outliers. Chapman and Hall, New York

Hekimoglu S, Erenoglu RC (2005) A test for Baarda's internal reliability theory. In: Proceedings of international symposium on "Modern technologies, education and professional practice in geodesy and related fields", Sofia, Bulgaria

Kim H-J (2006) On the ratio of two folded normal distributions. Commun Stat Theory Methods 35:965–977

Kim H-J (2014) Some distributional properties of ratio of two folded normals. Technical report, Statistics Department, Dongguk University, Seoul, Korea

Knight NL, Wang J, Rizos C (2010) Generalised measures of reliability for multiple outliers. J Geod. doi:10.1007/s00190-010-0392-4

Prószyński W (2008) The vector space of imperceptible observation errors: a supplement to the theory of network reliability. Geod Cartogr 57(1):3–19

Prószyński W (2010) Another approach to reliability measures for systems with correlated observations. J Geod 84:547–556

Schaffrin B (1997) Reliability measures for correlated observations. J Eng Surv 123:126–137

Teunissen PJG (1990) Quality control in integrated navigation systems. IEEE Aerosp Electron Syst Mag 5(7):35–41

Teunissen PJG (1996) Testing theory, an introduction. Delft University Press, Delft

Teunissen PJG (1998) Minimal detectable biases of GPS data. J Geod 72:236–244

Teunissen PJG (2000) Testing theory: an introduction. Delft University Press, Delft

Wang J, Chen Y (1994) On the reliability measure of observations. Acta Geodaet et Cartograph Sin Engl Edn 42–51

Wang J, Knight N (2012) New outlier separability test and its application in GNSS positioning. J Glob Position Syst 11(1):46–57

Wang J, Almagbile Y, Wu T, Tsujii T (2012) Correlation analysis for fault detection statistics in integrated GNSS/INS systems. J Glob Position Syst 11(2):89–99

Yang L, Wang J, Knight N, Shen Y (2013) Outlier separability analysis with a multiple alternative hypotheses test. J Geod 87(6):591–604