Survey Review

ISSN: 0039-6265 (Print) 1752-2706 (Online) Journal homepage: https://www.tandfonline.com/loi/ysre20

# A half-century of Baarda's concept of reliability: a review, new perspectives, and applications

Vinicius Francisco Rofatto, M. T. Matsuoka, I. Klein, M. R. Veronez, M. L. Bonimani & R. Lehmann

**To cite this article:** Vinicius Francisco Rofatto, M. T. Matsuoka, I. Klein, M. R. Veronez, M. L. Bonimani & R. Lehmann (2018): A half-century of Baarda's concept of reliability: a review, new perspectives, and applications, Survey Review, DOI: 10.1080/00396265.2018.1548118

**To link to this article:** https://doi.org/10.1080/00396265.2018.1548118

# A half-century of Baarda's concept of reliability: a review, new perspectives, and applications

**Vinicius Francisco Rofatto** *[1,2], **M. T. Matsuoka**[1,2], **I. Klein**[3,4], **M. R. Veronez**[5], **M. L. Bonimani**[1] and **R. Lehmann**[6]

Over the 50 years of its existence, Baarda's concept of reliability has been used as a standard practice for the quality control in geodesy and surveying. In this study, we analysed the pioneering work of Baarda (Publ Geod New Ser 2(4) 1967; Publ Geod New Ser 2(5) 1968) and recent studies on the subject. We highlighted that the advent of personal computers with powerful processors has rendered Monte Carlo method as an attractive and cost-effective approach for quality control purposes. We also provided an overview of the latest advances in the reliability theory for geodesy, with particular emphasis on Monte Carlo method.

**Keywords:** Reliability, Outlier, Hypothesis Testing, Quality Control, Minimal Detectable Bias, Minimal Identifiable Bias, Data snooping

## Inspiration and motivation

In many geodetic analysis tasks, large sets of observations are being recorded or sampled. It is nearly impossible that such datasets are free from outliers. Thus, it is important to identify outlying observations that may lead to model misspecification, biased parameter estimation, and incorrect results. Among the many definitions found in the literature on what an outlier is, we follow the statement given by Lehmann (2013): 'an outlier is an observation that is so probably caused by a gross error that it is better not used or not used as it is'. In geodesy, outliers are most often caused by gross errors and gross errors most often cause outliers (Lehmann 2013).

Over the 50 years of its existence, Baarda's (1967, 1968) groundbreaking work on outlier testing has been used as a standard practice for the quality control in geodesy and surveying. As part of his contribution, he proposed a procedure based on hypothesis testing for the detection of a single outlier in linear(ised) models, which he called data snooping. Most of conventional geodetic studies have a chapter on Baarda's data snooping procedure, e.g. Koch (1999) and Teunissen (2006). Furthermore, this procedure has become very popular and is routinely used in adjustment computations (Ghilani 2010). Although data snooping was introduced as a testing procedure for use in geodetic networks, it is a generally applicable method (Lehmann 2012).

Baarda's data snooping consists of screening each individual observation for a possible outlier (Kok 1984, Teunissen 2006). Baarda's $w$-test statistic for his data snooping is given by a normalised least-squares residual. This test, which is based on a linear mean-shift model, can also be derived as a particular case of the generalised likelihood ratio test (Kok 1984, Koch 1999, Teunissen 2006). In principle, Baarda's $w$-test only makes a decision between the null $\mathcal{H}_0$ and a single alternative hypothesis $\mathcal{H}_i$. In that case, rejection of $\mathcal{H}_0$ automatically implies acceptance of $\mathcal{H}_i$, and vice versa (Teunissen 2017, Imparato *et al.* 2018). Therefore, it concerns on outlier detection and not on outlier identification (Teunissen *et al.* 2017).

Based on the probability of rejecting a true null hypothesis (type I error – 'false alarm', denoted by $\alpha$) and the probability of rejecting a true alternative hypothesis (type II error - 'missed detection', denoted by $\beta$), Baarda (1968) also derived the concept of a marginally detectable outlier (Kok 1984), which is commonly called as the minimal detectable bias (MDB)—the term given by Teunissen (1989). The MDB of an alternative hypothesis is the smallest outlier that can lead to rejection of a null hypothesis for a given $\alpha$ and $\beta$. Thus, for each alternative hypothesis $\mathcal{H}_i$, the corresponding MDB can be easily computed. The set of MDBs describes the internal reliability, whereas their propagation into the parameters is said to describe the external reliability (Baarda 1968, Vaníček *et al.* 2001, Teunissen 2006).

Different from the case of the single alternative hypothesis above, Förstner (1983) developed the separability analysis between two alternative hypotheses, e.g. $\mathcal{H}_i$ and $\mathcal{H}_j$. In this case, rejection of $\mathcal{H}_0$ does not necessarily imply the correct identification of a particular alternative hypothesis. Consequently, there is also a probability of correctly rejecting the null hypothesis but wrongly accepting an alternative hypothesis. This type of wrong decision is called type III error (Hawkins 1980) or 'wrong

[1]Quality Control Lab (QCGeo), Geography Institute, Federal University of Uberlandia, Monte Carmelo, Brazil
[2]Graduate Program in Remote Sensing, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[3]Land Surveying Program, Federal Institute of Santa Catarina (IFSC), Florianopolis, Brazil
[4]Graduate Program in Geodetic Sciences, Federal University of Parana, Curitiba, Brazil
[5]Advanced Visualization Laboratory (Vizlab), Graduate Program in Applied Computing, Unisinos University, Sao Leopoldo, Brazil
[6]Faculty of Spatial Information, University of Applied Sciences Dresden, Dresden, Germany

*Corresponding author, email vfrofatto@gmail.com
This article has been republished with minor changes. These changes do not impact the academic content of the article.


Received 3 August 2018; accepted 8 November 2018
DOI 10.1080/00396265.2018.1548118

exclusion' (indicated by '$\kappa$' in this paper). In this respect, we discuss the identification of an outlier. For identification, the correlation coefficient (here, denoted by $\rho$) between Baarda's $w$-test should be considered during the data snooping procedure (Förstner 1983, Wang and Knight 2012, Yang *et al.* 2013, Prószyński 2015).

Investigations on outlier identification involve separability analyses between alternative hypotheses. Generally, a separability analysis reveals the degree to which the alternative hypotheses can be distinguished. Van Mierlo's (1975, 1979) studies were the first to show the possibility of describing the separability of alternative hypotheses on the so-called wrong exclusion in deformation monitoring. The current outlier separability theory is credited to Yang *et al.* (2013). They have extended the separability analysis from two alternative hypotheses (Förstner 1983, Li 1986) to multiple alternative hypotheses. Thus, for a system with $n$ measurements, the null hypothesis $\mathcal{H}_0$ should be tested against all $n$ alternative hypotheses.

Recently, Prószyński (2015) provided supplementation of the MDB by an outlier identifiability index to each observation and misidentifiability index as the maximum probability of wrong exclusion rate. Teunissen *et al.* (2017) introduced the concept of minimal identifiable bias (MIB) as the smallest outlier of an alternative hypothesis that can lead to its identification for a given probability. Imparato *et al.* (2018) showed that, under multiple testing, the MIBs should be larger than their MDBs in order to guarantee a successful outlier identification. There is no difference between MDB and MIB when only a single alternative hypothesis is involved (Teunissen *et al.* 2017, Imparato *et al.* 2018). This means that an outlier with a magnitude of MDB will hardly be identified. In $n$ alternative hypotheses, therefore, the MDB should increase with the correlation coefficient (Yang *et al.* 2013).

With regard to identifiability, Zaminpardaz and Teunissen (2018) presented the data snooping in the context of the DIA method for the detection, identification and adaptation of the mismodelling errors. In this study, the central role that is played by the partitioning of misclosure space, both in the formation of the decision probabilities and in the construction of the DIA estimator. They highlighted their difference by showing the difference between their corresponding contributing factors. They also demonstrated that the correct detection probabilities of the alternative hypotheses is not necessarily the same as that of their correct identification probabilities. And they showed, if two alternative hypotheses, say $\mathcal{H}_i$ and $\mathcal{H}_j$ are not distinguishable, that the testing procedure may have different levels of sensitivity to $\mathcal{H}_i$-biases compared to the same and $\mathcal{H}_j$-biases.

The articles quoted above, i.e. Yang *et al.* (2013), Prószyński (2015), Imparato *et al.* (2018), and Zaminpardaz and Teunissen (2018) have already pointed out that the probability density functions (pdf) under the multiple alternative hypotheses are complex and even practically impossible to obtain in a closed form. In that case, the critical region of the outlier test is also complicated. The critical region is the subset of the observations, for which the null hypothesis is rejected. Consequently, they used the Monte Carlo method (MC) to analyse their analytical expression.

The MC is a well known method of numerical integration. Different from determinist approach such as the trapezoidal rule, MC incorporates randomness by sampling random values from specified distributions. The basic idea is to approximate probability distributions by frequency distributions of computer random experiments performed using pseudo-random number generators (Koch 2007, Lehmann 2012). A computation with one set of pseudo-random number generators is a Monte Carlo experiment (Lehmann and Scheffler 2011). For more details about the MC, see, for example, Altiok and Melamed (2007), Robert and Casella (2013), Gamerman and Lopes (2006).

One can argue on the disadvantage of high computational complexity involved in a MC approach. In contrast to Baarda's era, we currently have fast computers and large data storage systems at our disposal. Therefore, such obstacles are no longer existing because computing power is not a bottleneck at present (Lehmann 2015). In this context, the MC has been extensively applied in geodesy (Lehmann 1994, Hekimoglu and Koch 1999, Koch 2007, Lehmann and Scheffler 2011, Aydin 2012, Lehmann 2012, Yang *et al.* 2013, Marx 2015, Prószyński 2015, Koch 2016, Klein *et al.* 2017, Lehmann and Voß-Böhme 2017, Rofatto *et al.* 2017, Imparato *et al.* 2018).

The situation becomes a little more complicated when Baarda's data snooping is applied in its iterative form. This procedure is called iterative data snooping (IDS) (Teunissen 2006, p. 135). In contrast to the well-defined theories (Baarda 1968, Förstner 1983, Wang and Knight 2012, Yang *et al.* 2013, Prószyński 2015, Imparato *et al.* 2018), we believe that the IDS procedure is a heuristic method, and therefore, there is no complete and rigorous reliability theory on it. Heuristics are understood as techniques designed for finding approximate solutions when classic methods fail to find any exact solution. In such case, an analytical model with a tractable solution for its probabilities is unknown, and therefore, one needs to resort to MC.

The purpose of this study is to provide an overview of the latest advances in the reliability theory for geodesy. Here, we analysed the pioneering work of Baarda (1968) and the recent studies on the subject (Lehmann 2012, Yang *et al.* 2013, Prószyński 2015, Lehmann and Lösler 2016, Lehmann and Voß-Böhme 2017, Teunissen *et al.* 2017, Imparato *et al.* 2018). Then, we pointed out new possibilities and perspectives on the issue, with a particular emphasis on the MC.

The paper is organised as follows. First, we provide the elements of Baarda's original data snooping testing procedure and its restrictions when there are multiple alternative hypotheses. Second, we demonstrate by an illustrative example how IDS works, and we point out that its probabilities should be computed by a MC. Third, we present a MC approach as an alternative for quality control. With regard to the MC, two problems are presented and we try to answer them. First, how can we find the optimal number of MC experiments for quality control purpose? Second, how do we obtain the probability levels of IDS? Finally, the concluding remarks are summarised at the end of this paper.

## Baarda's data snooping for outlier detection

### Null hypothesis versus alternative hypothesis: a binary case

The null hypothesis, which is also called the working hypothesis, corresponds to a supposedly valid model

describing the physical reality of the observations without the presence of an outlier. When it is assumed to be 'true', this model is used to estimate the unknown parameters, typically in a least-squares approach. Thus, the null hypothesis of the standard Gauss–Markov model in linear or linearised form is given by (Koch 1999)

$$\mathcal{H}_0{:}\mathrm{E}\{\boldsymbol{y}\} = Ax,\ \mathrm{D}\{\boldsymbol{y}\} = \boldsymbol{\Sigma}_{yy} \tag{1}$$

where $\mathrm{E}\{.\}$ is the expectation operator, $y \in \mathbb{R}^n$ the vector of measurements, $\boldsymbol{A} \in \mathbb{R}^{n\times u}$ the Jacobian matrix (also called design matrix) of full rank $u$, $x \in \mathbb{R}^u$ the unknown parameter vector, $\mathrm{D}\{.\}$ the dispersion operator, and $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}^{n\times n}$ the known positive definite covariance matrix of the measurements. In this case, the redundancy $r$ of the model in equation (1) is $r = n - u$.

Instead of $\mathcal{H}_0$, Baarda (1968) proposed a mean shift alternative hypothesis $\mathcal{H}_\mathrm{A}$, also referred to as model misspecification by Teunissen (2006), as follows:

$$\mathcal{H}_\mathrm{A}{:}\mathrm{E}\{\boldsymbol{y}\} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{C}\nabla,\ \mathrm{D}\{\boldsymbol{y}\} = \boldsymbol{\Sigma}_{yy} \tag{2}$$

with $\boldsymbol{C} \in \mathbb{R}^{n\times q}$ as the matrix of outlier model and $\nabla \in \mathbb{R}^q$ the unknown outlier vector. Under $\mathcal{H}_\mathrm{A}$, we restrict ourselves to $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}\end{bmatrix} \in \mathbb{R}^{n\times(u+q)}$ having a full column rank, such that $q \leq n - u = r$. The alternative to this $\mathcal{H}_\mathrm{A}$ is the variance inflation $\mathcal{H}_\mathrm{A}$. Up to now it is rarely used in geodesy because it is more difficult to derive a powerful test and a reliability theory for it.

The number of unknown outliers defines the alternative model, i.e.

$$\overbrace{i}^{\text{local test}} \leq q \leq \overbrace{n-u}^{\text{overall test}} \tag{3}$$

When $q = n - u = r$, an overall model test is performed. In that extreme case, the redundancy of the alternative hypothesis $\mathcal{H}_\mathrm{A}$ is $r = n - (u + q) = n - (u + n - u) \therefore r = 0$, and therefore, there is no need to specify any particular alternative hypothesis. In this case, the overall model test is based on the weighted sum-of-squared residuals $\|\hat{e}_0\|^2$, i.e.

$$\|\hat{e}_0\|^2 = \hat{\boldsymbol{e}}_0^\mathrm{T}\boldsymbol{\Sigma}_{yy}^{-1}\hat{\boldsymbol{e}}_0 \tag{4}$$

with $\hat{\boldsymbol{e}}_0$ being the least-squares residuals vector under $\mathcal{H}_0$. By setting an acceptance region $\mathcal{A}\alpha_0 \subset \mathbb{R}^n$ (Baarda 1968, Teunissen 2006),

$$\mathcal{A}\alpha_0 = \{y \in \mathbb{R}^n \mid \|\hat{e}_0\|^2 \leq \chi^2_{\alpha_0}(r, 0)\} \tag{5}$$

Thus, the overall test is given as

$$\text{Accept } \mathcal{H}_0 \text{ if } \|\hat{\boldsymbol{e}}_0\|^2 \leq \chi^2_{\alpha_0}(r, \lambda q), \text{ reject otherwise} \tag{6}$$

The critical value $\chi^2_{\alpha_0}(r, \lambda q)$ is computed from the central chi-squared distribution with $r = n - u$ degrees of freedom and type I error, also known as false alarm or level of significance, $\alpha_0$ (note: the index '0' represents the case of a single alternative hypothesis testing). The second argument of $\chi^2_{\alpha_0}$ is the non-centrality parameter $(\lambda q)$. The non-centrality parameter $\lambda q$ describes the discrepancy between $\mathcal{H}_0$ and $\mathcal{H}_\mathrm{A}$. The performance of that test can be described by its probabilities of committing type I errors and type II errors (also known as missed detection rate, $\beta_0$) as follows:

$$\begin{aligned}\text{Type I error: } &\mathrm{P}[\boldsymbol{y} \notin \mathcal{A}\alpha_0 \mid \mathcal{H}_0 = \textbf{\textit{true}}] = \alpha_0\\ \text{Type II error: } &\mathrm{P}[\boldsymbol{y} \in \mathcal{A}\alpha_0 \mid \mathcal{H}_0 = \textbf{\textit{false}}] = \beta_0\end{aligned} \tag{7}$$

The probability of type I error $\alpha_0$ is the probability of rejecting the null hypothesis when it is true, whereas the type II error $\beta_0$ is the probability of failing to reject the null hypothesis when it is false. The complement of type II error is the well-known power of the test (also known as correct detection) given by $\gamma_0 = 1 - \beta_0$.

For the overall model test, the weighted sum-of-squared residuals under $\mathcal{H}_\mathrm{A}$ follow the non-central chi-squared distribution with $r = n - u$, i.e.

$$\begin{aligned}\|\hat{e}_A\|^2 &\sim \chi^2_{\alpha_0}(r, \lambda q), \text{ where } \lambda q\\ &= \nabla^T\boldsymbol{C}^T\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{\hat{e}_0}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{C}\nabla\end{aligned} \tag{8}$$

where $\hat{\boldsymbol{e}}_\mathrm{A}$ is the least-squares residuals vector of $\mathcal{H}_0$ which has this distribution under $\mathcal{H}_\mathrm{A}$, $\lambda q$ the non-centrality parameter, and $\boldsymbol{\Sigma}_{\hat{e}_0}$ the variance matrix of the best linear unbiased estimator of $\hat{\boldsymbol{e}}_0$ under $\mathcal{H}_0$. We restrict ourselves to the data snooping by Baarda (1968). For more details about the overall model test, see Teunissen *et al.* (2017) and Imparato *et al.* (2018).

On the other hand, Baarda (1968) proposed a procedure to detect a single outlier in linear(ised) models, which he called data snooping. The purpose of his data snooping procedure is to screen each individual observation for an outlier. Therefore, in contrast to the overall model test, Baarda's data snooping is based on a local model test. In such case, $q = 1$ and the $n$ alternative hypothesis is given by

$$\begin{aligned}\mathcal{H}_i{:}\mathrm{E}\{\boldsymbol{y}\} &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}_i\nabla i\ \ (\forall i = 1, \ldots, n),\ \mathrm{D}\{\boldsymbol{y}\}\\ &= \boldsymbol{\Sigma}_{yy}\end{aligned} \tag{9}$$

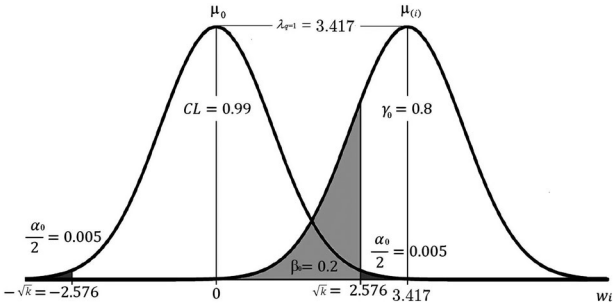
Now, the matrix $\boldsymbol{C}$ in equation (2) is reduced to a canonical unit vector $\boldsymbol{c}_i$, which consists exclusively of elements with values of 0 and 1, where 1 means that an $i$th outlier of magnitude $\nabla i$ affects an $i$th measurement and 0 otherwise. In that case, the rank $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{c}_i\end{bmatrix} \in \mathbb{R}^{n\times(u+1)}$ and the vector $\nabla$ in equation (2) reduces to a scalar $\nabla i$ in equation (9), i.e. $\boldsymbol{c}_i = \begin{bmatrix}0 & 0 & \ldots & \overbrace{1}^{i\text{th}} & 0 & \ldots & 0\end{bmatrix}^T$.

Baarda's $w$-test statistic for his data snooping is given by a normalised least-squares residual as follows (Baarda 1968):

$$wi = \frac{\boldsymbol{c}_i\boldsymbol{\Sigma}_{yy}^{-1}\hat{\boldsymbol{e}}_0}{\sqrt{\boldsymbol{c}_i{}^T\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{\hat{e}_0}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{c}_i}}\ \ (\forall i = 1, \ldots, n) \tag{10}$$

To verify if there are sufficient evidences to reject or not the null hypothesis, the test for binary case should be performed as

$$\begin{aligned}&\text{Accept } \mathcal{H}_0\\ &\text{if } |wi| \leq \sqrt{\chi^2_{\alpha_0}(q = 1, 0)}\\ &\qquad = \sqrt{k}, \text{ reject otherwise in favour of } \mathcal{H}_i\end{aligned} \tag{11}$$




**1 A non-centrality parameter of $\lambda_{q=1} = 3.417$ with $\alpha_0 = 0.01$ ($\sqrt{k} = 2.576$) lead to $\gamma_0 = 0.8$ (or $\beta_0 = 0.2$)**

with the acceptance region given as

$$\mathcal{A}^i_{\alpha_0} = \left\{ \boldsymbol{y} \in \mathbb{R}^n \mid |wi| \leq \sqrt{k} \right\} \tag{12}$$

with $k$ being the critical value from the chi-squared one-side distribution with $q = 1$ degree of freedom (note: the square root of a chi-squared one-side test is equivalent to that of a standard normal distribution two-tailed test). For example, for $\alpha_0 = 0.01$, we obtain $\sqrt{k} = 2.576$. In this case, if $|wi| > 2.576$ for some $yi$, one may reject $\mathcal{H}_0$ because $yi \notin \mathcal{A}^i_{\alpha_0}$.

Because Baarda's $w$-test in its essence is based on binary hypothesis testing, in which one decides between the null hypothesis $\mathcal{H}_0$ and a unique alternative hypothesis $\mathcal{H}_i$ of equation (9), it may lead to type I errors ($\alpha_0$) and type II errors ($\beta_0$). Instead of $\alpha_0$ and $\beta_0$, there is the confidence level (CL) and power of the test ($\gamma_0$), respectively. The first deals with the probability of accepting a true null hypothesis; the second, as already mentioned, with the probability of correctly accepting the alternative hypothesis.

The normalised least-squares residual $wi$ follows a standard normal distribution with the expectation that $\mu_0 = 0$ if $\mathcal{H}_0$ holds true (there is no outlier). On the other hand, if the system is contaminated with a single outlier at the $i$th position of the dataset (i.e. under $\mathcal{H}_i$), there is an outlier that causes the expectation of $wi$ to become $\lambda_{q=1}$(i.e. $\mu_i = \lambda_{q=1}$), which is the non-centrality parameter for $q = 1$. In this case, the non-centrality parameter $\lambda_{q=1}$ is given by

$$\lambda_{q=1} = \boldsymbol{c}_i{}^T \boldsymbol{\Sigma}^{-1}_{yy} \boldsymbol{\Sigma}_{\hat{e}_0} \boldsymbol{\Sigma}^{-1}_{yy} \nabla i^2 \tag{13}$$

Figure 1 shows the relationship among the parameters $\alpha_0$, $\beta_0$, CL, $\gamma_0$, $\lambda_{q=1}$, and $\sqrt{k} = \sqrt{\chi^2_{\alpha_0}(q = 1, 0)}$. For a given value of $\lambda_{q=1}$, the smaller the significance level $\alpha_0$, the smaller the power of the test $\gamma_0$. On the other hand, the larger the non-centrality parameter $\lambda_{q=1}$, the larger the power $\gamma_0$ for a given $\alpha_0$.

The non-centrality parameter $\lambda_{q=1}$ in equation (13) describes the discrepancy between the null $\mathcal{H}_0$ and the single alternative hypothesis $\mathcal{H}_i$ in equation (9). In other words, the non-centrality parameter $\lambda_{q=1}$ represents the expected mean shift of a specific $w$-test when there is an outlier. In such case, the term $\boldsymbol{c}_i{}^T \boldsymbol{\Sigma}^{-1}_{yy} \boldsymbol{\Sigma}_{\hat{e}_0} \Sigma^{-1}_{yy} \boldsymbol{c}_i$ becomes a scalar and the solution of the quadratic equation (13) is given by (Teunissen 2006)

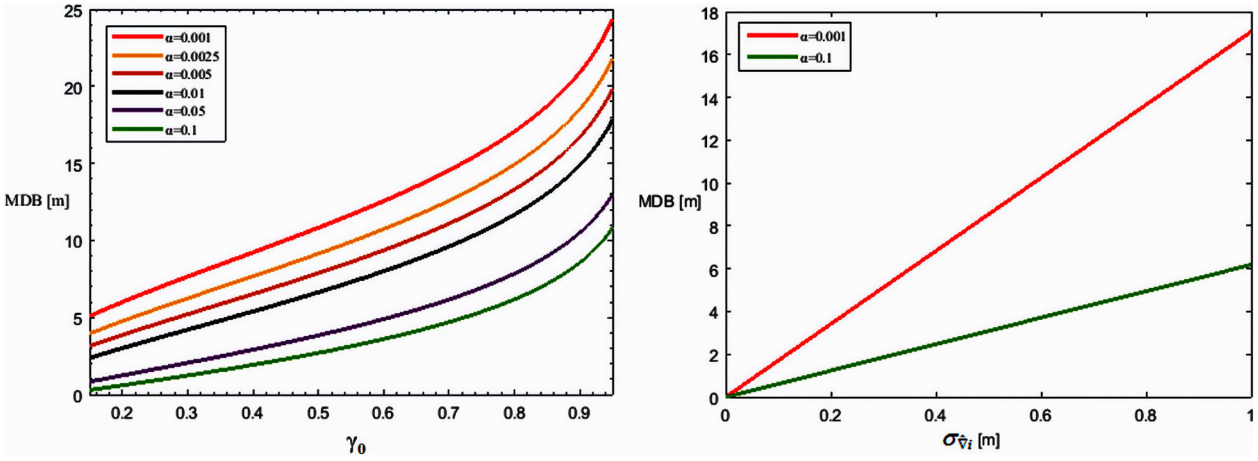
$$|\nabla i| = \mathrm{MDB}(i) = \sqrt{\frac{\lambda_{q=1}(\alpha_0,\ \gamma_0)}{\boldsymbol{c}_i{}^T \boldsymbol{\Sigma}^{-1}_{yy} \boldsymbol{\Sigma}_{\hat{e}_0} \Sigma^{-1}_{yy} \boldsymbol{c}_i}} \tag{14}$$

where $|\nabla i|$ is MDB($i$), which is computed for each of the $n$ alternative hypotheses according to equation (9).

Because

$$\sigma^2_{\hat{\nabla} i} = (\boldsymbol{c}_i{}^T \boldsymbol{\Sigma}^{-1}_{yy} \boldsymbol{\Sigma}_{\hat{e}_0} \Sigma^{-1}_{yy} \boldsymbol{c}_i)^{-1} \tag{15}$$

with $\sigma^2_{\hat{\nabla} i}$ being the variance of estimated outlier $\hat{\nabla} i$, MDB($i$) can also be computed by combining Eqs. (14)




**2 Left: MDB computed as a function of power of the test $\gamma_0$ for different values of type I error $\alpha_0$ (the estimation of the outlier precision was fixed as $\sigma_{\hat{\nabla} i} = 1$ m). Right: MDB as a function of $\sigma_{\hat{\nabla} i}$ for $\alpha_0 = 0.001$ and $\alpha_0 = 0.1$ (the power of the test was set as $\gamma_0 = 0.8$)**

and (15) as

$$\mathrm{MDB}(i) = \sigma_{\hat{\nabla} i}\sqrt{\lambda_{q=1}(\alpha_0,\ \gamma_0)},\ (\forall i = 1,\ \ldots,\ n) \quad (16)$$

where $\sigma_{\hat{\nabla} i}$ is the standard deviation of the estimated outlier $\hat{\nabla} i$.

Note that the non-centrality parameter $\lambda_{q=1}$ in equation (13) requires knowledge of the outlier size, which in practice is unknown. On the other hand, $\lambda_{q=1}$ can be computed as a function of $\alpha_0$, $\beta_0$, and $r = q = 1$. Baarda (1968) provided monograms for those interested in obtaining values as a function of $\alpha_0$, $\beta_0$, and $r$. Alternatively, we use the recursive algorithm based on the work by Aydin and Demirel (2005), namely Newton algorithm, in order to obtain the non-centrality parameter $\lambda_{q=1}$. The left side graph of Fig. 2 shows, for the case $q = 1$, that MDB becomes larger for smaller values of $\alpha_0$ and $\beta_0$ (or larger power of the test $\gamma_0$). It should also be noted that the more precise the estimated $\hat{\nabla} i$ (smaller $\sigma_{\hat{\nabla} i}$), the smaller the MDB (see right graph of Fig. 2).

The MDB was further investigated for a single outlier in a singular Gauss–Markov model (Wang and Chen 1999). There are also studies covering either independent or correlated measurements (Pelzer 1983, Prószyński 1994, Wang and Chen 1994, Schaffrin 1997, Teunissen 1998, Prószyński 2010).

For the case of $1 < q < n - u$, one assumes that more than one outlier is present in the dataset. In other words, it is possible to set up for the case of multiple outliers (Kok 1984, Ding and Coleman 1996, Teunissen 2006, Knight *et al.* 2010, Baselga 2011, Gui *et al.* 2011, Koch 2016, Klein *et al.* 2017). The multiple outlier case is covered at full length by Teunissen (2006) [For more details about alternative models, refer to Lehmann (2013) and Lehmann and Lösler (2016)].

Although Baarda's $w$-test belongs to the class of generalised likelihood ratio tests and has the property of being a uniformly most powerful invariant (UMPI) test when the null hypotheses is tested against a single alternative (Arnold 1981, Teunissen 2006, Kargoll 2007), this test may not necessarily be a UMPI when more than one alternative hypothesis are considered, as is the case of the data snooping procedure (Kargoll 2012, Lehmann and Voß-Böhme 2017, Teunissen *et al.* 2017).

## Baarda's data snooping for outlier identification

In this section, we will briefly review the two alternative hypotheses case (Förstner 1983) and the case of $n$ alternative hypotheses (Yang *et al.* 2013, Prószyński 2015, Teunissen *et al.* 2017, Imparato *et al.* 2018).

### Förstner method: two alternative hypotheses

After 15 years of existence, Baarda's theory was extended by Förstner (1983), who considered two alternative hypotheses instead of a single one. Under two alternative hypotheses, the non-centrality parameter is not only related to the sizes of $\alpha_0$ and $\beta_0$ as shown until the present, but it is also dependent on the correlation coefficient ($\rho$) between Baarda's $w$-test statistics, i.e. $wi$ and $wj$, as follows (Förstner 1983, Li 1986):

$$\rho_{wi,wj} = \frac{\boldsymbol{c}_i{}^T\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{\hat{e}_0}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{c}_j}{\sqrt{\boldsymbol{c}_i{}^T\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{\hat{e}_0}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{c}_i}\ \sqrt{\boldsymbol{c}_j{}^T\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{\hat{e}_0}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{c}_j}}(i \neq j) \quad (17)$$

Different from the MDB presented in the previous section, the correlation coefficient $\rho_{wi,wj}$ in equation (17) does not depend on the non-centrality parameter, which is a function of the probability levels $\alpha_0$, $\beta_0$, and $r$. Actually, it depends only on the design model (matrices $\boldsymbol{A}$ and $\boldsymbol{W}$) and the outlier models $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$. $\boldsymbol{W}$ is the weight matrix of the observations, taken as $\boldsymbol{W} = \sigma_0^2\boldsymbol{\Sigma}_{yy}^{-1}$, where $\sigma_0^2$ is the variance factor (Koch 1999, Lehmann 2012).

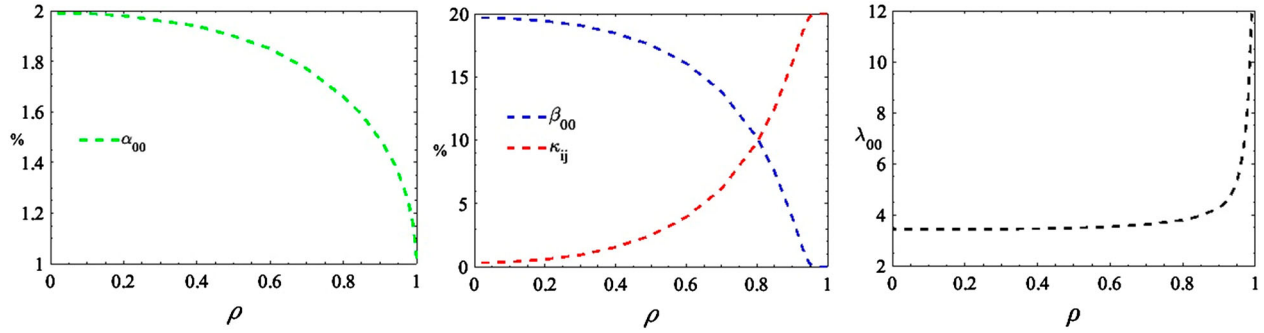
For the test with two alternative hypotheses, apart from type I (false alarm) and type II errors (missed detection), there is a third type of wrong decision when Baarda's data snooping is performed. Baarda's data snooping can also flag a non-outlying observation while the 'true' outlier remains in the dataset. We are referring to the type III error (Hawkins 1980), also called 'wrong exclusion' (Van Mierlo 1979, Förstner 1983, Li 1986, Yang *et al.* 2017). The determination of the type III error (here denoted by $\kappa_{ij}$) involves a separability analysis between the alternative hypotheses (Van Mierlo 1979, Förstner 1983, Li 1986, Wang and Knight 2012, Yang *et al.* 2013, Prószyński 2015, Yang *et al.* 2017). Therefore, we are now interested in the identification of the correct alternative hypothesis. Suppose that there is an outlier in the $i$th position of the dataset, the probability of committing a type III error ($\kappa_{ij}$) is given by

Type III error:

$$\mathrm{P}\Big[|wj| > |wi|,\ \ |wj| > \sqrt{k}\ (i \neq j)|\mathcal{H}_i = \ true\Big] = \kappa_{ij} \quad (18)$$

In contrast to the single alternative hypothesis, the acceptance region for the case of two alternative hypotheses, i.e. $\mathcal{H}_i$ and $\mathcal{H}_j$, is now given by the intersection of two individual acceptance regions, i.e. $\mathcal{A}\alpha_{00} = \mathcal{A}^i_{\alpha_0}\bigcap\mathcal{A}^j_{\alpha_0},\ (i \neq j)$. In this case, the intersection of the two individual acceptance regions has a smaller area compared to each of the individual ones (note: the index '00' represents the case of two alternative hypotheses testing). Thus, the type II error probability becomes smaller, i.e. $\beta_{00} \leq \beta_0$, whereas the type I error becomes larger, i.e. $\alpha_{00} > \alpha_0$. Now, the power of the test is the complement of the sum of type II and III errors, $\gamma_0 = 1 - (\beta_{00} + \kappa_{ij})$. Furthermore, the probabilities of committing type I, II, and III errors are essentially related to the dependencies of the alternative hypotheses. In this case, the larger the correlation coefficient, the larger the type III error and the smaller the type II error (Yang *et al.* 2013).

Based on the work by Yang *et al.* (2013), the left side graph of Fig. 3 shows that $\alpha_{00}$ becomes smaller when $\rho$ increases from 0 to 1, given $\alpha_0 = 0.01$ and $\beta_{00} + \kappa_{ij} = 0.2$. The middle graph of Fig. 3 shows that the larger the correlation coefficient ($\rho > 0.8$), the higher the type III error ($\kappa_{ij}$). When the correlation coefficient is very large, $\beta_{00}$ is effectively zero whereas $\kappa_{ij}$ is 0.2 (20.0%). Therefore, the larger the correlation coefficient, the larger the value of $\kappa_{ij}$ and the smaller the $\beta_{00}$. The right side




**3 Relationships of the error probabilities, non-centrality parameters and the correlation coefficient of two alternative hypotheses**

graph of Fig. 3 shows that the non-centrality parameter $\lambda_{00}$ for two alternative hypotheses increases as the correlation increases from 0 to 1.

## The case of multiple alternative hypotheses

In the local model test with a single alternative hypothesis $\mathcal{H}_i$ of equation (9), the sizes of type I and II errors are given. Under this assumption, the non-centrality parameter coinciding with the MDB can be obtained as a lower bound of the outlier that can be successfully detected (Yang *et al.* 2013). In practice, however, we do not have a single alternative hypothesis during the data snooping procedure, but we have multiple alternative hypotheses. Therefore, the data snooping procedure has an effect when it returns the largest absolute value among the *wi* (Teunissen 2006), i.e.

$$w = \max_{i\in\{1,\ldots,n\}} |wi| \tag{19}$$

The concept of multiple testing says that if $\mathcal{H}_0$ is rejected, among all $\mathcal{H}_i$'s the one should be accepted, which would have rejected $\mathcal{H}_0$with the least $\alpha$. In the case that all critical values are identical, it is most simple: $\mathcal{H}_i$ with the maximum test statistic should be accepted. In order to check its significance, the maximum value ($w$) should be compared with a critical value ($\sqrt{k}$). For instance, a $j$th observation is suspected to be an outlier (i.e. we accept $\mathcal{H}_j$) when

$$\begin{aligned} |wj| > |wi|\ \forall i,\ \ |wj| &> \sqrt{k} \\ &= \sqrt{\chi^2_{\alpha_0}(r = q = 1, 0)}(i \neq j) \end{aligned} \tag{20}$$

If none of the $n$ $w$-tests gets rejected, then we accept the null hypothesis $\mathcal{H}_0$. The data snooping procedure for the case of multiple alternative hypotheses is therefore given as

$$\text{Accept}\ \ \mathcal{H}_0\ \ \textit{if}\ \ w = \max_{i\in\{1,\ldots,n\}} |wi| \leq \sqrt{k} \tag{21}$$

Otherwise,

$$\text{Accept}\ \ \mathcal{H}i\ \ \text{if}\ \ w = \max_{i\in\{1,\ldots,n\}} |wi| > \sqrt{k} \tag{22}$$

The computational complexity increases considerably as the number of hypotheses increases (Yang *et al.* 2013). This means testing $\mathcal{H}_0$ against $\mathcal{H}_1$,$\mathcal{H}_2$,$\mathcal{H}_3$, …, and $\mathcal{H}n$. However, dealing with $n$ alternative hypotheses is not a trivial task for quality control purposes, because the higher the dimensionality of the alternative hypotheses, the more complicated the level probabilities associated with the data snooping procedure.

The acceptance region for $\mathcal{H}_0$ is the intersection of the n individual acceptance regions, i.e. $\mathcal{A}\alpha n = \cap_{i=1}^{n}\mathcal{A}^i_{\alpha_0}$, $(\forall i = 1, \ldots, n)$. Because the acceptance region of the multiple test $\mathcal{A}\alpha n$ is smaller than the single alternative hypothesis $\mathcal{A}^i_{\alpha_0}$ (equation 12), the probability of type II error through equation (21) will be

$$\beta n = \mathrm{P}[\boldsymbol{y} \in \mathcal{A}\alpha n\ |\mathcal{H}_i] \leq \mathrm{P}[\boldsymbol{y} \in \mathcal{A}\alpha_0\ \ |\mathcal{H}_i] = \beta_0 \tag{23}$$

where $\beta n$ is the type II error probability for the multiple alternative hypotheses, whereas $\beta_0$ is for the case of one alternative hypothesis (note: the index '$n$' represents the case of multiples alternative hypotheses).

Geometrically, the intersection of all $n$ individual acceptance regions has a smaller area than each of the individual ones. Hence, the type II error probability becomes smaller, i.e. $\beta n \leq \beta_0$. Therefore, we also have the following MDB inequality:

$$\mathrm{MDB}n \leq \mathrm{MDB}_0 = \sigma_{\hat{\nabla}i}\sqrt{\lambda_{q=1}(\alpha_0,\ \gamma_0)} \tag{24}$$

The right side of the inequality (24) is typically easy to compute using equation (16). Taking into consideration the geometric complexity of the acceptance region and the probability density function of the $w$-tests, the computation of the left side requires a MC [see Imparato *et al.* (2018) for more details]. In multiple tests, on the other hand, for type I error, we have

$$\alpha n = \mathrm{P}[\boldsymbol{y} \notin \mathcal{A}\alpha\boldsymbol{n}\ \ |\mathcal{H}_0] \geq \mathrm{P}[\boldsymbol{y} \notin \mathcal{A}\alpha_0\ \ |\mathcal{H}_0] = \alpha_0 \tag{25}$$

According to equation (23), the type I error will generally become larger, i.e. $\alpha n \geq \alpha_0$. By using the concept of Bonferroni (1936) and neglecting the dependence between $w$-tests, the type I error for multiple tests can be given as $1 - \alpha n = (1 - \alpha_0)^n$, where for small $\alpha_0$, it can be approximated as $\alpha n \leq n\alpha_0$. As pointed out by Imparato *et al.* (2018), this approximation works very well for a system with a high redundancy and/or low correlated $w$-tests. Lehmann (2012) used the MC by considering the dependencies between the least-squares residuals. In that case, as $n$ increases, the values of the correlation coefficients decrease.

In addition, a wrong exclusion is caused by a high correlation between Baarda's $w$-test statistics. Therefore, the separability analysis for controlling misidentification when there are multiple alternative hypotheses should also be considered. For the multiple alternative hypotheses case, the probability of committing different types of error are presented in Table 1.

**Table 1 Decisions for the multiple alternative hypotheses case (adapted fromYang *et al.* 2013)**

| | Result of test | | | | |
|---|---|---|---|---|---|
| | $\mathcal{H}_0$ | $\mathcal{H}_1$ | $\mathcal{H}_2$ | | $\mathcal{H}n$ |
| Reality ('unknown') | $\lvert wi\rvert \le \sqrt{k}, \forall i$ | $\lvert w_1\rvert > \sqrt{k}, \lvert w_1\rvert > \lvert wi\rvert, \forall i$ | $\lvert w_2\rvert > \sqrt{k}, \lvert w_2\rvert > \lvert wi\rvert, \forall i$ | ... | $\lvert wn\rvert > \sqrt{k}, \lvert wn\rvert > \lvert wi\rvert, \forall i$ |
| $\mathcal{H}_0$ | Correct decision $1-\alpha n$ | Type I Error $\alpha_{01}$ | Type I Error $\alpha_{02}$ | ... | Type I Error $\alpha_0 n$ |
| $\mathcal{H}_1$ | Type II Error $\beta_{10}$ | Correct identification $1-\beta_{11}$ | Type III Error $\kappa_{12}$ | ... | Type III Error $\kappa_1 n$ |
| $\mathcal{H}_2$ | Type II Error $\beta_{20}$ | Type III Error $\kappa_{21}$ | Correct identification $1-\beta_{22}$ | ... | Type III Error $\kappa_2 n$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $\mathcal{H}n$ | Type II Error $\beta n_0$ | Type III Error $\kappa n_1$ | Type III Error $\kappa n_2$ | ... | Correct identification $1-\beta nn$ |

Under multiple alternative hypotheses, the probabilities of type I errors in the data snooping procedure for outlier identification, when there are no outliers, are given by

$$\alpha_{0i} = \mathrm{P}\Big[|wi| > |wj|\forall j,\ |wi| > \sqrt{k}(i \neq j)|\mathcal{H}_0{:}\mathrm{true}\Big] = \int\limits_{|wi|>|wj|\forall j,|wi|>\sqrt{k}} f'_0\,\partial w_1 \ldots \partial wn \tag{26}$$

where $f'_0$ is the probability density function when the expectation of the multivariate Baarda's $w$-test statistics is zero ($\mu n = 0$).

The confidence level is

$$1-\alpha n = \mathrm{P}\Big[\cap_{i=1}^{n}|wi| \le \sqrt{k}|\ \ \mathcal{H}_0{:}\mathrm{true}\Big] = \int\limits_{|wi|>|wj|\forall j,|wi|>\sqrt{k}} f'_0\,\partial w_1 \ldots \partial wn \tag{27}$$

Based on the assumption that one outlier is in the $i$th position of the dataset, the probability of a correct identification is

$$1-\beta ii = \mathrm{P}\Big[|wi| > |wj|\ \ \forall j,\ |wi| > \sqrt{k}(i \neq j)|\mathcal{H}i{:}\ \mathrm{true}\Big] = \int\limits_{|wi|>|wj|\forall j,|wi|>\sqrt{k}} f_i i'\,\partial w_1 \ldots \partial wn \tag{28}$$

where $f_i i'$ is the probability density function when the expectation of the multivariate Baarda's $w$-test statistics is not equal to zero ($\mu n \neq 0$).

The probability of type II error for multiple testing is

$$\beta i0 = \mathrm{P}\Big[\cap_{i=1}^{n}|wi| \le \sqrt{k}|\mathcal{H}i{:}\mathrm{true}\Big] \tag{29}$$

and the size of type III error is given by

$$\sum_{i=1(i\neq j)}^{n}\kappa_{ij} = \sum_{i=1}^{n}\mathrm{P}\Big[|wj| > |wi|\ \ \forall i,\ |wj| > \sqrt{k}(i \neq j)|\ \ \mathcal{H}i{:}\ \mathrm{true}\Big] \tag{30}$$

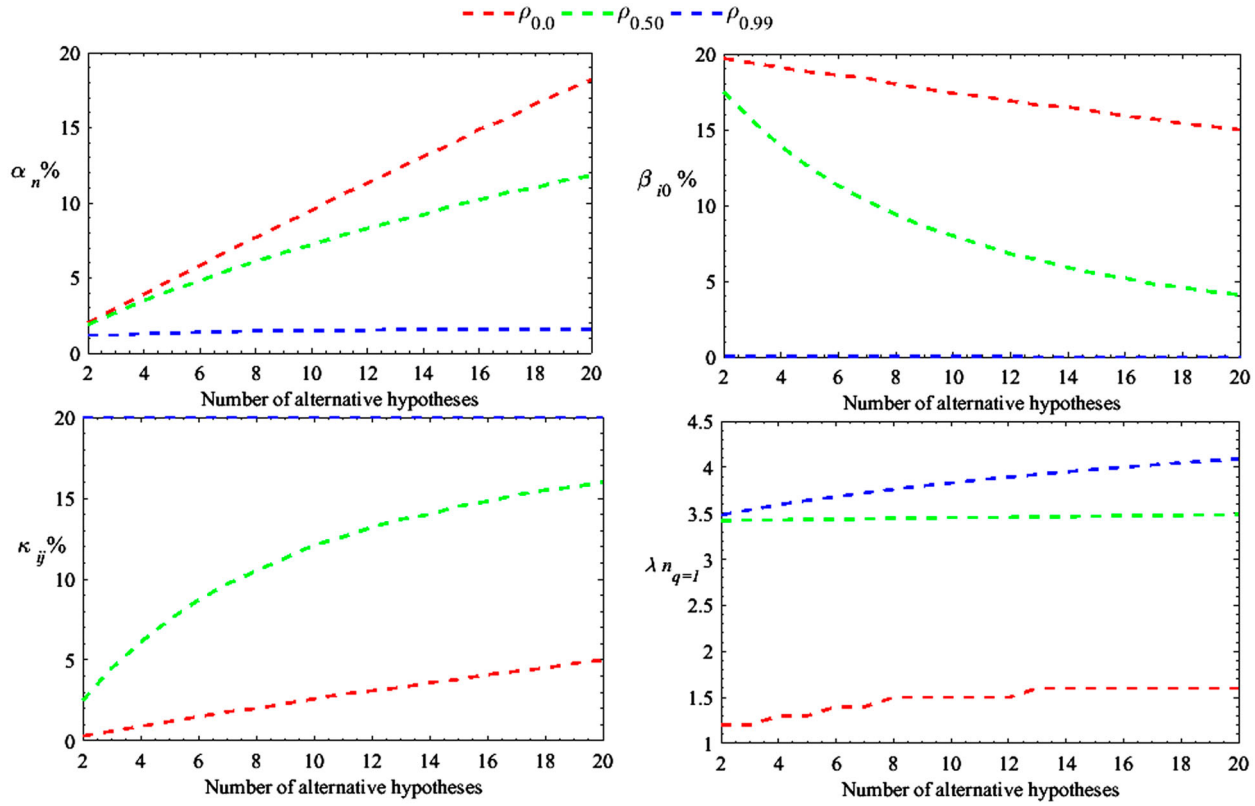
The probability levels associated with Baarda's data snooping under multiple alternative hypotheses by using an analytical formulation were described by Yang *et al.* (2013), as demonstrated above. However, these authors have already pointed out that the density functions and the critical region under multiple alternative hypotheses are extremely complex and even impossible to obtain deterministically. They used an empirical method based on numerical integration of random observation errors to analyse their analytical expression. The relationship between the number of observations ($n$ alternative hypotheses), the probabilities of committing different types of errors during data snooping procedure, and the non-centrality parameter for multiple alternative hypotheses (denoted by $\lambda n_{q=1}$) are shown in Fig. 4. The results were based on the experiments performed by Yang *et al.* (2013) in which the correlation coefficients between Baarda's $w$-test statistics were assumed to be the same.

From Fig. 4, although $\alpha n$ increases with the number of alternative hypotheses, $\alpha n$ decreases significantly as the correlation coefficient $\rho$ increases. The non-centrality parameter $\lambda n_{q=1}$ notably increases as the correlation coefficient $\rho$ increases and it also increases as the number of alternative hypotheses $n$ increases. Thus, in order to have a low probability of committing a type II or III error, a larger $\lambda n_{q=1}$ should be required. Gradually, the size of type II error diminishes as the number of alternative hypotheses increases. The larger the correlation coefficient, the lower the probability of committing $\beta i0$ and the higher the probability of committing type III error $\kappa_{ij}$.

In this context, Prószyński (2015) provided a supplementation of the MDB by an outlier identifiability index to each observation and a misidentifiability index as the maximum probability of type III error. Assuming that one outlier of MDB magnitude resides in the $i$th position of the dataset, the identifiability index (here denoted as $ID_{(\mathrm{i})}$) is defined in terms of conditional probability as follows:

$$ID_{(i)} = \mathrm{P}\left[\cap_{j=1}^{n}\frac{|wj|}{|wi|} < 1, (i \neq j)|\ \ \cup_{j=1}^{n}|wj| > \sqrt{k}\right] \tag{31}$$

Moreover, the misidentifiability index ($MID_{(\mathrm{i/\,j})}$) as the probability of identifying a $j$th non-outlying observation




**4 Relationships of the error probabilities, non-centrality parameters and the correlation coefficient of multiple alternative hypotheses**

instead of the contaminated *i*th observation is

$$MID_{(i/j)} = \mathrm{P}\left[\bigcap_{j=1}^{n} \frac{|wi|}{|wj|} < 1, (i \neq j) | \quad \cup_{i=1}^{n} |wi| > \sqrt{k}\right] \tag{32}$$

equation (31) indicates that the statistic of the outlier in the *i*th position dominates over each of the corresponding absolute values for the remaining observations within a set of suspected observations. equation (32), on the other hand, shows that the statistic of the non-outlying observation in the *j*th position is prevailing within a set of suspected observations. Thus, the ratio between the *w*-statistics for both equations (31 and 32) is smaller than one.

Prószyński (2015) then proposed the maximum value of $MID_{(i/j)}$, i.e. the maximum probability of committing type III errors. Furthermore, he formulated the following relationship based on Eqs. (31) and (32):

$$ID_{(i/j)} = 1 - \sum_{j=1, j\neq i}^{n-1} MID_{(i/j)} \tag{33}$$

with *n* as in Eqs. (31) and (32) being the number of observations (or alternative hypotheses). The indices in equation (33) show that the higher the probability of finding the contaminated observation (Wang and Knight 2012), the lower the probability of committing the type III error. The outliers of MDB magnitudes are often not identified. Thus, Prószyński (2015) also provided a multiplying factor, which indicates the degree of augmentation of MDB, to obtain the required level of outlier identifiability. This multiplying factor is based on another index called partial pseudo-identifiability index ($ID^*_{(i/j)}$), which is given by

$$ID^*_{(i/j)} = \mathrm{P}\left[\frac{|wj|}{|wi|} < 1, (i \neq j)\right] \tag{34}$$

An analogy can be found between the proposal by Prószyński (2015) and Wang and Knight (2012). In the approach by the latter, the concept of minimal separable bias (MSB), which is the magnitude of MDB increased by the multiplying factor to ensure identification of an outlier at a satisfactory confidence level, is presented. When the number of observations increases, the definitions of Eqs. (31), (32), (33), and (34) become more complex. Therefore, an empirical method based on MC of random observation errors was used to analyse the above analytical expression [for more details, see Prószyński (2015)].

In the context of identifiability of the hypotheses, Teunissen *et al.* (2017) recently introduced the concept of MIB as the smallest outlier that leads to its identification for a given correct identification rate. The detection and identification are equal in the case where we only have the one alternative hypothesis. However, under *n* alternative hypotheses (multiple testing), we have from equations (28–30)

$$\beta ii = \beta i0 + \sum_{i=1(i\neq j)}^{n} \kappa_{ij} \tag{35}$$

or

$$1 - \beta ii = \gamma_0 - \sum_{i=1(i\neq j)}^{n} \kappa_{ij} \therefore \gamma_0$$
$$= 1 - \beta ii + \sum_{i=1(i\neq j)}^{n} \kappa_{ij} \quad (36)$$

The probability of correct detection $\gamma_0$ (power of the test for a single alternative hypothesis) is the sum of the probability of correct identification $1 - \beta$ii (selecting a correct alternative hypothesis) and the probability of misidentification $\sum_{i=1(i\neq j)}^{n} \kappa_{ij}$ (selecting one of the $n-1$ other hypotheses). Thus, we have the inequality $1 - \beta ii \leq \gamma_0$ (Imparato *et al.* 2018), which is similar to equation (23). As a consequence of that inequality, the MIB will be larger than MDB:

$$\text{MIB}n \geq \text{MDB}_0 \quad (37)$$

The computation of MIB$n$ should be based on MC, because the acceptance region (as well as the critical region) for the multiple alternative hypotheses case is analytically intractable. In this respect, Imparato *et al.* 2018 showed how to compute the quantities MIB$n$ and MDB$n$. They found that the larger the size of the outlier and/or more precisely, the estimated outlier, the higher the probability of being correctly identified. In addition, increasing the type I error (i.e. reducing the acceptance region) leads to higher probabilities of correct identification. Furthermore, increasing the number of alternative hypotheses leads to a lower probability of correct identification. Therefore, the general characteristic of MIB$n$, MDB$n$, and $\text{MDB}_0$ is given as

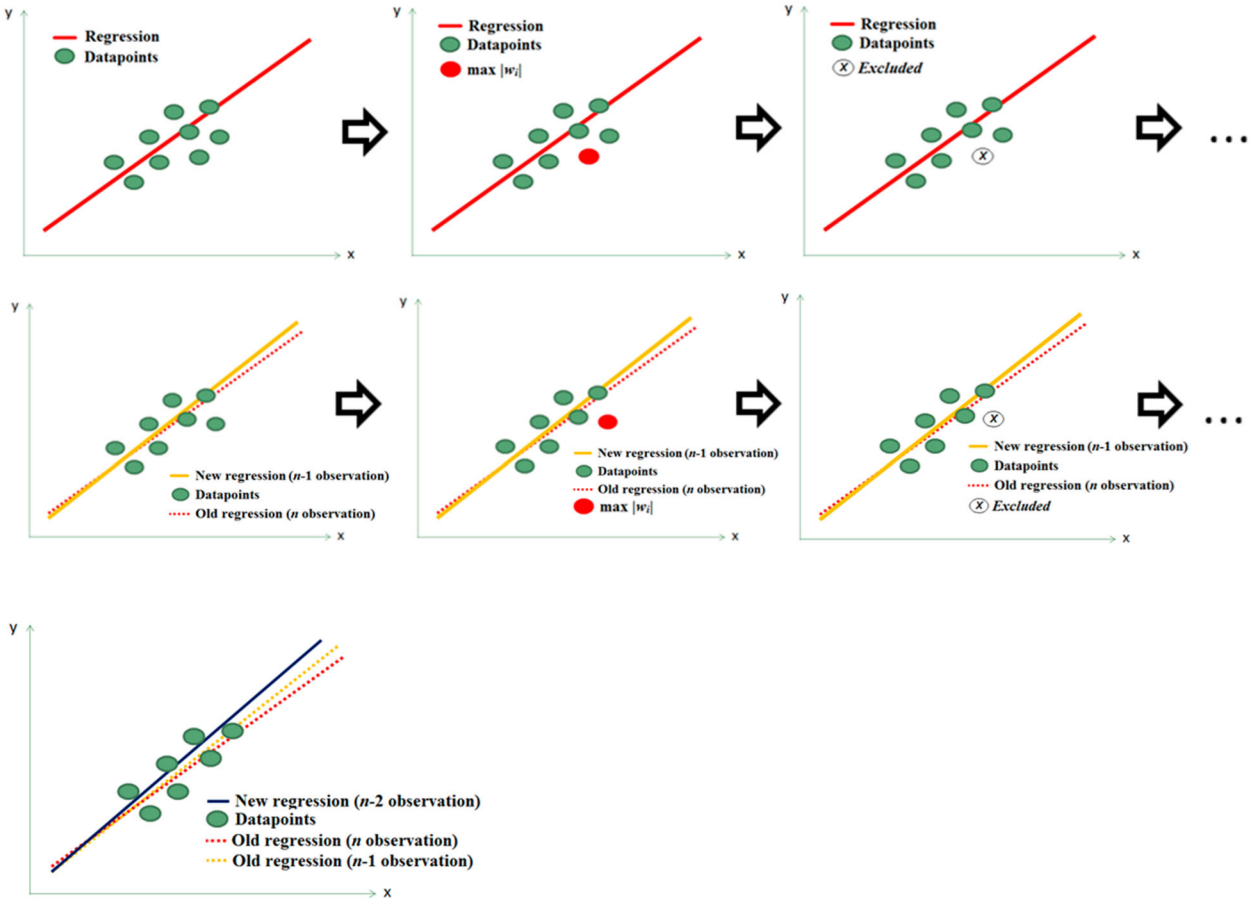
$$\text{MIB}n \geq \text{MDB}_0 \geq \text{MDB}n \quad (38)$$

There is no difference between MDB and MIB in the case of a single alternative hypothesis. As the number of alternative hypotheses increases, however, MDBs become smaller, whereas MIBs become larger. The relation in (38) shows that $\text{MDB}_0$ can be used as a safe upper bound for its multivariate MDB$n$. However, this is not the case for MIBs. For more details about identifiability see e.g. Zaminpardaz and Teunissen (2018).

## Data snooping with adaptation: an iterative procedure

From the MIB concept, one may evaluate, using an a priori analysis, the probability of identifying an outlier in the first adjustment run. In that case, the data snooping procedure is applied only once according to the detector (equation (19)). In practice, however, data snooping is applied iteratively in the process of estimation, identification, and adaptation. Fig. 5 illustrates an example of data snooping applied iteratively to a regression model for the case where there are two outliers in the dataset and it is assumed that false decisions, such as types I, II, and III errors, do not occur.

First, the least-squares residual vector is estimated and Baarda's $w$-test statistics are computed by equation (10).




**5 Data snooping procedure in its iterative form for the case where there are two outliers in the dataset**

Then, the detector given by equation (19) is applied to identify the most likely outlier (red mark in Fig. 5). The identified outlier is then excluded from the dataset (marked with '*X*' in Fig. 5) and the least-squares estimation adjustment is restarted without the rejected observation. Then, Baarda's *w*-test (equation 11) as well as the detector (equation 19) are again computed. Obviously, if redundancy permits, this procedure is repeated until no more (possible) outliers can be identified. This procedure is called IDS (Teunissen 2006, pp. 135).

The theories presented until the present are based on a single run of data snooping, i.e. without subsequent diagnostic operations such as removal or re-weighting of observations. However, the actual practice of data snooping is performed in its iterative process. Thus, as the case in actual practice, a reliability measure cannot be easily computed for quality control purposes. Consequently, MDB, *ID*, *ID**, *MID*, and MIB are valid only for the case where data snooping is run once, and they cannot be used as a diagnostic tool for IDS. Because an analytical formula is not easy to compute, a MC should be run to obtain the probabilities and reliability measures for IDS. The MC allows insights into these cases where analytical solutions are extremely complex to fully understand, are doubted for one reason or another, or are not available.

## The *p*-value approach to multiple outlier detection

Remember that Baarda's data snooping (10) searches for one outlier at a time. If one is detected, it is excluded and the procedure is repeated, etc. The question is why not searching for multiple outliers already in the first run. We would then have additional alternative hypotheses (2) with $q > 1$ and corresponding additional test statistics already in the first run of the data snooping. The problem is that for each $q$ we get a different critical value. This invalidates (19) as a criterion to identify the outliers because we can reasonably take the maximum only of test statistics relating to the same $q$, i.e. to the same critical value.

A solution of this problem is proposed by Lehmann and Lösler (2016) and is called '*p-value approach*'. In statistical hypothesis testing, the *p*-value or probability value or asymptotic significance is defined as the probability that, when the null hypothesis is true, the test statistic would be of greater magnitude than the actually obtained value *w*.

This approach is realised as follows:

(1) The procedure starts with an overall test (6). If $\mathcal{H}_0$ is accepted, the procedure stops without detected outlier.

(2) Up to some maximum number of outliers (here, denoted by $q_{\max}$) all alternative hypotheses (2) are set up and the correponding test statistics are evaluated. (This can be quite a time consuming operation because the number of such test statistics grows strongly with $q_{\max}$. But Lehmann and Lösler (2016) show that there are opportunities to save most of the computer time.

(3) Selecting some $\alpha_0$ in (7) and comparing the critical values with the test statistics can result in three cases:

(a) case A: No test statistic exceeds its critical value. This would contradict the result of the overall test.

(b) case B: Exactly one test statistic exceeds its critical value. This is the desired case, because the outliers are uniquely identified.

(c) case C: Many test statistics exceed their critical values. This is undesired, because the outliers are not uniquely identified.

(d) It depends on $\alpha_0$ which case occurs. The solution is now to tune $\alpha_0$, which is always to some degree debatable, in such a way that case B occurs. This is equivalent to computing the *p*-value for each $\mathcal{H}_i$ and taking the minimum. This identifies the outliers uniquely.

When $q_{\max} = 1$ is chosen, the *p*-value approach coincides with the result of (19). Lehmann and Lösler (2016) made numerical experiments with that approach in comparison to the iterative approach, which those authors call '*consecutive testing*'. Based on the numerical experiments it is not justified to conclude, which approach detects the outliers best, but it is demonstrated that they behave differently and sometimes even produce different results. To find the best approach in some practical sense, more experiences must be gained.

In the next sections, we focus on some issues associated with the Monte Carlo Method for quality control.

## Accuracy of Monte Carlo integration for quality control purposes

The larger the number of MC experiments (i.e. the sample size of random numbers generated), the better the approximation of an expected value in some stochastic processes. The theoretical convergence is only of the order of $m^{1/2}$ (where $m$ is the number of experiments performed). This order does not depend on the dimension of the dataset (Tanizaki 2004). However, until the present, no study has been conducted to evaluate empirically the accuracy of the MC for quality control purposes in geodesy. Generally, only the degree of dispersion of the MC is considered. Thus, an issue remains: how can we find the optimal number of MC experiments for quality control purpose? We use an exact theoretical reference given by a univariate minimal detectable bias $\mathrm{MDB}_0$ (equation 14).

The procedure to analyse the accuracy of MC can be briefly described as follows:

(1) Generate a vector of random errors. Typically, it is assumed that the random errors of the good measurements are normally distributed with expectation zero. Thus, the random errors are generated by using the well-known Box–Muller method (Box and Muller 1958) based on a multivariate normal distribution of the observations (actually, we used MATLAB's function *mvnrnd* here).

(2) Add the $\mathrm{MDB}_0$ computed for a given observation using the analytical expression (equation 14) to the corresponding random error. Thus, we have the total error (here, denoted by $\varepsilon$) as a combination of the random errors and the observation contaminated by its corresponding $\mathrm{MDB}_0$, i.e.

$$\varepsilon = e + \boldsymbol{c}_i \, \mathrm{MDB}_0(i) \tag{39}$$

where $\boldsymbol{e} \in \mathbb{R}^n$ is the random error generated from normal distribution $\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{yy})$ and $\boldsymbol{c}_i$ consists exclusively of elements with values of 0 and 1, where 1 means that an $i$th outlier of magnitude $\mathrm{MDB}_0(i)$ affects an $i$th measurement, and 0 otherwise.

(3) Compute the least-squares residual $\hat{\boldsymbol{e}}_{\boldsymbol{0}}$ vector as follows:

$$\hat{\boldsymbol{e}}_{\boldsymbol{0}} = \boldsymbol{R}\varepsilon, \quad \text{with} \quad \boldsymbol{R} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W} \tag{40}$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is the redundancy matrix, $\boldsymbol{W}$ is the weight matrix, and $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ is the identity matrix.
(4) Compute Baarda's $w$-statistics given by equation (10). If $|wi| > \sqrt{\chi^2_{\alpha_0}(r=1,\,0)} = \sqrt{k}$, then $\mathcal{H}_0$ is rejected in favour of $\mathcal{H}_i$.
(5) Perform the procedure described from 1 to 4 above for $m$ samples of random error vector $e_\nu$ with $\nu = \{1, \ldots, m\}$. By counting the number of correct detections in the generated samples, one can determine the corresponding power of the test. Thus, if $m$ is the total number of samples simulated, we count the number of times of correct detection $n_{CD}$ in which $|wi^{\,\nu}| > \sqrt{\chi^2_{\alpha_0}(r=1,\,0)} = \sqrt{k}$ for $\nu = \{1, \ldots, m\}$, and then approximate the probability of the power of the test as

$$\hat{\gamma}_0 = \frac{n_{CD}}{m}$$

where $m$ is known as the number of Monte Carlo experiments (or also referred to as Monte Carlo sample size).

One of the potential disadvantages of the MC is that a single run of a MC experiment does not indicate by itself the reliability of the results. For analysing the accuracy of MC, several trials for each sample size should be performed. By computing the average and standard deviations of the trials, one can analyse the accuracy as a function of the MC experiments. The average empirical probability of the power of the test $\bar{\gamma}_0$ should be as close as possible to the true value (i.e. $\gamma_0$). The average is given by

$$\bar{\gamma}_0 = \frac{1}{t}\sum_{i=1}^{t} \hat{\gamma}_{0(i)} \tag{42}$$

where $\hat{\gamma}_{0(i)}$is the empirical power of the test computed for each trial $i$ and for a given MC experiment $m$, and $t$ is the total number of trials. The variance can also be computed to evaluate the degree of variation of the MC experiment as follows:

$$\bar{s}^2_{\bar{\gamma}_0} = \frac{1}{t-1}\sum_{i=1}^{t} (\hat{\gamma}_{0(i)} - \bar{\gamma}_0)^2, \quad \text{for } i = 1, \ldots, t \tag{43}$$

with $\bar{s}^2_{\bar{\gamma}_0}$ being the variance and $\bar{s}_{\bar{\gamma}_0} = \pm(\bar{s}_{\bar{\gamma}_0})^{1/2}$ the standard deviation for a given MC experiment.

For simplicity, a simulated GNSS (Global Navigation Satellite System) network is considered. It has one control station (fixed) and five user stations with unknown 3D positions (X, Y, Z), totalling six minimally constrained stations. For each pair of stations, there are four of five baselines (see Klein *et al.* (2017) for more details of that geodetic network). Here, we choose a significance level of $\alpha_0 = 0.001$ and a power of the test $\gamma_0 = 0.8$. Based on those probabilities, the MDB was computed for the baseline MGIN-SPCA of that geodetic network (see equation (14)).We run $1.10^3$, $5.10^3$, $1.10^4$, $2.10^4$, $3.10^4$, $4.10^4$., $5.10^4$, $6.10^4$, $7.10^4$, $8.10^4$, $9.10^4$, and $10.10^4$ MC experiment. For each MC experiment, 1000 trials were performed to evaluate the MC.

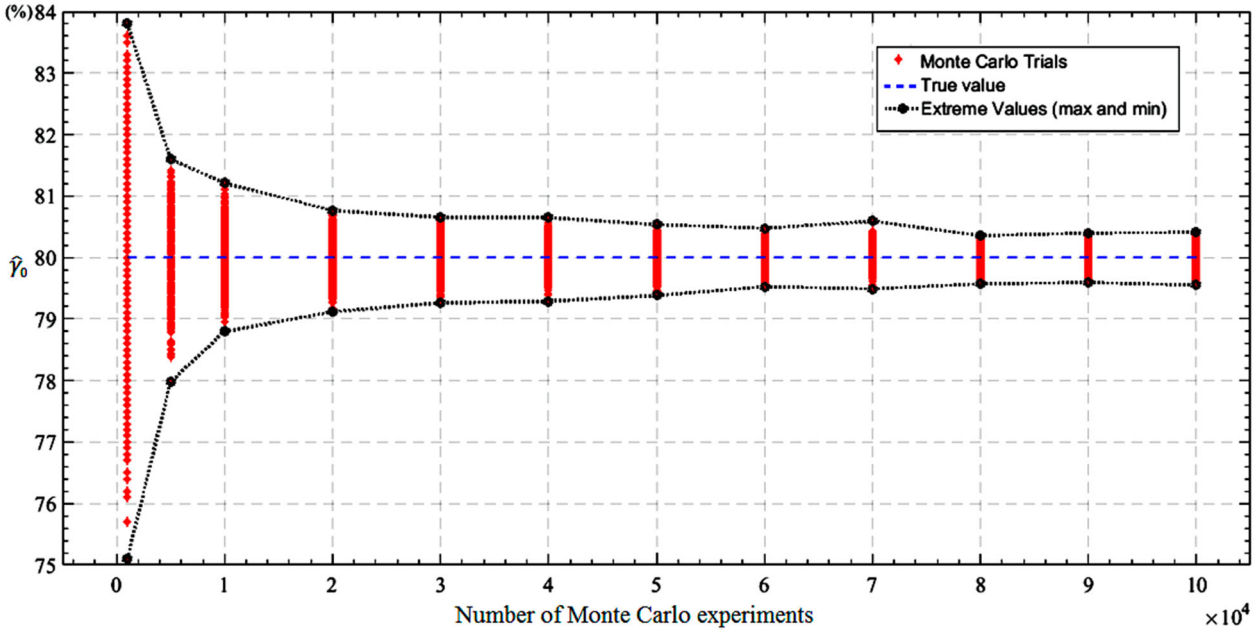
Table 2 presents the statistics of the power of the test computed by MC experiments as well as the mean elapsed time for each MC experiment. In our analyses, Intel Core i5-4200M and Matlab (R2017a) were used for the MC experiments. Fig. 6 shows the variation of the empirical power of the test computed by MC for each experiment.

According to Table 2 and Fig. 6, the greater the number of MC experiments, the more stable the standard deviation of the power of the test $\bar{s}_{\bar{\gamma}_0}$. This property of the MC method can thus be used as a direct way of determining the number of experiments for a given application. For the application of outlier identification, which is the topic in the next section, we suggest that the MC experiment should be chosen when $\hat{\gamma}_0$ in equation (41) fluctuates by no more than 0.1%. It should be noted that the tested experiments are not sufficient to compute the probabilities of order $10^{-3}$.

To ensure that the approximate value of $\hat{\gamma}_0$ is close to the true value $\gamma_0$ in the order of 0.1%, we also ran for $2.10^5$, $5.10^5$, and $10^6$ MC experiments. The latter two numerical experiments provided a degree of dispersion of the order of 0.06% and 0.04%, respectively. For the case of $2.10^5$ experiments, we obtained a standard deviation of $\pm$ 0.07%. Both $5.10^5$ and $10^6$ MC experiments did not present significant advantages over the 200,000. Therefore, we consider that the number of experiments

**Table 2 Statistics of MC experiments based on computation of the power of the test for 1000 trials**

| Number of MC experiments | $\bar{\gamma}_0$ ($\pm\,\bar{s}_{\bar{\gamma}_0}$) (%) | Min. $\hat{\gamma}_{0(i)}$ (%) | Max. $\hat{\gamma}_{0(i)}$ (%) | Mean elapsed time (s) |
|---|---|---|---|---|
| $1.10^3$ | 80.02 (±1.31) | 75.10 | 83.80 | $5.10^{-2}$ |
| $5.10^3$ | 80.01 (±0.56) | 77.98 | 81.60 | $2.10^{-1}$ |
| $1.10^4$ | 80.02 (±0.39) | 78.80 | 81.21 | $5.10^{-1}$ |
| $2.10^4$ | 80.01 (±0.28) | 79.12 | 80.77 | $9.10^{-1}$ |
| $3.10^4$ | 79.99 (±0.23) | 79.26 | 80.65 | 1.3 |
| $4.10^4$ | 80.00 (±0.21) | 79.29 | 80.65 | 1.8 |
| $5.10^4$ | 79.99 (±0.18) | 79.38 | 80.53 | 2.2 |
| $6.10^4$ | 80.00 (±0.16) | 79.53 | 80.48 | 2.6 |
| $7.10^4$ | 79.99 (±0.16) | 79.49 | 80.59 | 3.0 |
| $8.10^4$ | 79.99 (±0.14) | 79.57 | 80.36 | 3.4 |
| $9.10^4$ | 79.99 (±0.13) | 79.59 | 80.40 | 3.7 |
| $10.10^4$ | 80.00 (±0.13) | 79.55 | 80.41 | 4.6 |




**6 Variation of the empirical power of the test $\hat{\gamma}_0$ for each MC experiment for 1000 trials. The blue horizontal dashed line at $\hat{\gamma}_0$ = 80% is the true value of the power of the test used for calculating MDB ($\gamma_0 = 0.8$). The black dotted line shows the maximum and minimum values of $\hat{\gamma}_0$. As expected, a small number of MC experiments provide a wider spread of results**

of 200,000 can provide consistent results with sufficient numerical precision for outlier identification.

## Probabilities of IDS based on Monte Carlo integration

The probability of identifying an outlier is still a bottleneck in geodesy. The challenge is even greater for the case of IDS, i.e. when data snooping is applied iteratively. Unfortunately, the probability levels associated with IDS are not easy to study using analytical models owing to the paucity or lack of practically computable solutions (closed form or numerical). In contrast, a MC method can almost always be run to generate system histories that yield useful statistical information on system operation and performance measures (Altiok and Melamed 2007, Gamerman and Lopes 2006, Robert and Casella 2013).

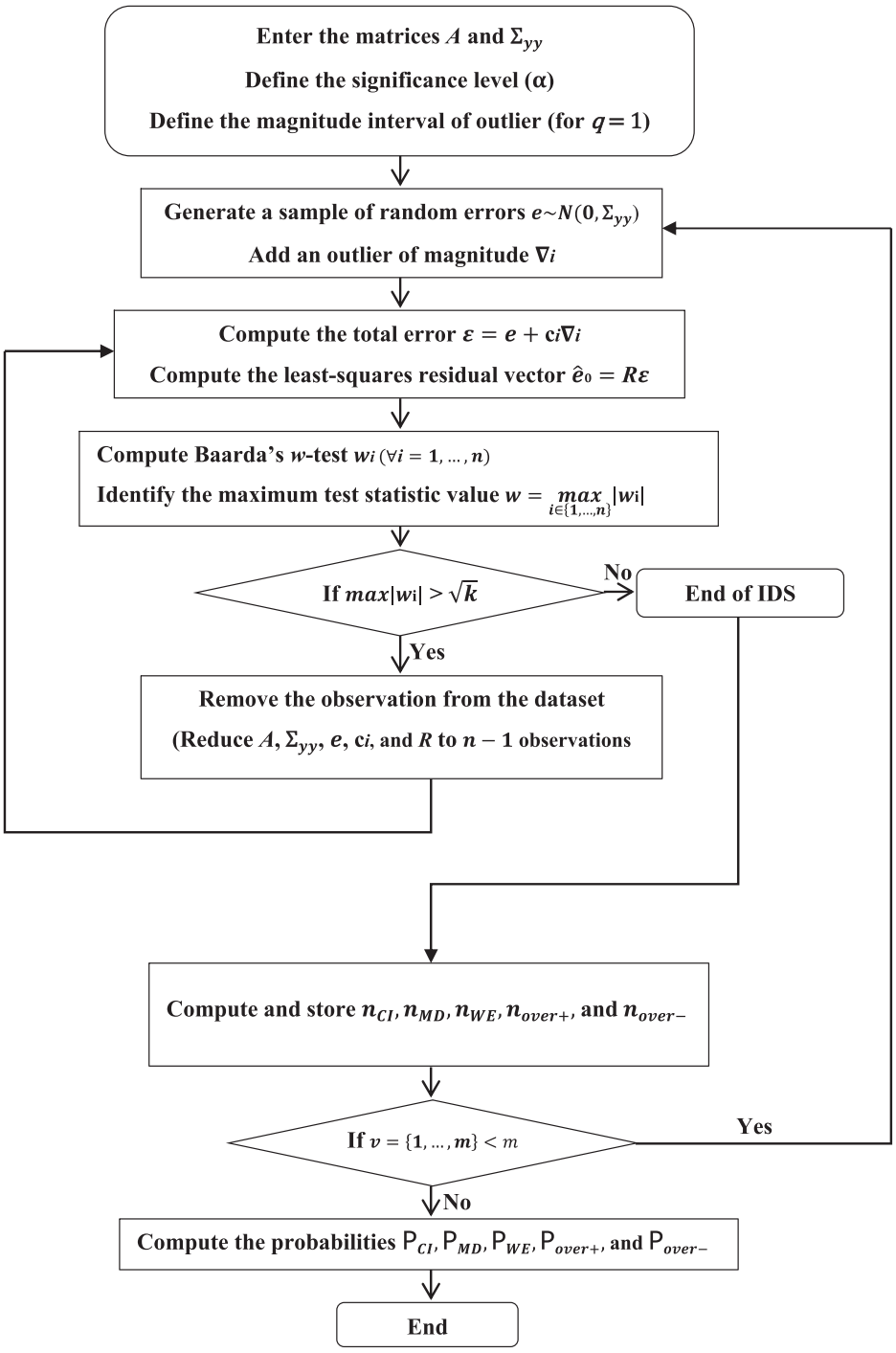
Recent studies by Rofatto *et al.* (2017) showed how to extract the probability levels associated with Baarda's IDS procedure by MC. Furthermore, they introduced two new classes of wrong decisions for IDS, which they called over-identification. One is the probability of IDS flagging simultaneously the outlier and good observations. Second is the probability of IDS flagging only the good observations as outliers (more than one) while the outlier remains in the dataset. Obviously, these two new false decisions could occur during the iterative process of estimation, identification, and exclusion, as is the case of IDS. However, more studies on the subject are still required. They have considered 10,000 experiments, which can not be ideal as shown in the previous section.

A procedure based on the MC is applied to compute the probability levels of IDS as follows (summarised as a flowchart in Fig. 7). In the first step, the design matrix $A \in \mathbb{R}^{n\times u}$ and the covariance matrix of the measurements $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}^{n\times n}$ are entered; then, the significance level ($\alpha$) and the magnitude intervals of simulated outliers are defined. The magnitude intervals of outliers are based on a standard deviation of observation (e.g. $|3\sigma$ to $9\sigma|$, where $\sigma$ is the standard deviation of observation. The random error vectors are synthetically generated based on a multivariate normal distribution, because the assumed stochastic model for random errors is based on a matrix covariance of the observations. Here, we use the Mersenne Twister algorithm (Matsumoto and Nishimura 1998) to generate a sequence of random numbers and Box–Muller to transform it into a normal distribution. On the other hand, the magnitude of the outlier (for $q = 1$) is selected based on magnitude intervals of the outliers for each Monte Carlo experiment. We use the continuous uniform distribution to select the outlier magnitude. Thus, similar to the previous section, the total error ($\varepsilon$) is a combination of the random errors, and its corresponding outlier is as follows:

$$\varepsilon = e + c_i \nabla \boldsymbol{i} \tag{43}$$

where $\boldsymbol{e} \in \mathbb{R}^n$ is the random error generated from normal distribution $\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{yy})$ and $c_i$ consists exclusively of elements with values of 0 and 1, where 1 means that an $i$th outlier of magnitude $\nabla \boldsymbol{i}$ affects an $i$th measurement, and 0 otherwise. Then, the least-squares residuals vector $\hat{\boldsymbol{e}}_{\boldsymbol{0}}$ is computed similarly using equation (40).

For IDS, the hypothesis of equation (9) for $q = 1$ (one outlier) is assumed and the corresponding test statistic is computed according to equation (10). Then, the maximum test statistic value is computed according to equation (19). After identifying the observation suspected as the most likely outlier, it is typically excluded from the model, and least-squares estimation and data snooping are applied iteratively until there are no further outliers identified in the dataset. The procedure should be performed for $m$ experiments of random error vectors with each experiment contaminated by an outlier.




**7 Flowchart of the procedure based on MC for computation of the probability levels of IDS for each observation**

By counting the number of correct identifications in the generated samples, one can determine the corresponding success rate. Thus, if $m$ is the total number of MC experiments, we count the number of times that the outlier is correctly identified (denoted as $n_{CI}$) so that $\max|wi^{\,v}| > \sqrt{\chi^2_{\alpha_0}(r = q = 1, 0)} = \sqrt{k}$ for $v = \{1, \ldots, m\}$, and then approximate the probability of correct identification ($P_{CI}$) as

$$P_{CI} \approx \frac{n_{CI}}{m} \tag{44}$$

As mentioned in the previous section, $m$ is known as the number of MC experiments. In addition, the error probabilities are also approximated as

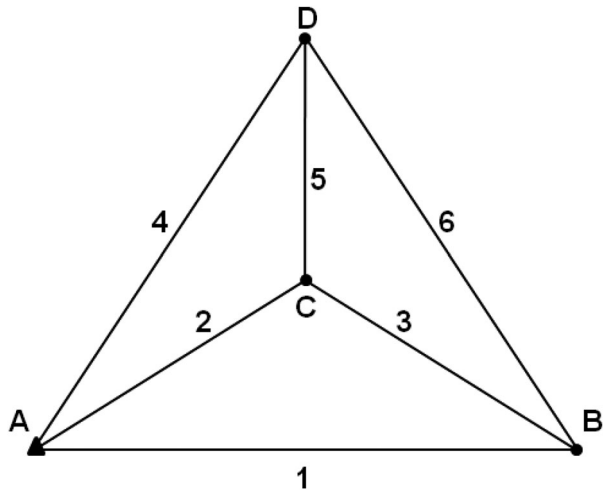
$$P_{MD} \approx \frac{n_{MD}}{m} \tag{45}$$

$$P_{WE} \approx \frac{n_{WE}}{m} \tag{46}$$

$$P_{over+} \approx \frac{n_{over+}}{m} \tag{47}$$

$$P_{over-} \approx \frac{n_{over-}}{m} \tag{48}$$

where $n_{MD}$ is the number of experiments in which the IDS does not detect the outlier ($P_{MD}$ represents the type II error,




**8 Simulated levelling network**

also referred to as missed detection probability); $n_{WE}$ is the number of experiments in which the IDS procedure flags a non-outlying observation while the 'true' outlier remains in the dataset ($P_{WE}$ is the probability of committing a type III error, i.e. wrong exclusion); $n_{over+}$ is the number of experiments where the IDS identifies correctly the outlying observation and others, and $P_{over+}$ corresponds to its probability; finally $n_{over-}$ represents the number of experiments where the IDS identifies more than one non-outlying observation, whereas the 'true outlier' remains in the dataset, and $P_{over-}$ is the probability corresponding to that error probability class.

As an example, the procedure based on MC experiments for the computation of probability levels of IDS is applied to the following simulated closed-levelling network, with one control (fixed) point (A) and three points with unknown heights (B, C, and D), totalling four minimally constrained points (see Fig. 8). Thus, there are $n = 6$ observations, $u = 3$ unknowns, and $n - u = 6 - 3 = 3$ redundant observations in this network. Observations 1, 2, 3, 4, 5, and 6 are assumed normally distributed, uncorrelated, and with nominal precisions (a priori standard deviation σ) of ±8 mm, ±5.6 mm, ±5.6 mm, ±8 mm, ±5.6 mm, and ±8 mm, respectively. The magnitude interval of outlier is from the minimum $3\sigma$ to maximum $6\sigma$, with an interval rate of $0.5\sigma$. Here, positive and negative outliers are considered for each observation. The significance level is taken as $\alpha = 0.001$. We ran 200,000 MC experiments for each observation and for each outlier magnitude interval, totalling 7,200,000 numerical experiments. Tables 3–8 presents the correct identification and error probabilities for one outlier and for each observation.

Tables 3–8 indicate that, in general, the greater the magnitude of outliers, the greater is the efficiency of IDS. In practice, as the magnitudes of outliers are unknown, one can define the probability of the correct identification in order to find the MIB of IDS for a given application. For the above geodetic network, if the probability of correct identification is taken as 0.85 (85%) and the type I error as 0.001 (0.1%), the MIB for observations 1, 4, and 6 is approximately $5.75\sigma$, whereas it is $7.25\sigma$ for observations 2, 3, and 5. It will also be dependent on the functional model (i.e. geometric configuration of the network).

Regarding the cases of misidentification rates, in general, an increase in the magnitude interval of outliers leads to a decrease in the missed detection rate (type II error). Thus, it is important to note that over-identification cases are practically absent for $\alpha = 0.001$. However, as shown by Rofatto *et al.* (2017), for a GNSS network, the probability of committing over-identification during the IDS depends more on the critical value than the outlier magnitude for a one-dimensional identification (one outlier). Furthermore, note that observations 1, 4, and 6 have the same behaviour of wrong exclusion rate (type III error). From $3\sigma$ to $4.5\sigma$, there is an increase in wrong exclusion rate, but there is a decrease from $4.5\sigma$ to $6\sigma$. Similarly, observations 2, 3, and 5 have an increase in wrong exclusion rate from $3\sigma$ to $5.5\sigma$ and a decrease from $5.5\sigma$

**Table 3 Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 1, $\alpha = 0.001$**

| Outlier magnitude interval | $P_{CI}$ % | $P_{MD}$ % | $P_{over+}$ % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 21.20 | 75.22 | 0.01 | 0.01 | 3.55 |
| $3.5\sigma$ to $4\sigma$ | 33.45 | 62.24 | 0.02 | 0.01 | 4.29 |
| $4\sigma$ to $4.5\sigma$ | 47.95 | 47.54 | 0.02 | 0.01 | 4.48 |
| $4.5\sigma$ to $5\sigma$ | 62.69 | 32.95 | 0.03 | 0.01 | 4.32 |
| $5\sigma$ to $5.5\sigma$ | 75.59 | 20.60 | 0.05 | 0.02 | 3.75 |
| $5.5\sigma$ to $6\sigma$ | 85.45 | 11.48 | 0.06 | 0.02 | 2.98 |

**Table 4 Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 2, $\alpha = 0.001$**

| Outlier magnitude interval | $P_{CI}$ % | $P_{MD}$ % | $P_{over+}$ % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 9.91 | 86.69 | 0.00 | 0.00 | 3.40 |
| $3.5\sigma$ to $4\sigma$ | 16.27 | 79.43 | 0.00 | 0.01 | 4.29 |
| $4\sigma$ to $4.5\sigma$ | 25.06 | 69.92 | 0.00 | 0.00 | 5.01 |
| $4.5\sigma$ to $5\sigma$ | 35.69 | 58.65 | 0.01 | 0.01 | 5.65 |
| $5\sigma$ to $5.5\sigma$ | 47.39 | 46.68 | 0.01 | 0.01 | 5.90 |
| $5.5\sigma$ to $6\sigma$ | 59.42 | 34.73 | 0.03 | 0.01 | 5.81 |

**Table 5** Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 3, $\alpha = 0.001$

| Outlier magnitude interval | $P_{CI}$ % | $P_{MD}$ % | $P_{over+}$ % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 9.83 | 86.84 | 0.00 | 0.00 | 3.33 |
| $3.5\sigma$ to $4\sigma$ | 16.57 | 79.21 | 0.00 | 0.00 | 4.21 |
| $4\sigma$ to $4.5\sigma$ | 25.11 | 69.69 | 0.01 | 0.01 | 5.19 |
| $4.5\sigma$ to $5\sigma$ | 35.47 | 58.87 | 0.01 | 0.01 | 5.65 |
| $5\sigma$ to $5.5\sigma$ | 47.55 | 46.49 | 0.01 | 0.01 | 5.94 |
| $5.5\sigma$ to $6\sigma$ | 59.46 | 34.75 | 0.02 | 0.01 | 5.75 |

**Table 6** Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 4, $\alpha = 0.001$

| Outlier magnitude interval | $P_{CI}$ % | $P_{MD}$ % | $P_{over+}$. % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 21.04 | 75.32 | 0.01 | 0.00 | 3.63 |
| $3.5\sigma$ to $4\sigma$ | 33.59 | 62.13 | 0.01 | 0.01 | 4.25 |
| $4\sigma$ to $4.5\sigma$ | 48.01 | 47.47 | 0.02 | 0.01 | 4.49 |
| $4.5\sigma$ to $5\sigma$ | 62.79 | 32.81 | 03 | 0.01 | 4.36 |
| $5\sigma$. to $5.5\sigma$ | 75.63 | 20.53 | 0.06 | 0.01 | 3.77 |
| $5.5\sigma$ to $6\sigma$ | 85.46 | 11.42 | 0.06 | 0.02 | 3.04 |

**Table 7** Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 5, $\alpha = 0.001$

| Outlier magnitude interval | $P_{CI}$ % | | $P_{over+}$ % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 9.82 | 86.81 | 00 | 0.00 | 3.36 |
| $3.5\sigma$ to $4\sigma$ | 16.30 | 79.51 | 0.00 | 0.00 | 4.19 |
| $4\sigma$ to $4.5\sigma$ | 25.06 | 69.97 | 0.01 | 0.00 | 4.96 |
| $4.5\sigma$ to $5\sigma$ | 35.60 | 58.83 | 0.01 | 0.01 | 5.55 |
| $5\sigma$ to $5.5\sigma$ | 47.45 | 46.56 | 0.02 | 0.01 | 5.96 |
| $5.5\sigma$ to $6\sigma$ | 59.32 | 34.79 | 0.03 | 0.01 | 5.84 |

onwards. Based on equation (17), we found that the correlations between the statistics of observations 1, 4, and 6 have the same order of $\rho = 0.3289$, whereas observations 2, 3, and 5 have the same value of $\rho = 0.5$. Therefore, as expected, the behaviour of type III error is associated with the correlation coefficient. This example shows how to compute MIB for the IDS case based on the MC. Obviously, MIB should be computed for a given probability of correct identification and significance level.

## Improving the statistical power of Baarda's outlier test by the Monte Carlo method

Baarda's test statistics, e.g. (4) or (10), have the property of being uniformly most powerful invariant (UMPI). Lehmann and Voß-Böhme (2017) showed that within the small class of test statistics having a common central or non-central chi squared distribution, they are even uniformly most powerful. This means that their optimality is restricted to only a small class of test statistics. Outside this class there might be a cosmos of (in some sense) more powerful tests.

Lehmann and Voß-Böhme (2017) proved the existence of such a test in the larger class of so-called generalised chi squared distributions. They also proposed a method to construct the corresponding test statistic:

(1) Starting with an initial guess of such a test statistic, which is governed by $n \times (n+1)/2$ parameters, the MC is used to compute the statistical power of the corresponding test.

(2) Those parameters are modified is such a way that the power is increasing, until an optimum is reached, i.e. (1) is embedded in an optimisation procedure.

Until now, four obstacles prevent the practical application of this method:

(1) By Monte Carlo integration we need to evaluate an $n \times (n+1)/2$-dimensional integral, which is numerically very costly.

(2) The function to be optimised in (2) is highly nonlinear and has $n \times (n+1)/2$ dimensions, which requires a time-consuming iterative process.

(3) Unlike Baarda's test in the class of chi squared test statistics, any optimal test statistic in the class of generalised chi squared test statistics is in general not uniform. This means that the optimal solution generally depends on the unknown true values of the parameters $\nabla i$.

(4) Each test requires the construction of a new optimal test statistic.

These obstacles seem to be practically insurmountable. Therefore, Lehmann and Voß-Böhme (2017) do not really propose a 'new outlier detection method', but at least show that in principle Baarda's test can be outperformed.

**Table 8** Probabilities of correct identification $P_{CI}$, missed detection $P_{MD}$ as well as the over-identifications $P_{over+}$ and $P_{over-}$ for one outlier in observation 6, $\alpha = 0.001$

| Outlier magnitude interval | $P_{CI}$ % | $P_{MD}$ % | $P_{over+}$ % | $P_{over-}$ % | $P_{WE}$ % |
|---|---|---|---|---|---|
| $3\sigma$ to $3.5\sigma$ | 21.00 | 75.39 | 0.01 | 0.00 | 3.60 |
| $3.5\sigma$ to $4\sigma$ | 33.74 | 62.11 | 0.01 | 0.01 | 4.14 |
| $4\sigma$ to $4.5\sigma$ | 48.11 | 47.38 | 0.02 | 0.01 | 4.48 |
| $4.5\sigma$ to $5\sigma$ | 62.76 | 32.87 | 0.03 | 0.01 | 4.31 |
| $5\sigma$ to $5.5\sigma$ | 75.56 | 20.56 | 0.04 | 0.02 | 3.82 |
| $5.5\sigma$ to $6\sigma$ | 85.47 | 11.44 | 0.06 | 0.02 | 3.02 |

## Final remarks

In this study, we provided an overview of the latest advances in the reliability theory for geodesy. We have shown that over the course of 50 years, several scientific concepts have been developed according to Baarda's initial ideas. Furthermore, we analysed the pioneering work of Baarda and the recent studies on the subject. Then, we pointed out new possibilities and perspectives on the issue, with a particular emphasis on the Monte Carlo integration.

We highlighted that Monte Carlo method is a primary tool for deriving solutions to complex (or analytically intractable) problems. Here, two problems were presented. First, how can we find the optimal number of Monte Carlo experiments for quality control purpose? We used the closed form of MDB given by Baarda (1968) to answer that question. We showed that the number of experiments of 200,000 could provide consistent results with sufficient numerical precision for outlier identification. However, it is important to mention that if computation time is a goal, one should employ adaptative techniques, such as the adaptive importance sampling algorithm.

Second, how do we obtain the probability levels of the IDS procedure? In this study, we used the Monte Carlo method as a key tool for studying the IDS procedure. We highlighted that, in contrast to the well-defined theories of reliability, the IDS procedure is a heuristic method, and therefore, there is no theoretical reliability measure for it. Hence, an analytical model with tractable solution is unknown, and therefore, one needs to resort to MC. Based on the work by Rofatto *et al.* (2017), we showed how to find the probability levels associated with IDS and how to obtain its MIB for each observation by means of the Monte Carlo method for a given correct identification probability and significance level.

Despite the countless contributions made over the last 50 years, there is continuing research on the topic, which paves the way for further investigations. In addition to the Monte Carlo, new tools have become available, such as genetic algorithms (Alma *et al.* 2011), swarm intelligence (Ye and Chen 2008), simulated annealing (Weisberg and Atkinson 1991), fuzzy approaches (Yousri *et al.* 2007), and metaheuristic algorithms (Koch *et al.* 2017), to handle datasets contaminated by outliers. Their potential for outlier treatment is not yet fully exploited in geodesy.

## Acknowledgement

We appreciate the comments and suggestions of the anonymous reviewers. We also appreciate the support of all members of the research group: 'Quality Control in Geodesy' (http://dgp.cnpq.br/dgp/espelhogrupo/3674873915161650). This study was financed by National Council for Scientific and Technological Development – CNPq (proc. n. 305599/2015-1).

## Funding

This study was financed by National Council for Scientific and Technological Development – CNPq (proc. n. 305599/2015-1).

## ORCID

*Vinicius Francisco Rofatto* http://orcid.org/0000-0003-1453-7530

## References

Alma, Ö.G., Kurt, S., and Ugur, A., 2011. Genetic algorithms for outlier detection in multiple regression with different information criteria. *Journal of statistical computation and simulation*, 81 (1), 29–47. (ISSN 0094-9655).

Altiok, T., and Melamed, B., 2007. *Simulation modeling and analysis with arena*. 1st ed. London: Academic Press.

Arnold, S., 1981. *The theory of linear models and multivariate analysis*. 1st ed. New York: Wiley.

Aydin, C., 2012. Power of global test in deformation analysis. *J Surv. Eng.*, 138, 51–56.

Aydin, C., and Demirel, H., 2005. Computation of Baarda's lower bound of the non-centrality parameter. *Journal of geodesy*, 78, 437–441. doi:10.1007/s00190-004-0406-1.

Baarda, W., 1967. Statistical concepts in geodesy. *Netherlands geodetic commission, publ. on geodesy, new series*, 2 (4), 1–74. https://www.ncgeo.nl/index.php/en/publicatiesgb/publications-on-geodesy/item/2514-pog-08-w-baarda-statistical-concepts-in-geodesy.

Baarda, W., 1968. A testing procedure for use in geodetic networks. *Netherlands geodetic commission, publ. on geodesy, new series*, 2 (5), 1–97. https://ncgeo.nl/index.php?option=com_k2&view=item&id=2515:pog-09-w-baarda-a-testing-procedure-for-use-in-geodetic-networks&Itemid=350&lang=en.

Baselga, S., 2011. Nonexistence of rigorous tests for multiple outlier detection in least-squares adjustment. *Journal of surveying engineering*, 137, 109–112.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Instituto Superiore de Scienze Economiche e Commerciali de Firenze*, 8, 3–62.

Box, G.E., and Muller, M.E., 1958. A note on the generation of random normal deviates. *The analls of mathematical statistics*, 29 (2), 610–611.

Ding, X., and Coleman, R., 1996. Multiple outlier detection by evaluating redundancy contributions of observations. *Journal of geodesy*, 70, 489–498.

Förstner, W. 1983. Reliability and discernability of extended Gauss-Markov models. In: Seminar on mathematical models to outliers and systematic errors, Deutsche Geodätische Kommision, Series A, no. 98. Munich, Germany, 79–103. ISSN 978376681802.

Gamerman, D., and Lopes, H.F., 2006. *Markov chain monte carlo: stochastic simulation for Bayesian inference*. 2nd ed. London, UK: Chapman & Hall/CRC Texts in Statistical Science.

Ghilani, C.D., 2010. *Adjustment computations: spatial data analysis*. 5th ed. New Jersey: John Wiley & Sons.

Gui, Q., *et al.*, 2011. A Bayesian unmasking method for locating multiple gross errors based on posterior probabilities of classification variables. *Journal of geodesy*, 85, 191–203.

Hawkins, D.M., 1980. *Identification of outliers*. London: Chapman & Hall.

Hekimoglu, S., and Koch, K.R., 1999. How can reliability of the robust methods be measured? *In*: M.O. Altan, and L. Gründig, ed. *Third Turkish-German joint geodetic days*. Istanbul: Third Turkish-German Joint Geodetic Days, 179–196.

Imparato, D., Teunissen, P.J.G., and Tiberius, C.C.J.M., 2018. Minimal Detectable and Identifiable Biases for quality control. *Survey review*, 1–11. doi:10.1080/00396265.2018.1437947. (Latest Articles).

Kargoll, B., 2007. *On the theory and application of model misspecification tests in geodesy. Dissertation*. Universitats und Landesbibliothek Bonn.

Kargoll, B. 2012. On the theory and application of model misspecification tests in geodesy, Series C, vol. 674, German Geodetic Commission, Munich, Germany.

Klein, I., *et al.*, 2017. An approach to identify multiple outliers based on sequential likelihood ratio tests. *Survey review*, 49, 449–457.

Knight, N.L., Wang, J., and Rizos, C., 2010. Generalised measures of reliability for multiple outliers. *Journal of geodesy*, 84, 625–635.

Koch, K.R., 1999. *Parameter estimation and hypothesis testing in linear models*. 2nd ed. Berlin: Springer.

Koch, K.R., 2007. *Introduction to Bayesian statistics*. 2nd ed. Berlin: Springer.

Koch, K.R., 2016. Minimal detectable outliers as measures of reliability. *Journal of geodesy*, 89, 483–490.

Koch, I.E., *et al.*, 2017. Least trimmed squares estimator with redundancy constraint for outlier detection in GNSS networks. *Expert Systems With Applications*, 88, 230–237.

Kok, J.J., 1984. *On data snooping and multiple outlier testing*. Rockville, United States: US Department of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, Charting and Geodetic Services. https://babel.hathitrust.org/cgi/pt?id=osu.32435075524116;view=1up;seq=1.

Lehmann, R. 1994. Adjustment in non-linear models by means of the adaptive Monte-Carlo-Integration. Allgemeine Vermessungsnachrichten, vol 7/1994. Herbert Wichmann Verlag GmbH Heidelberg (in German).

Lehmann, R., 2012. Improved critical values for extreme normalized and studentized residuals in Gauss-Markov models. *Journal of geodesy*, 86, 1137–1146.

Lehmann, R., 2013. On the formulation of the alternative hypothesis for geodetic outlier detection. *Journal of geodesy*, 87, 373–386.

Lehmann, R., 2015. Observation error model selection by information criteria vs. normality testing. *Studia Geophysica et Geodaetica*, 59, 489–504.

Lehmann, R., and Lösler, M., 2016. Multiple outlier detection: hypothesis tests versus model selection by information criteria. *Journal of Surveying Engineering*, 142 (4), doi:10.1061/(ASCE)SU.1943-5428.0000189.

Lehmann, R., and Scheffler, T., 2011. Monte Carlo based data snooping with application to a geodetic network. *Journal of geodesy*, 5, 123–134.

Lehmann, R., and Voß-Böhme, A., 2017. On the statistical power of Baarda's outlier test and some alternative. *Journal of geodetic science*, 7, 68–78.

Li, D.R., 1986. Trennbarkeit und Zuverlässigkeit bei zwei verschiedenen alternativhypothesen im Gauß–Markoff-Modell. *Zeitschrift für Vermessungswesen*, 3, 114–128. Heft.

Marx, C., 2015. Outlier detection by means of Monte Carlo estimation including resistant scale estimation. *Journal of applied geodesy*, 9, 123–142.

Matsumoto, M., and Nishimura, T., 1998. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM transactions on modeling and computer simulation*, 8, 3–30.

Pelzer, H. 1983. Detection of errors in the functional adjustment model. Deutsche Geodatische Kommision Seminar on mathematical models of geodetic photogrammetric point determination with regard to outliers and systematic errors, 61-70, Germany, ISSN: 9783769681802.

Prószyński, W., 1994. Criteria for internal reliability of linear least squares models. *Bulletin géodésique*, 68, 162–167.

Prószyński, W., 2010. Another approach to reliability measures for systems with correlated observations. *Journal of geodesy*, 84, 547–556.

Prószyński, W., 2015. Revisiting Baarda's concept of minimal detectable bias with regard to outlier identifiability. *Journal of geodesy*, 89, 993–1003.

Robert, C., and Casella, G., 2013. *Monte carlo statistical methods*. Berlin: Springer.

Rofatto, V.F., Matsuoka, M.T., and Klein, I., 2017. An attempt to analyse Baarda's iterative data snooping procedure based on Monte Carlo simulation. *South African journal of geomatics*, 6, 416–435.

Schaffrin, B., 1997. Reliability measures for correlated observations. *Journal of Surveying Engineering*, 123, 126–137.

Tanizaki, H., 2004. *Computational methods in statistics and econometrics*. 1st ed. New York: Marcel Dekker.

Teunissen, P.J.G., 1989. Quality control in integrated navigation systems. *IEEE aerospace and electronic systems magazine*, 5 (7), 35–41.

Teunissen, P.J.G., 1998. Minimal detectable biases of GPS data. *Journal of geodesy*, 72 (4), 236–244.

Teunissen, P.J.G., 2006. *Testing theory: an introduction*. 2nd ed. Delft, Netherlands: Delft University Press. Series on Mathematical Geodesy and Positioning.

Teunissen, P.J.G., *et al.* 2017. Distributional theory for the DIA method. *Journal of geodesy*, 91, 1–22. doi:10.1007/s00190-017-1045-7.

Vaníček, P., Craymer, M.R., and Krakiwsky, E.J., 2001. Robustness analysis of geodetic horizontal networks. *Journal of geodesy*, 75, 199–209.

Van Mierlo, J., 1975. Statistical analysis of geodetic networks designed for the detection of crustal movements. In: G.J. Borradaile, A.R. Ritsema, H.E. Rondeel, and O.J. Simon, ed. *Progress in geodynamics*. Amsterdam: North-Holland, 52–61.

Van Mierlo, J., 1979. Statistical analysis of geodetic measurements for the investigation of crustal movements. *Tectonophysics*, 52, 457–467.

Wang, J., and Chen, Y., 1994. On the reliability measure of observations. *Acta Geodaetica et Cartographica Sinica*, 23 (4), 42–51. English Edition.

Wang, J., and Chen, Y., 1999. Outlier detection and reliability measures for singular adjustment models. *Geomat. Res. Aust.*, 71, 57–72.

Wang, J., and Knight, N., 2012. New outlier separability test and its application in GNSS positioning. *The Journal of global positioning systems*, 11 (1), 46–57.

Weisberg, S., and Atkinson, A.C., 1991. Simulated annealing for the detection of Multiple Outliers using least squares and least median of squares fittin. In: W. Stahel, and S. Weisberg, ed. *Directions in robust statistics and diagnostics*. New York: Springer-Verlag, 7–20.

Yang, L., *et al.*, 2013. Outlier separability analysis with a multiple alternative hypotheses test. *J Geodesy*, 87 (6), 591–604.

Yang, L., *et al.*, 2017. Extension of internal reliability analysis regarding separability analysis. *Journal of Surveying Engineering*, 143 (3), 04017002, doi:10.1061/(ASCE)SU.1943-5428.0000220.

Ye, D., and Chen, Z., 2008. A new algorithm for high-dimensional outlier detection based on constrained particle swarm intelligence. *Lecture notes in computer science*, 5009, 516–523.

Yousri, N.A., Ismail, M.A., and Kamel, M.S., 2007. *Fuzzy outlier analysis: a combined clustering - outlier detection approach*. 2007 IEEE international conference on systems, man and cybernetics, 412–418. doi:10.1109/ICSMC.2007.4413873.

Zaminpardaz, S., and Teunissen, P.J.G., 2018. DIA-datasnooping and identifiability. *Journal of geodesy*. doi:10.1007/s00190-018-1141-3.