remote sensing

MDPI

Article

A Monte Carlo-Based Outlier Diagnosis Method for Sensitivity Analysis

Vinicius Francisco Rofatto [1,2,*], Marcelo Tomio Matsuoka [1,2,3,4], Ivandro Klein [5,6], Maurício Roberto Veronez [4] and Luiz Gonzaga da Silveira, Jr. [4]

1 Graduate Program in Remote Sensing, Federal University of Rio Grande do Sul, Porto Alegre 91501970, Brazil; tomio@ufu.br
2 Institute of Geography, Federal University of Uberlandia, Monte Carmelo 38500-000, Brazil
3 Graduate Program in Agriculture and Geospatial Information, Federal University of Uberlândia, Monte Carmelo 38500-000, Brazil
4 Graduate Program in Applied Computing, Unisinos University, Av. Unisinos, 950, São Leopoldo 93022-000, Brazil; veronez@unisinos.br (M.R.V.); lgonzagajr@gmail.com (L.G.d.S.J.)
5 Department of Civil Construction, Federal Institute of Santa Catarina, Florianopolis 88020-300, Brazil; ivandroklein@gmail.com
6 Graduate Program in Geodetic Sciences, Federal University of Paraná, Curitiba 81531-990, Brazil
* Correspondence: vfrofatto@gmail.com or vinicius.rofatto@ufu.br

Received: 24 January 2020; Accepted: 4 March 2020; Published: 6 March 2020

Abstract: An iterative outlier elimination procedure based on hypothesis testing, commonly known as *Iterative Data Snooping* (*IDS*) among geodesists, is often used for the quality control of modern measurement systems in geodesy and surveying. The test statistic associated with *IDS* is the extreme normalised least-squares residual. It is well-known in the literature that critical values (quantile values) of such a test statistic cannot be derived from well-known test distributions but must be computed numerically by means of Monte Carlo. This paper provides the first results on the Monte Carlo-based critical value inserted into different scenarios of correlation between outlier statistics. From the Monte Carlo evaluation, we compute the probabilities of correct identification, missed detection, wrong exclusion, over-identifications and statistical overlap associated with *IDS* in the presence of a single outlier. On the basis of such probability levels, we obtain the Minimal Detectable Bias (MDB) and Minimal Identifiable Bias (MIB) for cases in which *IDS* is in play. The MDB and MIB are sensitivity indicators for outlier detection and identification, respectively. The results show that there are circumstances in which the larger the Type I decision error (smaller critical value), the higher the rates of outlier detection but the lower the rates of outlier identification. In such a case, the larger the Type I Error, the larger the ratio between the MIB and MDB. We also highlight that an outlier becomes identifiable when the contributions of the measures to the wrong exclusion rate decline simultaneously. In this case, we verify that the effect of the correlation between outlier statistics on the wrong exclusion rate becomes insignificant for a certain outlier magnitude, which increases the probability of identification.

Keywords: probability; hypothesis testing; outlier detection; monte carlo; quality control; control system; reliability; random number generators

1. Introduction

In recent years, Outlier Detection has been increasingly applied in sensor data processing [1–9]. Despite the countless contributions made over the years, there is continuing research on the subject, mainly because there has been an increase in computational power. One can argue that computational complexity is becoming high because of the era of information overload. However, this limitation

has been overcome over the years, mainly by the rapid development of computers, which now allow advanced computational techniques to be used efficiently on personal computers or even on handheld computers [10]. Therefore, computational complexity is no longer a bottleneck because we have fast computers and large data storage systems at our disposal [11,12].

Here, we assume that an outlier is a measurement that is so likely to be caused by a blunder that it is better to either not use it or not use it as it is [13]. Failure to identify an outlier can jeopardise the reliability level of a system. Because of its importance, outliers must be appropriately treated to ensure the quality of data analysis.

Two categories of advanced techniques for the treatment of a dataset contaminated by outliers have often been developed and applied in various situations: Robust Adjustment Procedures (see, e.g., [14–18]) and Statistical Hypothesis Testing (see, e.g., [2,12,19–23]). The first one is an estimation technique that is not unduly affected by outliers or other small departures from model assumptions. Classes of this technique include M-estimates (which follow from maximum likelihood considerations), L-Estimates (which are linear combinations of order statistics), and R-Estimates (based on statistical rank tests). Some classes of such robust adjustment methods, as well as their properties, are well known, while other methods are still being researched (see, e.g., L_1-norm estimation [24], M-estimation [25–27], R-estimation [28–30] and those based on meta-heuristics [31]). Besides the undoubted advantages of Robust Estimation, here, we focus on the hypothesis test-based outlier. The following advantages of the outlier test were mentioned by [32]:

1. It is an opportunity to investigate the causes of outliers;
2. Identified outliers can be remeasured; and
3. If the outliers are discarded from the measurements, then standard adjustment software, which operates according to the least-squares criterion, can be used.

In this paper, we consider iterative data snooping (*IDS*), which is the most common procedure found in the geodetic practice [12,33]. Most conventional geodetic studies have a chapter on *IDS* (see, e.g., [34,35]). *IDS* has also become very popular and is routinely used in adjustment computations [36]. It is important to mention that *IDS* is not restricted to the field of geodetic statistics but is a generally applicable method [22].

IDS is an iterative outlier elimination procedure, which combines estimation, testing and adaptation [37]. Parameter estimation is often conducted in the sense of the least-squares estimation (LSE). Assuming that no outlier exists, the LSE is the best linear unbiased estimator (BLUE) [35]. The LSE has often been used in several remote sensing applications (see, e.g., [38–41]). However, outliers can inevitably occur in practice and cause the loss of the LSE BLUE-property. Then, hypothesis testing is performed with the aim of identifying any outliers that may be present in the dataset. After its identification, the suspected outlier is then excluded from the dataset as a corrective action (i.e., adaptation), and the LSE is restarted without the rejected measurement. If model redundancy permits, this procedure is repeated until no more (possible) outliers can be identified (see, e.g., [35], p. 135). Although here we restrict ourselves to the case of one outlier at a time, *IDS* can also be applied to the case of multiple (simultaneous) outliers [42]. For more details about multiple (simultaneous) outliers, refer to [43–45].

Of particular importance for quality control purposes are decision probabilities. Probability levels have already been described in the literature for the case in which data snooping is run once (i.e., only one single estimation and testing), as well as for the case in which the outlier is parameterised in the model (see, e.g., [2,19–21,23,37,46,47]). For such cases, the probability of correct detection ($\mathcal{P}_{CD}$) and correct identification ($\mathcal{P}_{CI}$) and their corresponding Minimal Detectable Bias (MDB) and Minimal Identifiable Bias (MIB) have already been described for data snooping [37,46].

The MDB is defined as the smallest value of an outlier that can be detected given a certain $\mathcal{P}_{CD}$. The MDB is an indicator of the sensitivity of data snooping to outlier detection and not to outlier identification. On the other hand, the MIB is defined as the smallest value of an outlier that can

be identified given a certain $\mathcal{P}_{CI}$; i.e., the MIB is an indicator of the sensitivity of data snooping to outlier identification. It is important to highlight that "outlier detection" only informs us whether or not there might have been at least one outlier. However, the detection does not tell us which measurement is an outlier. The localisation of the outlier is a problem of "outlier identification". In other words, "outlier identification" implies the execution of a search among the measurements for the most likely outlier.

However, both the MDB and MIB cannot be used as a diagnostic tool when *IDS* is in play. In this contribution, we highlight the fact that the correct outlier identification for *IDS* is not only dependent on the correct/missed detection and wrong exclusion but also other decision probabilities.

The evaluation of the probability levels associated with *IDS* is not a trivial task. When used for data snooping for a single run, the probabilities of *IDS* are multivariate integrals over complex regions [2,47]. This complexity is due to the fact that *IDS* is not only based on multiple hypothesis testing but also on multiple rounds of estimation, testing and exclusion. Because an analytical formula is not easy to compute, the Monte Carlo method should be run to obtain the probabilities and the minimal bias (MDB and MIB) indicators for *IDS*. The Monte Carlo method provides insights into these cases, in which analytical solutions are too complex to fully understand, are doubted for one reason or another or are not available [12]. The Monte Carlo method for quality control purposes has already been applied in geodesy (see, e.g., [2,10,22,23,33,46,48–51]). For in-depth coverage of Monte Carlo methods, consult, for instance, [52–54].

Recent studies by Rofatto et al. [12,55] provide an algorithm based on Monte Carlo to determine the probability levels associated with *IDS*. In that case, five classes of decisions for *IDS* are described, namely, the probability of correct identification ($\mathcal{P}_{CI}$), the probability of missed detection ($\mathcal{P}_{MD}$), the probability of wrong exclusion ($\mathcal{P}_{WE}$), the probability of over-identification positive ($\mathcal{P}_{over+}$), and the probability of over-identification negative ($\mathcal{P}_{over-}$), defined as follows:

- $\mathcal{P}_{CI}$: The probability of correctly identifying and removing an outlying measurement;
- $\mathcal{P}_{MD}$: The probability of not detecting the outlier (i.e., Type II decision error for *IDS*);
- $\mathcal{P}_{WE}$: The probability of identifying and removing a non-outlying measurement while the '*true*' outlier remains in the dataset (i.e., Type III decision error for *IDS*);
- $\mathcal{P}_{over+}$: The probability of correctly identifying and removing the outlying measurement and others and
- $\mathcal{P}_{over-}$: The probability of identifying and removing more than one non-outlying measurement while the 'true outlier' remains in the dataset.

However, the procedure used by these authors [12,55] does not allow the user to control the Type I decision error (denoted by α'). The probability level α' (known as the significance level of a test) defines the size of a test and is often called the "false alarm probability". In this paper, we highlight the fact that the test statistic associated with *IDS* does not have a known distribution, and therefore, its critical values (i.e., the percentile of its probability distribution) cannot be taken from well-known statistical tables (e.g., normal distribution).

Here, the critical value is computed by Monte Carlo such that a user-defined Type I decision error α' for *IDS* is warranted. In other words, the Type I decision error α' is effectively user-controlled when both the functional and stochastic parts of the model are taken into account. To do so, we employ the Monte Carlo method because the critical region of the test statistic associated with *IDS* is too complicated. The critical region is the subset of the measurements for which the null hypothesis $\mathcal{H}_0$ is rejected [12]. Therefore, the false alarm rate can be user-controlled by setting the appropriate size of the critical region.

We show that one of the advantages of having critical values based on the distribution test of *IDS* is that the dependencies between the least-squares residuals are captured by Monte Carlo simulation. In this paper, we present the first results on the Monte Carlo-based critical value in two different

scenarios of correlation between outlier test statistics. We also discuss this issue in the context of the well-known Bonferroni correction [56] to control the Type I decision error α' for *IDS*.

Moreover, herein, a new class of decision is taken into account when *IDS* is performed, which corresponds to the probability of simultaneously flagging two (or more) measurements as outliers. We call this the probability of "*statistical overlap*" ($\mathcal{P}_{ol}$). This means that $\mathcal{P}_{ol}$ occurs in cases in which one alternative hypothesis has the same distribution as another one. In other words, these hypotheses cannot be distinguished; i.e., they are nonseparable, and an outlier cannot be identified [37].

We also investigate the probabilities of making correct decisions and the risks of incorrect decisions when *IDS* is performed in the presence of an outlier in two different scenarios of correlation between outlier test statistics. On the basis of the probability levels associated with *IDS* (i.e., $\mathcal{P}_{CI}$, $\mathcal{P}_{MD}/\mathcal{P}_{CD}$, $\mathcal{P}_{WE}$, $\mathcal{P}_{over+}$, $\mathcal{P}_{over-}$ and $\mathcal{P}_{ol}$), we also show how to find the two sensitivity indicators MDB and MIB for *IDS*. We also analyse the relationship between the sensitivity indicators MDB and MIB for *IDS*.

2. Binary Hypothesis Testing versus Multiple Hypothesis Testing: True Data Snooping

Random measurement errors in a system are unavoidable. The stochastical properties of measurement errors are directly associated with the assumption of the probability distribution of these errors. In geodesy and many other scientific branches, the well-known normal distribution is one of the most used measurement error models. Its choice is further justified by both the central limit theorem and the maximum entropy principle. Some alternative measurement error models can be found in [11].

Therefore, the null hypothesis, denoted by $\mathcal{H}_0$, is formulated under the condition that random errors are normally distributed with expectation zero. In other words, the model associated with the null hypothesis $\mathcal{H}_0$ consists of the one believed to be valid under normal working conditions, i.e., in the absence of outliers. When it is assumed to be '*true*', this model is used to estimate unknown parameters, usually in a least-squares approach. Thus, the null hypothesis $\mathcal{H}_0$ of the standard Gauss–Markov model in the linear or linearised form is given by [34]

$$\mathcal{H}_0 : \mathbb{E}\{\boldsymbol{y}\} = \boldsymbol{A}\boldsymbol{x} + \mathbb{E}\{\boldsymbol{e}\} = \boldsymbol{A}\boldsymbol{x};\ \mathbb{D}\{\boldsymbol{y}\} = \boldsymbol{Q}_{\boldsymbol{e}} \tag{1}$$

where $\mathbb{E}\{.\}$ is the expectation operator, $\mathbb{D}\{.\}$ is the dispersion operator, $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ is the vector of measurements, $\boldsymbol{A} \in \mathbb{R}^{n\times u}$ is the Jacobian matrix (also called the design matrix) of full rank u, $\boldsymbol{x} \in \mathbb{R}^{u\times 1}$ is the unknown parameter vector, $\boldsymbol{e} \in \mathbb{R}^{n\times 1}$ is the unknown vector of measurement errors and $\boldsymbol{Q}_{\boldsymbol{e}} \in \mathbb{R}^{n\times n}$ is the positive-definite covariance matrix of the measurements $\boldsymbol{y}$.

Under normal working conditions (i.e., $\mathcal{H}_0$), the measurement error model is then given by

$$\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{Q}_{\boldsymbol{e}}), \tag{2}$$

Here, we confine ourselves to the case in which $\boldsymbol{A}$ and $\boldsymbol{Q}_{\boldsymbol{e}}$ have full column rank.

The best linear unbiased estimator (BLUE) of $\boldsymbol{e}$ under $\mathcal{H}_0$ is the well-known estimated least-squares residual vector $\hat{\boldsymbol{e}} \in \mathbb{R}^{n\times 1}$, which is given by

$$\begin{aligned}
\hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}} \\
&= \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y}) \\
&= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{W}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{e})) \\
&= \boldsymbol{e} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{e}) \\
&= (\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W})\boldsymbol{e} \\
&= \boldsymbol{R}\boldsymbol{e},
\end{aligned} \tag{3}$$

with $\hat{\boldsymbol{x}} \in \mathbb{R}^{u\times 1}$ being the BLUE of $\boldsymbol{x}$ under $\mathcal{H}_0$; $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is the known matrix of weights, taken as $\boldsymbol{W} = {\sigma_0}^2\boldsymbol{Q}_{\boldsymbol{e}}^{-1}$, where σ_0^2 is the variance factor, $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ is the identity matrix and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is known

as the redundancy matrix. The $\boldsymbol{R}$ matrix is an orthogonal projector that projects onto the orthogonal complement of the range space of $\boldsymbol{A}$.

We restrict ourselves to regular models, and therefore, the degrees of freedom r (redundancy) of the model under $\mathcal{H}_0$ (Equation (1)) is

$$r = rank(\boldsymbol{Q}_{\hat{e}}) = n - rank(\boldsymbol{A}) = n - u, where \tag{4}$$

$$\boldsymbol{Q}_{\hat{e}} = \boldsymbol{Q}_e - {\sigma_0}^2 \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A})^{-1} \boldsymbol{A}^T \tag{5}$$

On the other hand, an alternative model is proposed when there are doubts about the reliability level of the model under $\mathcal{H}_0$. Here, we assume that the validity of the null hypothesis $\mathcal{H}_0$ in Equation (1) can be violated if the dataset is contaminated by outliers. The model in an alternative hypothesis, denoted by $\mathcal{H}_A$, is to oppose Equation (1) by an extended model that includes the unknown vector $\boldsymbol{\nabla} \in \mathbb{R}^{q \times 1}$ of deterministic bias parameters as follows ([20,35]):

$$\mathcal{H}_A : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{C}\boldsymbol{\nabla} + \boldsymbol{e} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\nabla} \end{pmatrix} + \boldsymbol{e}, \tag{6}$$

where $\boldsymbol{C} \in \mathbb{R}^{n \times q}$ is the matrix that relates bias parameters, i.e., the values of the outliers to observations. We restrict ourselves to the matrix $(\boldsymbol{A} \quad \boldsymbol{C})$ having full column rank, such that

$$r = rank \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} \end{pmatrix} = u + q \leq n \tag{7}$$

One of the most used procedures based on hypothesis testing for outliers in linear (or linearised) models is the well-known data snooping method [19,20]. This procedure consists of screening each individual measurement for the presence of an outlier [42]. In that case, data snooping is based on a local model test, such that $q = 1$, and therefore, the n alternative hypothesis is expressed as

$$\mathcal{H}_A^{(i)} : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{c_i}\nabla_i + \boldsymbol{e} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{c_i} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \nabla_i \end{pmatrix} + \boldsymbol{e}, \forall i = 1, \cdots, n \tag{8}$$

Now, matrix $\boldsymbol{C}$ in Equation (6) 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 the ith bias parameter of magnitude ∇_i affects the ith measurement, and 0 means otherwise. In that case, the rank of $(\boldsymbol{A} \quad \boldsymbol{c_i}) \in \mathbb{R}^{n \times (u+1)}$ and the vector $\boldsymbol{\nabla}$ in Equation (6) reduces to a scalar ∇_i in Equation (8), i.e., $\boldsymbol{c_i}$=$\begin{pmatrix} 0 & 0 & 0 & \cdots & 1^{i_{th}} & 0 & \cdots & 0 \end{pmatrix}^T$. When $q = n - u$, an overall model test is performed. For more details about the overall model test, see, for example, [46,47].

Note that the alternative hypothesis $\mathcal{H}_A^{(i)}$ in Equation (8) is formulated under the condition that the outlier acts as a systematic effect by shifting the random error distribution under $\mathcal{H}_0$ by its own value [13]. In other words, the presence of an outlier in a dataset can cause a shift of the expectation under $\mathcal{H}_0$ to a nonzero value. Therefore, hypothesis testing is often employed to check whether the possible shifting of the random error distribution under $\mathcal{H}_0$ by an outlier is, in fact, a systematic effect (bias) or merely a random effect. This hypothesis test-based approach is called the *mean-shift model* [20]. The mean-shift model has been widely employed in a variety of applications, such as structural deformation analyses, sensor data processing, the integrity monitoring of GNSS (Global Navigation Satellite System) models and the design and quality control of geodetic networks (see, e.g., [1,3,6,8,12,19–22,45,51,57–61]). The alternative to the mean-shift model is variance inflation. Until now, it has been rarely used in geodesy because it is more difficult to derive a powerful test and a reliability theory for it [12,13,62].

2.1. Binary Hypothesis Testing

In the context of the mean-shift model, the test statistic involved in data snooping is given by the normalised least-squares residual, denoted by $\boldsymbol{w_i}$. This test statistic, also known as Baarda's w-test, is given as follows:

$$\boldsymbol{w_i} = \frac{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \hat{\boldsymbol{e}}}{\sqrt{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_i}}}, \forall i = 1, \cdots, n \tag{9}$$

Then, a test decision is performed as [63]

$$Accept \;\; \mathcal{H}_0 \;\; if \;\; |\boldsymbol{w_i}| \leq k, \; reject \; otherwise \; in \; favour \; of \; \mathcal{H}_A^{(i)} \tag{10}$$

Note that the decision rule (10) says that if the Baarda's w-test statistic is larger than some critical value k, i.e., a percentile of its probability distribution, then we reject the null hypothesis in favour of the alternative hypothesis. This is a special case of testing the null hypothesis $\mathcal{H}_0$ against only one single alternative hypothesis $\mathcal{H}_A^{(i)}$, and therefore, the rejection of the null hypothesis automatically implies the acceptance of the alternative hypothesis and vice versa [46,47]. In other words, the outlier detection automatically implies outlier identification and vice versa. This is because the formulation of the alternative hypothesis $\mathcal{H}_A^{(i)}$ is based on the condition that an outlier exists and is located at a pre-specified position in the dataset. In other words, the alternative hypothesis in a binary test says that "*a specific measurement is an outlier*".

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 only one single alternative hypothesis $\mathcal{H}_A^{(i)}$ of (8), it may lead to wrong decisions of Type I and Type II. The probability of a Type I Error α_0 is the probability of rejecting the null hypothesis $\mathcal{H}_0$ when it is true, whereas the probability of a Type II error β_0 is the probability of failing to reject the null hypothesis $\mathcal{H}_0$ when it is false (note: the index '0' represents the case in which a single hypothesis is tested). Instead of α_0 and β_0, there is the confidence level $CL = 1 - \alpha_0$ and the power of the test $\gamma_0 = 1 - \beta_0$, respectively. The first deals with the probability of accepting a true null hypothesis $\mathcal{H}_0$; the second addresses the probability of correctly accepting the alternative hypothesis $\mathcal{H}_A^{(i)}$. In that case, given a probability of a Type I decision error α_0, we find the critical value k_0 as follows:

$$k_0 = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{11}$$

where Φ^{-1} denotes the inverse of the cumulative distribution function (cdf) of the two-tailed standard normal distribution $\boldsymbol{N}(0, 1)$.

The normalised least-squares residual w_i follows a standard normal distribution with the expectation that $\mathbb{E}\{w_i\} = 0$ 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 ith location of the dataset (i.e., under $\mathcal{H}_A^{(i)}$), then the expectation of w_i is

$$\mathbb{E}\{w_i\} = \sqrt{\lambda_0} = \sqrt{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_i} \nabla_i^2} \tag{12}$$

where λ_0 is the non-centrality parameter for $q = 1$. Note, therefore, that there is an outlier that causes the expectation of w_i to become $\sqrt{\lambda_0}$. The square-root of the non-centrality parameter $\sqrt{\lambda_0}$ in Equation (12) represents the expected mean shift of a specific w-test. In such a case, the term $\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_i}$ in Equation (12) is a scalar, and therefore, it can be rewritten as follows [64]:

$$|\nabla_i| = MDB_{0_{(i)}} = \sqrt{\frac{\lambda_0}{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_i}}}, \; \forall i = 1, \cdots, n \tag{13}$$

where $|\nabla_i|$ is the Minimal Detectable Bias ($MDB_{0_{(i)}}$) for the case in which there is only one single alternative hypothesis, which can be computed for each of the n alternative hypotheses according to Equation (8).

For a single outlier, the variance of an estimated outlier, denoted by $\sigma^2_{\nabla_i}$, is

$$\sigma^2_{\nabla_i} = \left(\boldsymbol{c_i}^T \boldsymbol{Q}_e^{-1} \boldsymbol{Q}_{\hat{e}} \boldsymbol{Q}_e^{-1} \boldsymbol{c_i}\right)^{-1}, \ \forall i = 1, \cdots, n \tag{14}$$

Thus, the MDB can also be written as

$$MDB_{0_{(i)}} = \sigma_{\nabla_i} \sqrt{\lambda_0}, \ \forall i = 1, \cdots, n \tag{15}$$

where $\sigma_{\nabla_i} = \sqrt{\sigma^2_{\nabla_i}}$ is the standard deviation of estimated outlier ∇_i.

The MDB in Equations (13) or (15) of an alternative hypothesis is the smallest-magnitude outlier that can lead to the rejection of the null hypothesis $\mathcal{H}_0$ for a given α_0 and β_0. Thus, for each model of the alternative hypothesis $\mathcal{H}_A^{(i)}$, the corresponding MDB can be computed [12,49,65]. The limitation of this MDB is that it was initially developed for the binary hypothesis testing case. In that case, the MDB is a sensitivity indicator of Baarda's w-test when only one single alternative hypothesis is taken into account. In this article, we are confined to multiple alternative hypotheses. Therefore, both the MDB and MIB are computed by considering the case of multiple hypothesis testing.

2.2. Multiple Hypothesis Testing

The alternative hypothesis in Equation (8) has been formulated under the assumption that the measure y_i for some fixed i is an outlier. From a practical point of view, however, we do not know which measurement is an outlier. Therefore, a more appropriate alternative hypothesis would be [22] "*There is at least one outlier in the vector of measurements* y_i". Now, we are interested in knowing which of the alternative hypotheses may lead to the rejection of the null hypothesis with a certain probability. This means testing $\mathcal{H}_0$ against $\mathcal{H}_A^{(1)}, \mathcal{H}_A^{(2)}, \mathcal{H}_A^{(3)}, \ldots, \mathcal{H}_A^{(n)}$. This is known as multiple hypothesis testing (see, e.g., [1,2,12,21,23,37,46,66–69]). In that case, the test statistic coming into effect is the maximum absolute Baarda's w-test value (denoted by max-w), which is computed as [12]

$$\text{max-}w = \max_{i \in \{1, \cdots, n\}} |\boldsymbol{w_i}| \tag{16}$$

The decision rule for this case is given by

$$\begin{aligned} &Accept \ \ \mathcal{H}_0 \ \ if \ \ \text{max-}w \leq \hat{k} \\ Otherwise, & \\ &Accept \ \ \mathcal{H}_A^{(i)} \ \ if \ \ \text{max-}w > \hat{k} \end{aligned} \tag{17}$$

The decision rule in 17 says that if none of the n w-tests get rejected, then we accept the null hypothesis $\mathcal{H}_0$. If the null hypothesis $\mathcal{H}_0$ is rejected in any of the n tests, then one can only assume that detection occurred. In other words, if the max-w is larger than some percentile of its probability distribution (i.e., some critical value $\hat{k}$), then there is evidence that there is an outlier in the dataset. Therefore, "outlier detection" only informs us whether the null hypothesis $\mathcal{H}_0$ is accepted or not.

However, the detection does not tell us which alternative hypothesis $\mathcal{H}_A^{(i)}$ would have led to the rejection of the null hypothesis $\mathcal{H}_0$. The localisation of the alternative hypothesis, which would have rejected the null hypothesis, is a problem of "outlier identification". Outlier identification implies the execution of a search among the measurements for the most likely outlier. In other words, one seeks to find which of Baarda's w-test is the maximum absolute value max-w and if that max-w is greater than some critical value $\hat{k}$.

Therefore, the data snooping procedure of screening measurements for possible outliers is actually an important case of multiple hypothesis testing and not single hypothesis testing. Moreover, note that outlier identification only happens when outlier detection necessarily exists; i.e., "outlier identification" only occurs when the null hypothesis $\mathcal{H}_0$ is rejected. However, correct detection does not necessarily imply correct identification [2,12,46].

3. Probability Levels of Data Snooping for a Single Run under Multiple Alternative Hypotheses

Two sides of the multiple testing problem can be formulated: one under the reality of the null hypothesis $\mathcal{H}_0$, i.e., the event that there is no outlier in the dataset, and another one coinciding with the alternative hypothesis $\mathcal{H}_A^{(i)}$, i.e., the event that there is an outlier. The probability levels associated with data snooping for both events are presented in Table 1.

Table 1. Probability levels associated with data snooping under multiple alternative hypotheses.

Reality Unknown	**Result of the Test**				
	$\boldsymbol{\mathcal{H}_0}$	$\boldsymbol{\mathcal{H}_A^{(1)}}$	$\boldsymbol{\mathcal{H}_A^{(2)}}$	$\cdots$	$\boldsymbol{\mathcal{H}_A^{(n)}}$
$\mathcal{H}_0$	Correct decision $1-\alpha'$	Type I Error α_{01}	Type I Error α_{02}	$\cdots$	Type I Error α_{0n}
$\mathcal{H}_A^{(1)}$	Type II error β_{10}	Correct identification $1-\beta_{11}$	Type III error κ_{12}	$\cdots$	Type III error κ_{1n}
$\mathcal{H}_A^{(2)}$	Type II error β_{20}	Type III error κ_{21}	Correct identification $1-\beta_{22}$	$\cdots$	Type III error κ_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathcal{H}_A^{(n)}$	Type II error β_{n0}	Type III error κ_{n1}	Type III error κ_{n2}	$\cdots$	Correct identification $1-\beta_{nn}$

3.1. On the Scenario Coinciding with the Null Hypothesis $\mathcal{H}_0$

For the scenario coinciding with the null hypothesis $\mathcal{H}_0$, there is the probability of incorrectly identifying at least one alternative hypothesis. This type of wrong decision is known as the *family-wise error rate* (*FWE*). The *FWE* is defined as

$$FWE = \alpha_{0i} = \mathcal{P}\left(|\boldsymbol{w_i}| > |\boldsymbol{w_j}|\ \forall j,\ |\boldsymbol{w_i}| > k(i \neq j) \;\middle|\; \mathcal{H}_0 : true\right),\ \forall i = 1, \ldots, n \tag{18}$$

The probability of accepting the null hypothesis in test i is $1-\alpha$, $\forall i = 1, \ldots, n$, where α is the significance level or size of the test for single hypothesis testing. The classical and well-known procedure to control the *FWE* is the Bonferroni correction [56]. If all tests are mutually independent, then the probability that a true $\mathcal{H}_0$ is accepted in each test is approximately

$$(1-\alpha)^n = 1-\alpha' \tag{19}$$

where α' is the Type I Error for the entire dataset. Thus, we have

$$\alpha = 1-(1-\alpha')^{1/n} \tag{20}$$

which is approximately

$$\alpha = \frac{\alpha'}{n} \tag{21}$$

The quantity in Equation (21) is just equal to the upper bound of the Bonferroni inequality, i.e., $\alpha' \leq n\alpha$ [56].

Controlling the FWE at a pre-specified level α' corresponds to controlling the probability of a Type I decision error when carrying out a single test. In other words, one uses a global Type I Error rate α'

that combines all tests under consideration instead of an individual error rate α that only considers one test at a time [69]. In that case, the critical value k_{bonf} is computed as

$$k_{bonf} = \Phi^{-1}\left(1 - \frac{\alpha'}{2n}\right) \tag{22}$$

For single hypothesis testing, given a probability of a Type I decision error α_0, it is easier for us to find the critical value using Equation (11). On the other hand, the rate of Type I decision errors for multiple testing, α', cannot be directly controlled by the user. One can argue about the application of Bonferroni [56] using Equation (22). However, Bonferroni is a good approximation for the case in which alternative hypotheses are independent. In practice, however, the test results always depend on each other to some degree because we always have a correlation between w-tests. The correlation coefficient between any Baarda's w-test statistic (denoted by ρ_{w_i,w_j}), such as $\boldsymbol{w_i}$ and $\boldsymbol{w_j}$, is given by [21]

$$\rho_{w_i,w_j} = \frac{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_j}}{\sqrt{\boldsymbol{c_i}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_i}} \sqrt{\boldsymbol{c_j}^T \boldsymbol{Q_e}^{-1} \boldsymbol{Q_{\hat{e}}} \boldsymbol{Q_e}^{-1} \boldsymbol{c_j}}}, \forall (i \neq j) \tag{23}$$

The correlation coefficient ρ_{w_i,w_j} can assume values within the range $[-1, 1]$.

Here, the extreme normalised residuals max-w (i.e., maximum absolute) in Equation (16) are treated directly as a test statistic. Note that when using Equation (16) as a test statistic, the decision rule is based on a one-sided test of the form max-$w \leq \hat{k}$. However, the distributions of max-w cannot be derived from well-known test distributions (e.g., normal distribution). Therefore, critical values cannot be taken from a statistical table but must be computed numerically. This problem has already been addressed by Lehmann [22]. In that case, the dependencies between residuals are not neglected because the critical values are based on the distribution of max-w, which depends on the correlation between w-test statistics ρ_{w_i,w_j}.

According to Equation (23), the correlation ρ_{w_i,w_j} depends on the matrices $\boldsymbol{A}$ and $\boldsymbol{Q_e}$, and therefore, the distribution of max-w also depends on these matrices. In other words, the critical value depends on the uncertainty of the measurement sensor and the mathematical model of the problem.

In order to guarantee the user-defined Type I decision error α' for data snooping, the critical value must be computed by Monte Carlo.

The key of Monte Carlo is artificial random numbers (*ARN*) [70], which are called 'artificial' because the random numbers are generated using a deterministic process. A random number generator is a technology designed to generate a deterministic sequence of numbers that do not have any pattern and therefore appear to be random. It is '*random*' in the sense that the sequence of numbers generated passes statistical tests for randomness. For this reason, random number generators are typically referred to as pseudo-random number generators (PRNGs).

A PRNG simulates a sequence of independent and identically distributed (i.i.d.) numbers chosen uniformly between 0 and 1. PRNGs are part of many machine learning and data mining techniques. In a simulation, a PRNG is implemented as a computer algorithm in some programming language and is made available to the user via procedure calls or icons [71]. A good generator produces numbers that are not distinguishable from truly random numbers in limited computation time. This is particularly true for Mersenne Twister, a popular generator with the long period length of $2^{199371}-1$ [72].

In essence, Monte Carlo replaces random variables with computer *ARN*, probabilities with relative frequencies and expectations with arithmetic means over large sets of such numbers [12]. A computation with one set of *ARN* is a Monte Carlo experiment [33].

The procedure to compute the critical value of max-w is given step-by-step as follows:

1. Specify the probability density function (pdf) of the w-test statistics. The pdf assigned to the w-test statistics under an $\mathcal{H}_0$-distribution is

$$(w_1, w_2, w_3, \cdots, w_n)^T \sim N(\mathbf{0}, \boldsymbol{\mathcal{R}}_w) \tag{24}$$

where $\boldsymbol{\mathcal{R}}_w \in \mathbb{R}^{n \times n}$ is the correlation matrix with the main diagonal elements equal to 1, and the off-diagonal elements are the correlation between the w-test statistics computed by Equation (23).

2. In order to have w-test statistics under $\mathcal{H}_0$, uniformly distributed random number sequences are produced by the Mersenne Twister algorithm, and then they are transformed into a normal distribution by using the Box–Muller transformation [73]. Box–Muller has already been used in geodesy for Monte Carlo experiments [22,33,74]. Therefore, a sequence of m random vectors from the pdf assigned to the w-test statistics is generated according to Equation (24). In that case, we have a sequence of m vectors of the w-test statistics as follows:

$$\left[(w_1, w_2, w_3, \cdots, w_n)^{T^{(1)}}, (w_1, w_2, w_3, \cdots, w_n)^{T^{(2)}}, \cdots, (w_1, w_2, w_3, \cdots, w_n)^{T^{(m)}}\right] \tag{25}$$

3. Compute the test statistic by Equation (16) for each sequence of w-test statistics. Thus, we have

$$\left(\max_{i \in \{1,\cdots,n\}} |\boldsymbol{w_i}|^{(1)}, \max_{i \in \{1,\cdots,n\}} |\boldsymbol{w_i}|^{(2)}, \cdots, \max_{i \in \{1,\cdots,n\}} |\boldsymbol{w_i}|^{(m)}\right) \tag{26}$$

4. Sort in ascending order the maximum test statistic in Equation (26), getting a sorted vector $\tilde{w}$, such that

$$\tilde{w}^{(1)} < \tilde{w}^{(2)}, \tilde{w}^{(3)}, \cdots, < \tilde{w}^{(m)} \tag{27}$$

The sorted values $\tilde{w}$ in Equation (27) provide a discrete representation of the cumulative density function (cdf) of the maximum test statistic max-w.

5. Determine the critical value $\hat{k}$ as follows:

$$\hat{k} = \tilde{w}_{[(1-\alpha') \times m]} \tag{28}$$

where [.] denotes rounding down to the next integer that indicates the position of the selected elements in the ascending order of $\tilde{w}$. This position corresponds to a critical value for a stipulated overall false alarm probability α'. This can be done for a sequence of values α' in parallel.

It is important to mention that the probability of a Type I decision error for multiple testing α' is larger than that of Type I for single testing α_0. This is because the critical region in multiple testing is larger than that in single hypothesis testing.

3.2. On the Scenario Coinciding with the Alternative Hypothesis $\mathcal{H}_A^{(i)}$

The other side of the multiple testing problem is the situation in which there is an outlier in the dataset. In that case, apart from Type I and Type II errors, there is a third type of wrong decision associated with Baarda's w-test. Baarda's w-test can also flag a non-outlying observation while the '*true*' outlier remains in the dataset. We are referring to the Type III error [67], also referred to as the probability of wrong identification ($\mathcal{P}_{WI}$). The description of the Type III error (denoted by κ_{ij} in Table 1) involves a separability analysis between alternative hypotheses [2,21,23,66]. Therefore, we are now interested in the identification of the correct alternative hypothesis. In that case, the non-centrality parameter in Equation (12) is not only related to the sizes of Type I and Type II decision errors but also dependent on the correlation coefficient ρ_{w_i,w_j} given by Equation (23).

On the basis of the assumption that one outlier is in the ith position of the dataset (i.e., $\mathcal{H}_A^{(i)}$ is '*true*'), the probability of a Type II error (also referenced as the probability of "*missed detection*", denoted by $\mathcal{P}_{MD}$) for multiple testing is

$$\mathcal{P}_{MD} = \beta_{i0} = \mathcal{P}\left(\bigcap_{i=1}^{n} |\boldsymbol{w_i}| \leq \hat{k} \;\middle|\; \mathcal{H}_A^{(i)} : true\right), \tag{29}$$

and the size of a Type III wrong decision (also called "*misidentification*", denoted by $\mathcal{P}_{WI}$) is given by

$$\mathcal{P}_{WI} = \sum_{i=1}^{n} \kappa_{ij} = \sum_{i=1}^{n} \mathcal{P}\left(|\boldsymbol{w_j}| > |\boldsymbol{w_i}| \; \forall i, \; |\boldsymbol{w_j}| > \hat{k}(i \neq j) \;\middle|\; \mathcal{H}_A^{(i)} : true\right) \tag{30}$$

On the other hand, the probability of correct identification (denoted by $\mathcal{P}_{CI}$) is

$$\mathcal{P}_{CI} = 1 - \beta_{ii} = \mathcal{P}\left(|\boldsymbol{w_i}| > |\boldsymbol{w_j}| \; \forall j, \; |\boldsymbol{w_i}| > \hat{k}(i \neq j) \;\middle|\; \mathcal{H}_A^{(i)} : true\right) \tag{31}$$

with

$$1 - \mathcal{P}_{CI} = \beta_{ii} = \beta_{i0} + \sum_{i=1}^{n} \kappa_{ij}, \; for \; (i \neq j) \tag{32}$$

Note that the three probabilities of missed detection $\mathcal{P}_{MD}$, wrong identification $\mathcal{P}_{WI}$ and correct identification $\mathcal{P}_{CI}$ sum up to unity: i.e., $\mathcal{P}_{MD} + \mathcal{P}_{WI} + \mathcal{P}_{CI} = 1$.

The probability of correct detection $\mathcal{P}_{CD}$ is the sum of the probability of correct identification $\mathcal{P}_{CI}$ (selecting a correct alternative hypothesis) and the probability of misidentification $\mathcal{P}_{WI}$ (selecting one of the $n-1$ other hypotheses), i.e.,

$$\mathcal{P}_{CD} = \mathcal{P}_{CI} + \mathcal{P}_{WI} \tag{33}$$

The probability of wrong identification $\mathcal{P}_{WI}$ is identically zero, $\mathcal{P}_{WI} = 0$, when the correlation coefficient is exactly zero, $\rho_{w_i,w_j} = 0$. In that case, we have

$$\mathcal{P}_{CD} = \mathcal{P}_{CI} = 1 - \mathcal{P}_{MD} \tag{34}$$

The relationship given in Equation (34) would only happen if one neglected the nature of the dependence between alternative hypotheses. In other words, this relationship is valid for the special case of testing the null hypothesis $\mathcal{H}_0$ against only one single alternative hypothesis $\mathcal{H}_A^{(i)}$.

Since the critical region in multiple hypothesis testing is larger than that in single hypothesis testing, the Type II decision error (i.e., $\mathcal{P}_{MD}$) for the multiple test becomes smaller [12]. This means that the correct detection in binary hypothesis testing (γ_0) is smaller than the correct detection $\mathcal{P}_{CD}$ under multiple hypothesis testing, i.e.,

$$\mathcal{P}_{CD} > \gamma_0 \tag{35}$$

Detection is easier in the case of multiple hypothesis testing than single hypothesis testing. However, the probability of correct detection $\mathcal{P}_{CD}$ under multiple testing is spread out over all alternative hypotheses, and therefore, identifying is harder than detecting. From Equation (33), it is also noted that detection does not depend on identification. However, outlier identification depends on correct outlier detection. Therefore, we have the following inequality:

$$\mathcal{P}_{CI} \leq \mathcal{P}_{CD} \tag{36}$$

Note that the probability of correct identification $\mathcal{P}_{CI}$ depends on the probability of missed detection $\mathcal{P}_{MD}$ and wrong identification $\mathcal{P}_{WI}$ for the case in which data snooping is run only once, i.e., a single round of estimation and testing. However, in this paper, we deal with data snooping in its iterative form (i.e., *IDS*), and therefore, the probability of correct identification $\mathcal{P}_{CI}$ depends on other decision rules.

4. On the Probability Levels of Iterative Outlier Elimination and Its Sensitivity Indicators

In the previous section, probability levels are described for the case in which the data snooping procedure is applied only once according to the detector given by Equation (16). In practice, however, data snooping is applied iteratively: after identification and elimination of a single outlier, the model is reprocessed, and outlier identification is restarted. This procedure of iterative outlier elimination is known as iterative data snooping (*IDS*) [35]. Therefore, *IDS* is not only a case of multiple hypothesis testing but also a case of multiple runs of estimation, testing and adaptation. In that case, adaptation consists of removing a possible outlier.

Rofatto et al. [12,55] showed how to compute the probability levels associated with the *IDS* procedure. They introduced two new classes of wrong decisions for *IDS*, namely, over-identification positive and over-identification negative. The first is the probability of *IDS* flagging the outlier and good observations simultaneously. The second is the probability of *IDS* flagging only the good observations as outliers (more than one) while the outlier remains in the dataset.

This paper extends the current decision errors of *IDS* for the case in which there is a single outlier in the dataset. In addition to the probability levels described so far, there is the probability that the detector in (16) simultaneously flags two (or more) observations during a round of *IDS*. Here, this is referred to as *statistical overlap*. Statistical overlap occurs when two (or more) Baarda's w-test statistics are equal. For instance, if the correlation coefficient between two w-test statistics (ρ_{w_i,w_j}) were exactly 1.00 (or -1.00), i.e., if $\rho_{w_i,w_j} = \pm 1.00$, then the alternative hypothesis, say, $\mathcal{H}_A^{(i)}$, would have the same distribution as another one, $\mathcal{H}_A^{(j)}$. Th would mean that those hypotheses would not be distinguished, i.e., they would not be separable, and an outlier would not be identified [2]. Note that the correlation ρ_{w_i,w_j} provides an indication of whether or not the system redundancy is sufficient to identify an outlier. When the correlation coefficient between two w-test statistics is exactly 1.00 (or -1.00), i.e., $\rho_{w_i,w_j} = \pm 1.00$, a statistical overlap $\mathcal{P}_{ol}$ is expected to occur. We further discuss $\mathcal{P}_{ol}$ when we present the results.

In contrast to the data snooping single run, the success rate of correct detection $\mathcal{P}_{CD}$ for *IDS* depends on the sum of the probabilities of correct identification $\mathcal{P}_{CI}$, wrong exclusion ($\mathcal{P}_{WE}$), over-identification cases ($\mathcal{P}_{over+}$ and $\mathcal{P}_{over-}$), and statistical overlap ($\mathcal{P}_{ol}$), i.e.,

$$\mathcal{P}_{CD} = 1 - \mathcal{P}_{MD} = \mathcal{P}_{CI} + \mathcal{P}_{WE} + \mathcal{P}_{over+} + \mathcal{P}_{over-} + \mathcal{P}_{ol} \tag{37}$$

It is important to mention that the probability of correct detection is the complement of the probability of missed detection. Note from Equation (39) that the probability of correct detection $\mathcal{P}_{CD}$ is available even for cases in which the identification rate is null, $\mathcal{P}_{CI} = 0$. However, the probability of correct identification ($\mathcal{P}_{CI}$) necessarily requires that the probability of correct detection $\mathcal{P}_{CD}$ be greater than zero. For the same reasons given for the data snooping single run in the previous section, detecting is easier than identifying. In that case, we have the following relationship for the success rate of correct outlier identification $\mathcal{P}_{CI}$:

$$\mathcal{P}_{CI} = \mathcal{P}_{CD} - (\mathcal{P}_{WE} + \mathcal{P}_{over+} + \mathcal{P}_{over-} + \mathcal{P}_{ol}), \tag{38}$$

such as

$$\exists(\mathcal{P}_{CI}) \in [0,1] \iff (\mathcal{P}_{CD}) > 0 \tag{39}$$

It is important to mention that the wrong exclusion $\mathcal{P}_{WE}$ describes the probability of identifying and removing a non-outlying measurement while the '*true*' outlier remains in the dataset. In other words, $\mathcal{P}_{WE}$ is the Type III decision error for *IDS*). The overall wrong exclusion $\mathcal{P}_{WE}$ is the result of the sum of each individual contribution to $\mathcal{P}_{WE}$, i.e.,

$$\mathcal{P}_{WE} = \sum_{i=1}^{n-1} \mathcal{P}_{WE(i)} \tag{40}$$

We can also compute a weighting factor, denoted by $p_{i(\mathcal{P}_{WE})}$, for each individual contribution to $\mathcal{P}_{WE}$ as follows:

$$p_{i(\mathcal{P}_{WE})} = \frac{\mathcal{P}_{WE(i)}}{\mathcal{P}_{WE}}, \forall i = 1, \ldots, n-1, \tag{41}$$

so that

$$\sum_{i=1}^{n-1} p_{i(\mathcal{P}_{WE})} = \frac{\sum_{i=1}^{n-1} \mathcal{P}_{WE(i)}}{\mathcal{P}_{WE}} \tag{42}$$

The weighting factor $p_{i(\mathcal{P}_{WE})}$ is within a range of [0,1].

On the basis of the probability levels of correct detection $\mathcal{P}_{CD}$ and correct identification $\mathcal{P}_{CI}$, the sensitivity indicators of minimal biases—Minimal Detectable Bias (MDB) and Minimal Identifiable Bias (MIB)—for a given α' can be computed as follows:

$$MDB = \arg\min_{\nabla_i} \mathcal{P}_{CD}(\nabla_i) > \tilde{\mathcal{P}}_{CD}, \forall i = 1, \cdots, n \tag{43}$$

$$MIB = \arg\min_{\nabla_i} \mathcal{P}_{CI}(\nabla_i) > \tilde{\mathcal{P}}_{CI}, \forall i = 1, \cdots, n \tag{44}$$

Equation (43) gives the smallest outlier ∇_i that leads to its detection for a given correct detection rate $\tilde{\mathcal{P}}_{CD}$, whereas (44) provides the smallest outlier ∇_i that leads to its identification for a given correct identification rate $\tilde{\mathcal{P}}_{CI}$.

As a consequence of the inequality in (36), the MIB will be larger than MDB, i.e., $MIB \geq MDB$. For the special case of having only one single alternative hypothesis, there is no difference between the MDB and MIB [46]. The computation of MDB_0 is easily performed by Equations (13) or (15), whereas the computation of the MDB in Equation (43) and the MIB in Equation (44) must be computed using Monte Carlo because the acceptance region (as well as the critical region) for the case of multiple alternative hypotheses is analytically intractable.

The non-centrality parameter for detection ($\lambda_{q=1}^{(MDB)}$) and identification ($\lambda_{q=1}^{(MIB)}$) for *IDS* can be computed similarly to Equation (12) as follows, respectively:

$$\lambda_{q=1}^{(MDB)} = \frac{MDB_{(i)}^2}{\sigma_{\nabla_i}^2} \tag{45}$$

$$\lambda_{q=1}^{(MIB)} = \frac{MIB_{(i)}^2}{\sigma_{\nabla_i}^2} \tag{46}$$

Thus,

$$\frac{MIB_{(i)}}{MDB_{(i)}} = \sqrt{\frac{\lambda_{q=1}^{(MIB)}}{\lambda_{q=1}^{(MDB)}}} \tag{47}$$

Note from Equation (47) that the relationship between the non-centrality parameters for detection ($\lambda_{q=1}^{(MDB)}$) and identification ($\lambda_{q=1}^{(MIB)}$) do not depend on the variance (or standard deviation) of estimated outlier $\sigma_{\nabla_i}^2$.

In the case of *IDS*, the power depends not only on the rate of Type II and Type III decision errors but also on the rate of over-identifications and the probability of statistical overlap. In the next section, we provide a procedure for computing the errors and success rates associated with *IDS*.

5. Computational Procedure for the Estimation of Success and Failure Rates of Iterative Outlier Elimination

After finding the critical value $\hat{k}$ by the process described in Section 3.1, the procedure based on Monte Carlo is also applied to compute the probability levels of *IDS* when there is an outlier in the dataset as follows (summarised as a flowchart in Figure 1).

First, random error vectors are synthetically generated on the basis of a multivariate normal distribution because the assumed stochastic model for random errors is based on the matrix covariance of the observations. Here, we use the Mersenne Twister algorithm [72] to generate a sequence of random numbers and Box–Muller [73] to transform it into a normal distribution.

The magnitude intervals of simulated outliers are user-defined. The magnitude intervals are based on the standard deviation of the observation, e.g., $|3\sigma|$ to $|6\sigma|$, where σ is the standard deviation of the observations. Since the outlier can be positive or negative, the proposed algorithm randomly selects the signal of the outlier (for $q = 1$). Here, we use the discrete uniform distribution to select the signal of the outlier. Thus, the total error ($\boldsymbol{\varepsilon}$) is a combination of random errors, and its corresponding outlier is as follows:

$$\boldsymbol{\varepsilon} = \boldsymbol{e} + \boldsymbol{c_i}\nabla_i \tag{48}$$

In Equation (48), $\boldsymbol{e}$ is the random error generated from the normal distribution according to Equation (2), and the second part $\boldsymbol{c_i}\nabla_i$ is the additional parameter that describes the alternative model according to Equation (8). Next, we compute the least-squares residuals vector according to Equation (3), but now we use the total error ($\boldsymbol{\varepsilon}$) as follows:

$$\hat{\boldsymbol{e}} = \boldsymbol{R}\boldsymbol{\varepsilon} \tag{49}$$

For *IDS*, the hypothesis of (8) for $q = 1$ (one outlier) is assumed, and the corresponding test statistic is computed according to (9). Then, the maximum test statistic value is computed according to Equation (16). Now, the decision rule is based on the critical value $\hat{k}$ computed by Monte Carlo (see the steps (24)–(28) from Section 3.1). After identifying the measurement suspected to be the most likely outlier, it is excluded from the model, and least-squares estimation (LSE) and data snooping are applied iteratively until there are no further outliers identified in the dataset. Every time that a measurement suspected to be the most likely outlier is removed from the model, we check whether the normal matrix $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$ is invertible or not. If the determinant of $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$ is 0, $det|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}| = 0$, then there is a necessary and sufficient condition for a square matrix $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$ to be non-invertible. In other words, we check whether or not there is a solution available in the sense of ordinary LSE after removing a possible outlier.

The *IDS* procedure is performed for m experiments of random error vectors for each experiment contaminated by an outlier in the ith measurement. Therefore, for each measurement contaminated by an outlier, there are $v = 1, \ldots, m$ experiments.

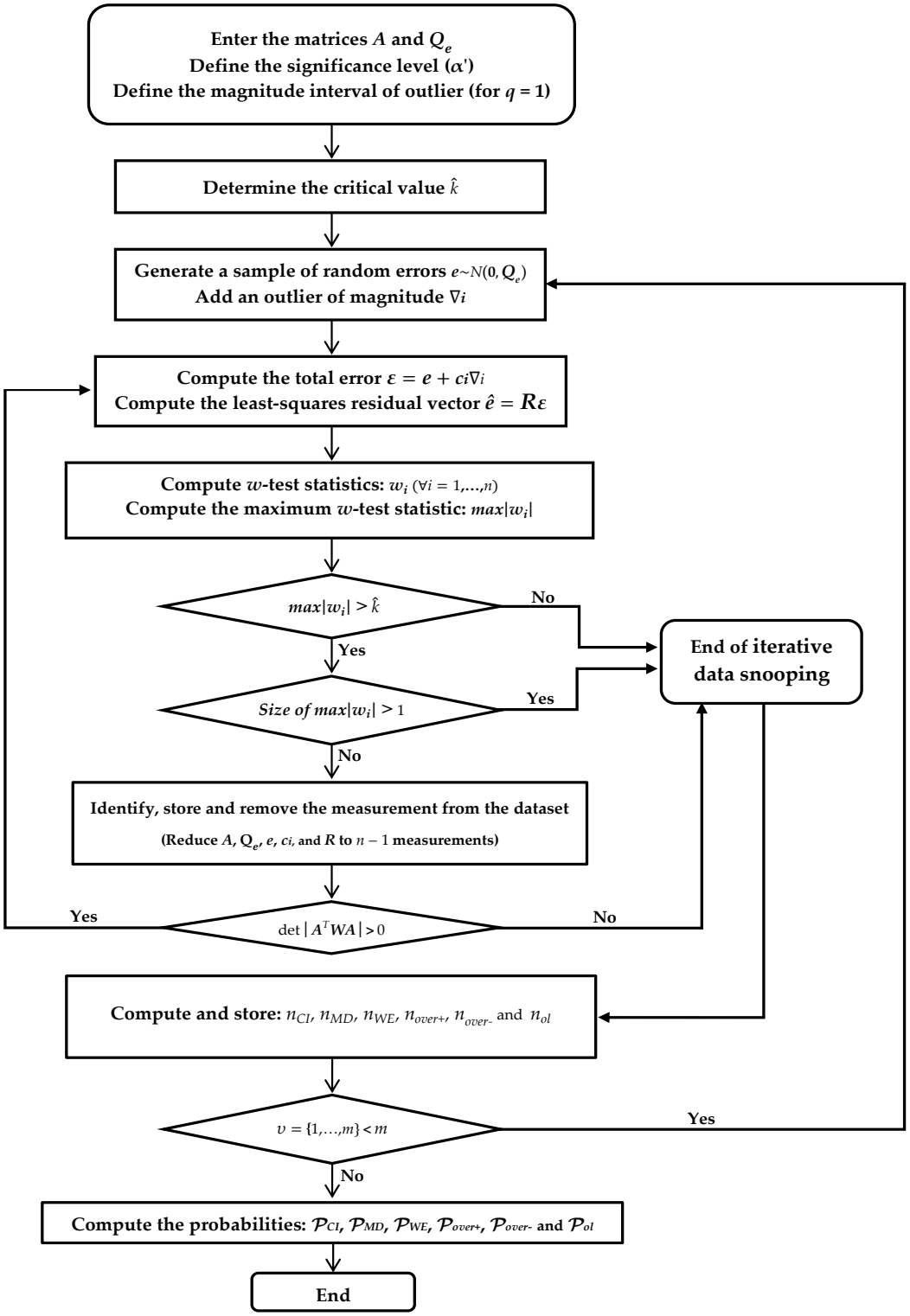

Figure 1. Flowchart of the algorithm to compute the probability levels of Iterative Data Snooping (*IDS*) for each measurement in the presence of an outlier.

The probability of correct identification $\mathcal{P}_{CI}$ is obtained by the ratio between the correct identification cases and possible cases. Thus, if m is the total number of Monte Carlo experiments

(possible cases), then we count the number of times that the outlier is correctly identified (denoted as n_{CI}) and then approximate the probability of correct identification $\mathcal{P}_{CI}$ as

$$\mathcal{P}_{CI} = \frac{n_{CI}}{m} \tag{50}$$

Similar to Equation (50), false decisions are computed as

$$\mathcal{P}_{MD} = \frac{n_{MD}}{m} \tag{51}$$

$$\mathcal{P}_{WE} = \frac{n_{WE}}{m} \tag{52}$$

$$\mathcal{P}_{over_+} = \frac{n_{over_+}}{m} \tag{53}$$

$$\mathcal{P}_{over_-} = \frac{n_{over_-}}{m} \tag{54}$$

$$\mathcal{P}_{ol} = \frac{n_{ol}}{m} \tag{55}$$

where:

- n_{MD} is the number of experiments in which *IDS* does not detect the outlier ($\mathcal{P}_{MD}$ corresponds to the rate of missed detection);
- n_{WE} is the number of experiments in which the *IDS* procedure flags and removes only one single non-outlying measurement while the '*true*' outlier remains in the dataset ($\mathcal{P}_{WE}$ is the wrong exclusion rate);
- n_{over_+} is the number of experiments in which *IDS* correctly identifies and removes the outlying measurement and others, and $\mathcal{P}_{over_+}$ corresponds to its probability;
- n_{over_-} represents the number of experiments in which *IDS* identifies and removes more than one non-outlying measurement, whereas the '*true outlier*' remains in the dataset ($\mathcal{P}_{over_-}$ is the probability corresponding to this error probability class); and
- n_{ol} is the number of experiments in which the detector in Equation (16) flags two (or more) measurements simultaneously during a given iteration of *IDS*. Here, this is referred to as the number of statistical overlap n_{ol}, and $\mathcal{P}_{ol}$ corresponds to its probability.

In contrast to [12], in this paper, the probability levels associated with *IDS* are evaluated for each observation individually and for each outlier magnitude. Furthermore, we take care to control the *family-wise error rate*. In the next sections, we show the application of the algorithm described in Figure 1 to compute statistical quantities for *IDS*.

6. On the Probability Levels of Iterative Outlier Elimination

In this experiment, we considered two closed levelling networks: one with a low correlation between residuals and another one with a high correlation. For the low correlation case, we used the network given by Rofatto et al. [60], whereas for the high correlation, we chose the network from Knight et al. [43]. Figures 2a,b show the configuration of these networks, respectively.

Figure 3 shows an example of levelling for a single line. The equipment used to measure the level difference is an electronic digital level. In this case, the instrument is designed to operate by employing electronic digital image processing and is used in conjunction with a special barcoded staff (also called a barcode rod). After an operator accomplishes rough levelling with a bull's eye bubble, an electronic digital level can be used to digitally obtain the barcoded staff reading. The result can be recorded manually in a levelling field-book or automatically stored in the data collector of the digital level. A simple height difference is obtained from the difference between the backsight staff reading and the foresight staff reading. An example of a "*digital level – barcode staff*" system is displayed in Figure 4. For more details about digital levels, see, for example [75–77].

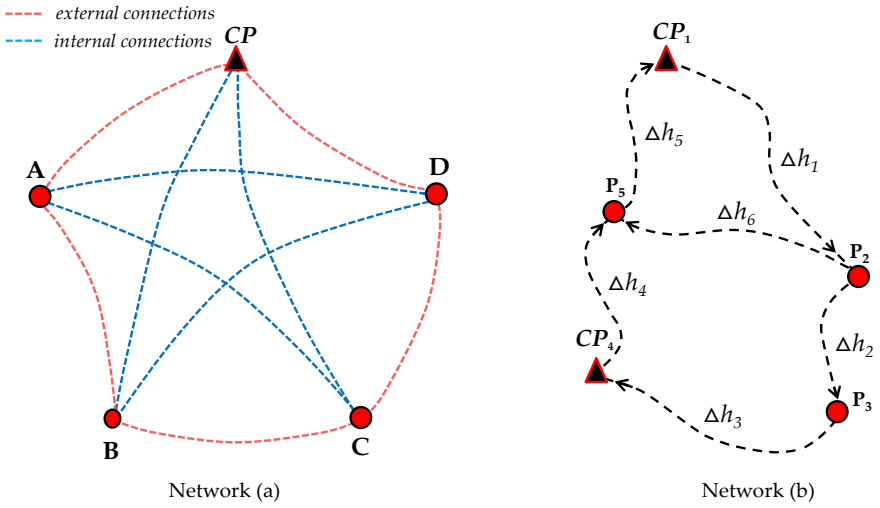

Figure 2. Levelling geodetic networks: (**a**) Levelling network adapted from [60] with a low correlation between *w*-test statistics; (**b**) Levelling network adapted from [43] with a high correlation between *w*-test statistics.

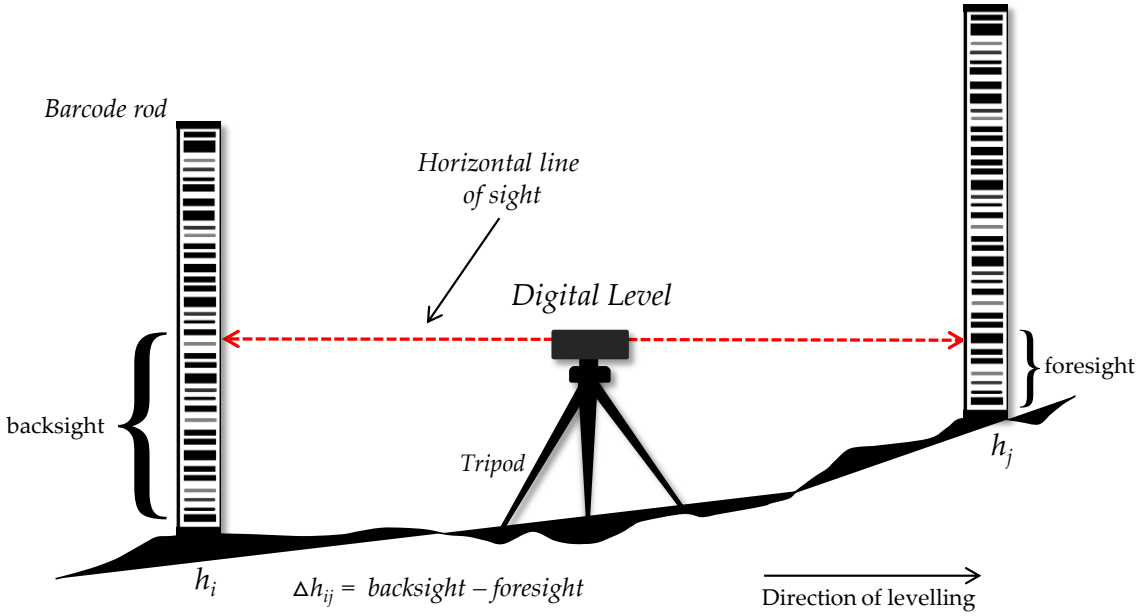

Figure 3. Example of levelling for a single line: the digital level instrument is placed between the points whose height difference is to be found. Special barcoded staffs are placed at these points and then sighted through the digital level instrument.

Figure 4. Example of a *digital level – barcode staff* system.

There are several levelling lines available for a levelling geodetic network. In the absence of outliers, i.e., under $\mathcal{H}_0$, the model for levelling a geodetic network can be written in the sense of the standard Gauss–Markov model in Equation (1) as follows:

$$\Delta h_{i-j} + e_{\Delta h_{i-j}} = h_j - h_i, \tag{56}$$

where Δh_{i-j} is the height difference measured from point i to j, and $e_{\Delta h_{i-j}}$ is the random error associated with the levelling measurement. Generally, one of these points has a known height h, from which the height of another point is determined. The point with the known height is referred to here as the control point or benchmark (denoted by CP).

From network (**a**) in Figure 2, we have one control point and four points with unknown heights (A, B, C and D) for a total of four minimally constrained points. The control point is fixed (hard constraint or non-stochastic variable). In that case, matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \tag{57}$$

In Figure 2, the red dashed lines are the external connections among the points of the network (**a**), whose measurements are Δh_{A-CP}, Δh_{A-B}, Δh_{B-C}, Δh_{C-D} and Δh_{D-CP}, whereas the blue dashed lines are the internal connections, whose measurements consist of Δh_{A-D}, Δh_{A-C}, Δh_{B-CP}, Δh_{B-D}

and Δh_{C-CP}. The distances of the external and internal connections are considered 240 m and 400 m, respectively. The equipment used is a spirit level with a nominal standard deviation for a single staff reading of 0.02 mm/m. Lines of sight distances are kept at 40 m. Each partial height difference, in turn, involves one instrument setup and two sightings: forward and back. Thus, the standard deviation for each total height difference equals

$$\begin{aligned} \sigma_{\Delta h_{i-j}} &= \sqrt{2p} \times 40\,\text{m} \times \frac{0.02\,\text{mm}}{m} \\ &= \sqrt{2p} \times 0.8\,\text{mm}, \end{aligned} \tag{58}$$

where p is the number of partial height differences. In this case, each total height difference between external or internal connections is made of, respectively, three or five partial height differences. The readings are assumed to be uncorrelated and have equal uncertainty. In this case, the standard deviations of the measures for external and internal connections are 1.96 mm and 2.53 mm, respectively.

On the other hand, from network (**b**) in Figure 2, there are two control stations (fixed) and three user-stations with unknown heights. Matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix} \tag{59}$$

For network (**b**), we have the following measurements: $\Delta h_1 = \Delta h_{CP_1-P_2}$, $\Delta h_2 = \Delta h_{P_2-P_3}$, $\Delta h_3 = \Delta h_{P_3-CP_4}$, $\Delta h_4 = \Delta h_{CP_4-P_5}$, $\Delta h_5 = \Delta h_{P_5-CP_1}$ and $\Delta h_6 = \Delta h_{P_2-P_5}$. In this case, the covariance matrix of the measurements (metro units) is given by [43]

$$\boldsymbol{Q}_e = \begin{pmatrix} 5.5 & 3.7 & 0.3 & -3.2 & -0.5 & 0.1 \\ 3.7 & 3.9 & 0.0 & -0.8 & -0.6 & -0.7 \\ 0.3 & 0.0 & 0.8 & -1.4 & 0.1 & 0.8 \\ -3.2 & -0.8 & -1.4 & 5.4 & -0.3 & -2.1 \\ -0.5 & -0.6 & 0.1 & -0.3 & 0.2 & 0.3 \\ 0.1 & -0.7 & 0.8 & -2.1 & 0.3 & 1.4 \end{pmatrix} \tag{60}$$

The correlation coefficients ρ_{w_i,w_j} between w-test statistics were computed for both network (**a**) and network (**b**) according to Equation (23). Table 2 provides the correlation ρ_{w_i,w_j} for network (**a**), and Table 3 provides it for network (**b**). In general, the correlation ρ_{w_i,w_j} for network (**b**) is much higher than that for network (**a**).

From Table 2, we observe that the maximum correlation is $\rho_{w_i,w_j} = \pm 0.4146$ for network (**a**) (i.e., $\rho_{w_i,w_j} = \pm 41.46\%$). In this case, as the correlation coefficient is less than 50%, the missed detection rate $\mathcal{P}_{MD}$ is expected to be larger than the other decision errors of *IDS*.

From Table 3, it is expected that the wrong exclusion rate $\mathcal{P}_{WE}$ will be significantly more pronounced than other wrong decisions of *IDS*. This is because of the high correlation between test statistics for network (**b**). Note also that the correlation coefficient between the second (Δh_2) and third Δh_3 measurement is exactly 1.00 (i.e., $\rho_{w_i,w_j} = 100\%$). This means that if one of these measurements is an outlier, then its corresponding w-test statistics will overlap. Therefore, an outlier can never be identified if it occurs in one of these measurements, but it can be detected.

Table 2. Correlation matrix of w-test statistics for levelling network (**a**).

Δh_{i-j}	$A-CP$	$A-B$	$C-B$	$D-C$	$CP-D$	$D-A$	$C-A$	$CP-B$	$D-B$	$C-CP$
$A-CP$	1.0000	−0.4146	−0.0488	−0.0488	−0.4146	−0.3464	−0.3134	−0.3464	−0.0660	−0.3134
$A-B$		1.0000	0.4146	0.0488	0.0488	−0.3134	−0.3464	0.3464	0.3134	0.0660
$C-B$			1.0000	0.4146	0.0488	−0.0660	−0.3464	−0.3134	−0.3464	0.3134
$D-C$				1.0000	0.4146	−0.3134	0.3134	−0.0660	−0.3464	−0.3464
$CP-D$					1.0000	0.3464	0.0660	−0.3134	0.3134	−0.3464
$D-A$						1.0000	−0.2565	−0.0223	−0.2565	0.0223
$C-A$							1.0000	0.0223	−0.0223	0.2565
$CP-B$								1.0000	−0.2565	−0.2565
$D-B$									1.0000	−0.0223
$C-CP$										1.0000

Table 3. Correlation matrix of w-test statistics for levelling network (**b**).

Δh_{i-j}	Δh_1	Δh_2	Δh_3	Δh_4	Δh_5	Δh_6
Δh_1	1.00	−0.41	−0.41	0.96	0.98	0.97
Δh_2		1.00	1.00	−0.36	−0.50	−0.61
Δh_3			1.00	−0.36	−0.50	−0.61
Δh_4				1.00	0.98	0.93
Δh_5					1.00	0.98
Δh_6						1.00

In the following subsections, we compute and analyse the probability levels associated with *IDS* for two cases. In the first part, the dataset is considered to be free of outliers, whereas, in the second one, there is an outlier in the dataset.

6.1. Scenario Coinciding with the Null Hypothesis: Absence of Outliers

In this context, we investigated the extent to which the Bonferroni correction deviates from the distribution of max-w on the basis of Monte Carlo. For this purpose, we considered the following Type I decision error rates: $\alpha' = 0.001$, $\alpha' = 0.0027$, $\alpha' = 0.01$, $\alpha' = 0.025$, $\alpha' = 0.05$ and $\alpha' = 0.1$. It is important to mention that the probability of a Type I Error of $\alpha' = 0.0027$ corresponds to a critical value of $k = 3$ in the case of a single test, which is known as the 3σ-rule [32]. We also set up $m = 200{,}000$ Monte Carlo experiments, as proposed by [12]. For each α', we computed the corresponding critical value according to Bonferroni from Equation (22) and on the basis of Monte Carlo from the procedure described in Section 3, specifically from step (24) to (28). Both methods were applied for networks (**a**) and (**b**) in Figure 2. The result is displayed in Figure 5.

From Figure 5, we observe that the critical values computed from Monte Carlo are always smaller than those values computed by Equation (22) for both networks. This is because matrix $\boldsymbol{R}$ in Equation (3) promotes the correlation between residuals. Note that matrix $\boldsymbol{R}$ depends on the network geometry given by matrix $\boldsymbol{A}$. This means that we will always have some degree of correlation between residuals. If we neglect the correlation between residuals, as in (22), then we are assuming that there is no association between residuals. Thus, we overestimate the probability of max-w by using (22), and the dashed curve in Figure 5 is always above the solid curve. Therefore, with Bonferroni, the user has no full control over Type I Errors. We point out some particularities as follows.

The Bonferroni method can only be used with a good approximation to control the Type I Error of *IDS* for the case in which we have a measurement system with high redundancy and low residual correlation (i.e., low correlation between w-test statistics ρ_{w_i,w_j}) and for small α'. This is the case for network (**a**). On the other hand, Bonferroni does not work well for network (**b**). In the latter case, the effect is more pronounced because of the low redundancy and high residual correlation. In general, under the condition that the correct decision is made when the null hypothesis $\mathcal{H}_0$ is accepted, it is less likely to get a large max-w than the prediction by (22).

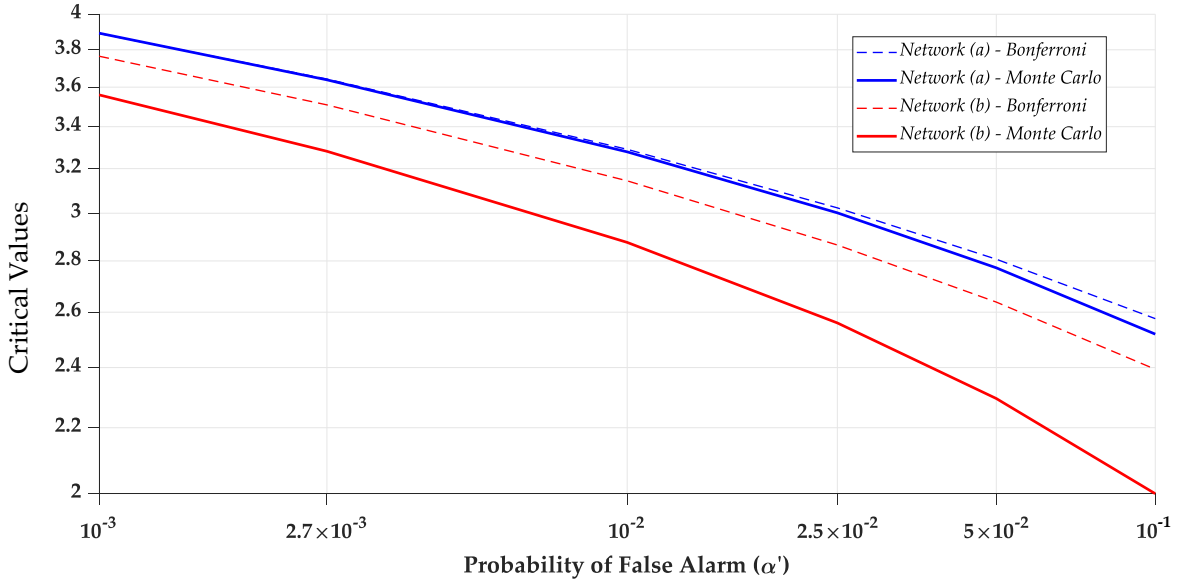

Figure 5. Critical values of max-w for both levelling networks (**a**) in blue and (**b**) in red. *Solid curves* were obtained from the Monte Carlo procedure in Section 4 with m = 200,000 Monte Carlo experiments. Dashed curves were obtained from Bonferroni in (22).

Here, we treated the extreme (i.e., maximum absolute) normalised residuals max-w in (16) directly as a test statistic. Figure 6 shows the distribution of max-w for both networks (**a**) and (**b**). We observe that the distribution of max-w for network (**a**) gets closer to a normal distribution than network (**b**). This means that the smaller the correlation between residuals, the smaller the Type I decision error rate α' and, therefore, the larger the critical values. Figure 7 shows the cumulative distribution of max-w for both networks (**a**) and (**b**). Figure 7 reveals that, for the same level of error probability, the critical value for network (**a**) is always smaller than the one computed for network (**b**).

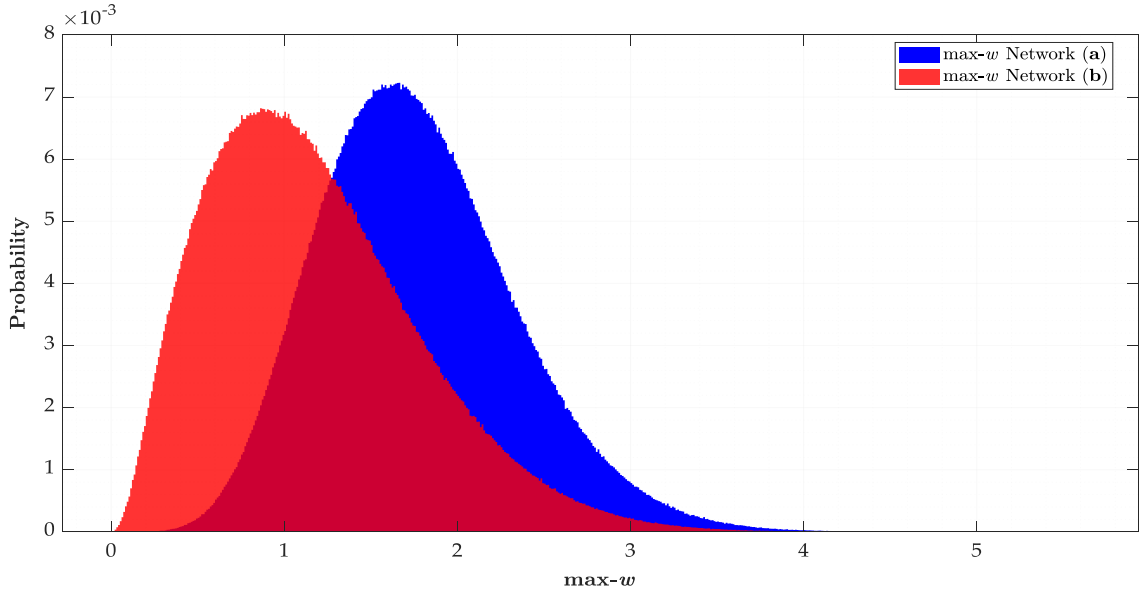

Figure 6. Probability histograms of max-w for networks (**a**) in blue and (**b**) in red.

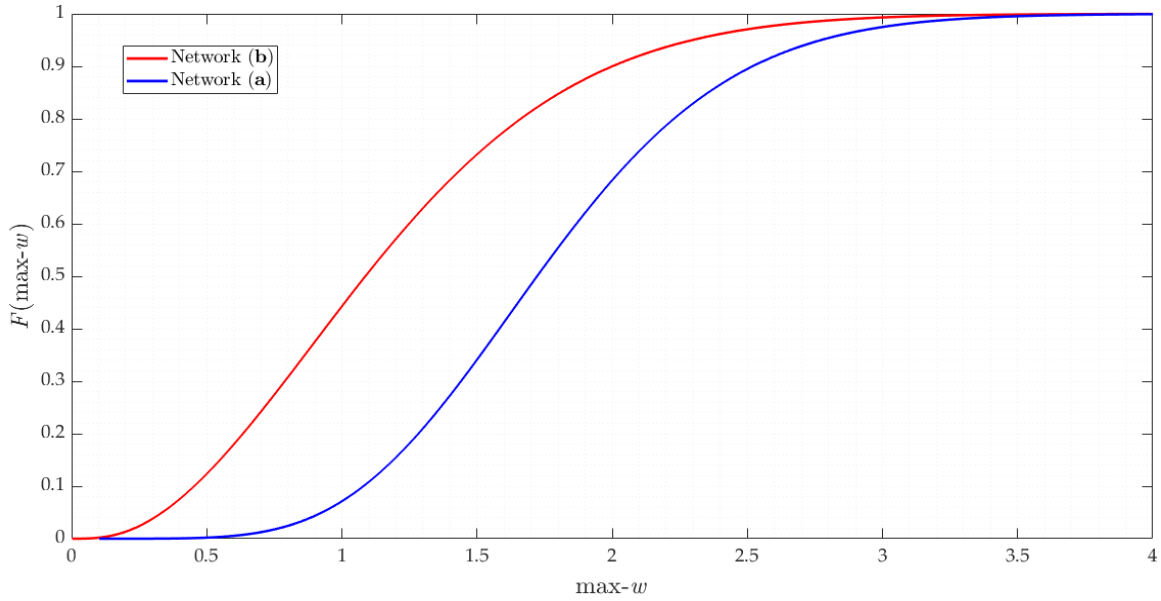

Figure 7. Cumulative frequency of max-w for networks (**a**) in blue and (**b**) in red.

Until now, our outcomes have been investigated under the condition that $\mathcal{H}_0$ is true. In the next subsection, we analyse the probability levels of *IDS* in the presence of an outlier. In that case, the decision rule is based on critical values from the max-w distribution, as detailed in Table 4. It is important to note that the critical value 3 of the 3σ-rule is not valid for a multiple test case. In fact, the critical values for outlier identification (i.e., a multiple test) depend on the geometry of the network and the sensor uncertainty. Therefore, the probability associated with the 3σ-rule for network (**a**) will be close to $\alpha' = 0.025$, and that for network (**b**) will be close to $\alpha' = 0.0067$.

Table 4. Critical values from the Bonferroni and Monte Carlo procedures for networks (**a**) and (**b**).

α'	k_{bonf} from (22) for Net (a)	k_{bonf} from (22) for Net (b)	$\hat{k}$ from (28) for Net (a)	$\hat{k}$ from (28) for Net (b)
0.001	3.89	3.76	3.89	3.56
0.0027	3.64	3.51	3.64	3.28
0.01	3.29	3.14	3.28	2.88
0.025	3.02	2.87	3	2.56
0.05	2.81	2.64	2.77	2.29
0.1	2.58	2.39	2.52	2

6.2. Scenario Coinciding with an Alternative Hypothesis: Presence of an Outlier

In this scenario, there is an outlier in the dataset. Thus, the correct decision is made when the alternative hypothesis $\mathcal{H}_A^{(i)}$ from Equation (8) is accepted. In this step, we computed the probability levels of *IDS* using the procedure in Section 5. The decision rule is based on critical values computed by Monte Carlo. These values are presented in Table 4. We arbitrarily defined the outlier magnitude from $|3\sigma|$ to $|8\sigma|$ for network (**a**) and $|1\sigma|$ to $|12\sigma|$ for network (**b**).

The sensitivity indicators MDB and MIB were also computed according to Equations (43) and (44), respectively. The success rates for outlier detection and outlier identification were taken as equal to 0.8, i.e., $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 0.8$, respectively.

6.2.1. Geodetic Network with Low Correlation between Residuals

We start from network (**a**) with a low correlation between residuals. We observe that there is a high degree of homogeneity for network (**a**). This can be explained by the redundancy numbers, denoted by (r_i). The redundancy numbers are the elements of the main diagonal of matrix $\boldsymbol{R}$ in Equation (3). The redundancy number (r_i) is an internal reliability measure that represents the ability

of a measurement to project the measurement error in the least-squares residuals. Then, the higher the number of redundancy, the higher the resistance of the measurement to outliers.

The redundancy numbers for the measurements constituting external connections are identical and equal to $r_i = 0.519$, whereas the measurements constituting external connections are also identical but equal to $r_i = 0.681$. Consequently, the probability levels associated with *IDS* are practically identical for both external and internal connections. Thus, we subdivided our result into two parts: mean values of the probability levels for external connections and mean values of the probability levels for internal connections.

Figure 8 shows the probability of correct identification ($\mathcal{P}_{CI}$) and correct detection ($\mathcal{P}_{CD}$) in the presence of an outlier for both external and internal connections. In general, we observe that the larger the Type I Error α' (or the lower the critical value $\hat{k}$), the higher the rate of correct detection $\mathcal{P}_{CD}$. This is not fully true for outlier identification $\mathcal{P}_{CI}$.

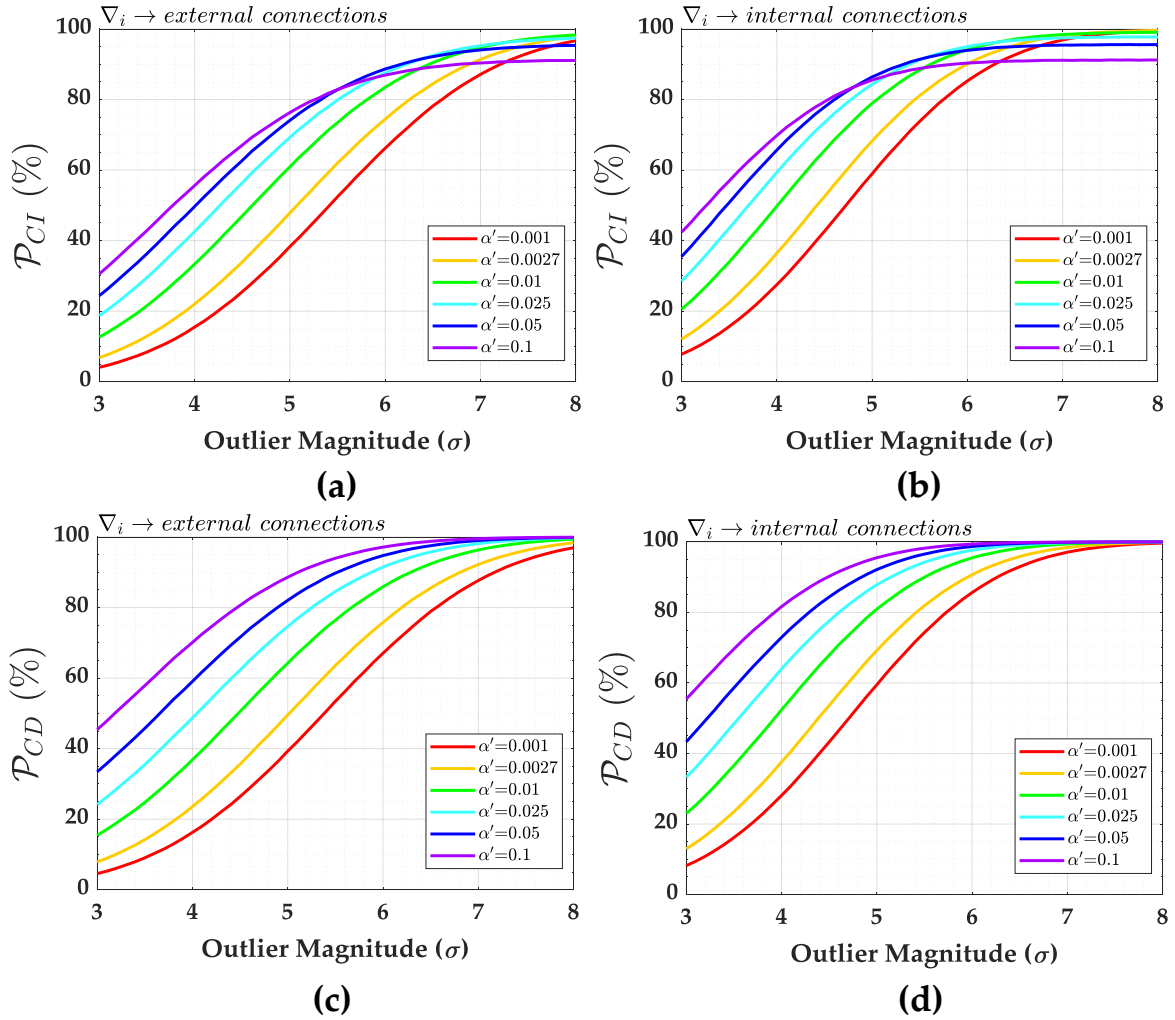

Figure 8. Probability of correct identification ($\mathcal{P}_{CI}$) and correct detection ($\mathcal{P}_{CD}$) for network (**a**): (**a**) $\mathcal{P}_{CI}$ for the measurements constituting external connections. (**b**) $\mathcal{P}_{CI}$ for the measurements constituting internal connections. (**c**) $\mathcal{P}_{CD}$ for the measurements constituting external connections. (**d**) $\mathcal{P}_{CD}$ for the measurements constituting internal connections.

We observe that the probability of correct identification $\mathcal{P}_{CI}$ becomes constant from a certain outlier magnitude. Moreover, the larger the α', the faster the success rate at which outlier identification $\mathcal{P}_{CI}$ stabilizes. In other words, the larger the α', the higher the $\mathcal{P}_{CI}$, but only up to a certain limit of outlier magnitude. After this bound, there is an inversion: the larger the α', the lower the probability

of correct identification $\mathcal{P}_{CI}$. This can be explained by the following: (i) the larger the α', the larger the critical region (or the smaller the acceptance region) of the working hypothesis $\mathcal{H}_0$; (ii) the larger the critical region, the smaller the size of the test; (iii) the smaller the size of the test, the less likely the hypothesis test will identify a small difference. In other words, there is no significant difference among the probabilities of correct identification $\mathcal{P}_{CI}$ for outliers lying within a certain location of the critical region. Therefore, the probabilities of correct identification $\mathcal{P}_{CI}$ for those outliers are practically identical.

Let us take the probability of correct identification $\mathcal{P}_{CI}$ for external connections in Figure 8 as an example. If the user chooses $\alpha' = 0.1$, then the test will be limited to 90% of the acceptance region. In this case, an outlier of 6.6σ will have a practically identical probability of correct identification of $\mathcal{P}_{CI} = 90\%$ of an outlier of 8σ (or greater than 8σ). However, if one chooses an $\alpha' = 0.001$ (99.9% of acceptance region), then a 6.6σ outlier would not be identified at the same rate as an 8σ outlier. Therefore, in that case, the Type I decision error α' (or the critical value $\hat{k}$) restricts the maximum rate of correct outlier identification $\mathcal{P}_{CI}$.

Note also that there are no significant differences between detection $\mathcal{P}_{CD}$ and identification $\mathcal{P}_{CI}$ rates for small Type I decision errors (see, e.g., $\alpha' = 0.001$ and $\alpha' = 0.0027$).

Furthermore, the probabilities of correct detection $\mathcal{P}_{CD}$ and identification $\mathcal{P}_{CI}$ are greater for internal than external connections. For an outlier of 4.5σ and $\alpha' = 0.1$, for instance, the probability of correct identification is $\mathcal{P}_{CI} = 67\%$ for external connections, whereas, for internal connections, it is $\mathcal{P}_{CI} = 80\%$.

Next, we compared the sensitivity indicators MDB and MIB by considering a success rate of 0.8 (80%) for both outlier identification and outlier detection, i.e., $\tilde{\mathcal{P}}_{CI} = \tilde{\mathcal{P}}_{CD} = 80\%$ (see Equations (43) and (44). The user can also find the MDB and MIB for other success rates. The result is displayed in Figure 9. We observe that the larger the Type I decision error α', the more that the MDB deviates from the MIB. In other words, the MIB stabilizes for a certain α', whereas the MDB continues to decrease. It is harder to identify than it is to detect an outlier. Therefore, the MIB will always be greater than or equal to the MDB.

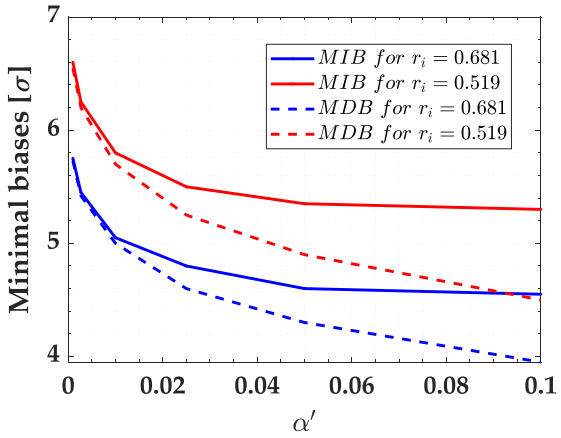

Figure 9. Minimal Detectable Bias (MDB) and Minimal Identifiable Bias (MIB) for both external and internal connections and for each α'.

The standard deviations of estimated outlier σ_{∇_i} for external and internal connection measurements are 2.7 mm and 3 mm, respectively. These σ_{∇_i} values were obtained by means of the square-root in Equation (14). Note from Equations (45) and (46) that the higher the accuracy of the outlier estimate, the lower the MDB and MIB, respectively. However, note from Equation (47) that the relationship between the MIB and MDB does not depend on $\sigma^2_{\nabla_i}$. This is true when the outlier is treated as bias. In other words, if outliers are treated as bias, then they act like systematic errors

by shifting the random error distribution by their own value [13]. The result for $\tilde{\mathcal{P}}_{CI} = \tilde{\mathcal{P}}_{CD} = 0.8$ is summarised in Table 5.

As can be seen from Table 5, in general, the MIB does not deviate too much from the MDB. This is because of a low correlation between residuals. The difference becomes larger when the Type I decision error α' is increased. Note, for instance, that the MDB and MIB are practically identical for Type I decision errors of $\alpha' = 0.001$ and $\alpha' = 0.0027$. In other words, an outlier is detected and identified with the same probability level when there is a low correlation between residuals and for small α'. Therefore, we observe that the larger the α', the greater the difference between the MIB and MDB. In this case, the difference between the MIB and MDB is governed by the user-defined α'.

Table 5. Relationship between the MDB and MIB for network (**a**) by considering $\tilde{\mathcal{P}}_{CI} = \tilde{\mathcal{P}}_{CD} = 0.8$.

α'	(External Connections)			(Internal Connections)		
	$\lambda_{q=1}^{(MDB)}$	$\lambda_{q=1}^{(MIB)}$	MIB/MDB	$\lambda_{q=1}^{(MDB)}$	$\lambda_{q=1}^{(MIB)}$	MIB/MDB
0.001	22.27	22.61	1.01	22.36	22.52	1.00
0.0027	19.95	20.27	1.01	20.01	20.23	1.01
0.01	16.86	17.46	1.02	17.03	17.37	1.01
0.025	14.3	15.7	1.05	14.41	15.69	1.04
0.05	12.46	14.85	1.09	12.59	14.41	1.07
0.1	10.51	14.58	1.18	10.63	14.10	1.15

From Table 6, it can also be noted that the MIB is higher for internal than external connections. This is because internal connections are less precise than external connections. Therefore, the effect on the heights (model parameters) of an unidentified outlier is greater if the outlier magnitude is equal to the MIB of the internal connections. However, from Figure 8, we observe that it would be easier to identify an outlier if it occurred in the measurements that constitute internal connections than if it occurred in external connections.

Table 6. MIBs for each α' and for $\tilde{\mathcal{P}}_{CI} = 80\%$.

α'	MIB (m) (External Connections)	MIB (m) (Internal Connections)
0.001	0.0129	0.0145
0.0027	0.0122	0.0138
0.01	0.0114	0.0128
0.025	0.0108	0.0121
0.05	0.0105	0.0116
0.1	0.0104	0.0115

It is important to mention that both the MDB and MIB are 'invariant' with respect to the control point position *CP*. This is a well-known fact and can already follow from the MDB and MIB definitions in Equations (45) and (46), respectively, which show that both the MDB and MIB are driven by the variance matrices of the measurements and adjusted residuals.

Figure 10 provides the result for the Type III decision error ($\mathcal{P}_{WE}$). In the worst case, we have $\mathcal{P}_{WE} = 0.12$ (12%) for $|3\sigma|$. In general, $\mathcal{P}_{WE}$ is larger for external than internal connections. This is linked to the fact that the residual correlation ρ_{w_i,w_j} in Table 2 is higher for external than internal connections. Furthermore, the larger the Type I error rate α', the larger the $\mathcal{P}_{WE}$ for both internal and external connections. Because of the low probability of $\mathcal{P}_{WE}$ decision errors for network (**a**), the user may opt for a larger α' so that the Type II decision error $\mathcal{P}_{MD}$ is as small as possible. Thus, it is possible to guarantee a high outlier identification rate. This kind of analysis can be performed, for instance, during the design stage of a geodetic network (see, e.g., [60]).

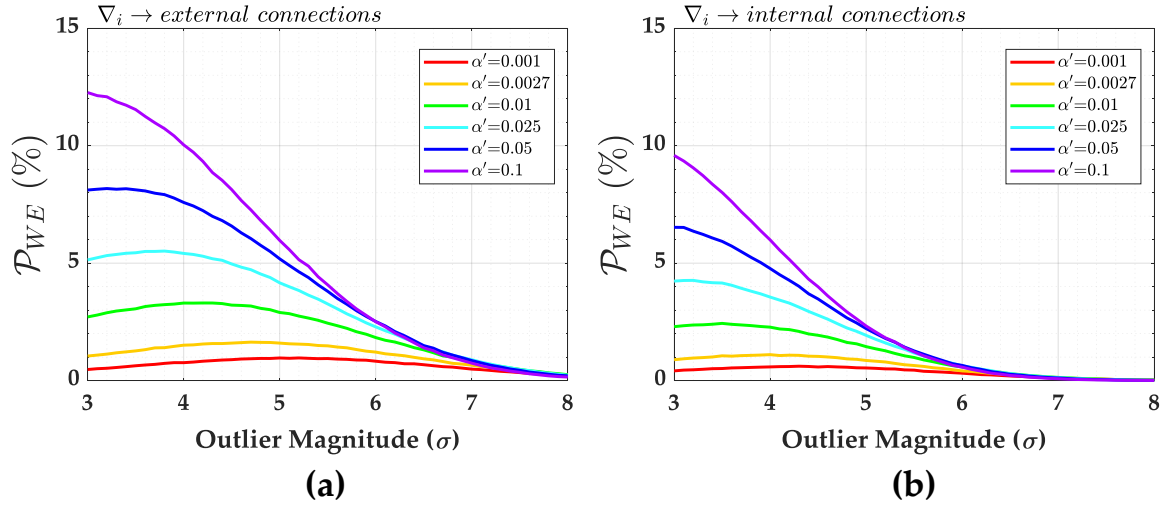

Figure 10. Probability of committing a Type III decision error ($\mathcal{P}_{WE}$) for network (**a**): (**a**) $\mathcal{P}_{WE}$ for external connections. (**b**) $\mathcal{P}_{WE}$ for internal connections.

Figure 10 gives only the overall rate of $\mathcal{P}_{WE}$. Figure 11, on the other hand, displays the individual contributions to $\mathcal{P}_{WE}$ according to Equation (40) for $\alpha' = 0.1$. As expected, the higher the correlation coefficient between w-test statistics ρ_{w_i,w_j}, the greater the contribution of the measurement to $\mathcal{P}_{WE}$ (see, e.g., [2]). In that case, we can also verify from Figure 12 that the larger the redundancy number r_i, the smaller the $\mathcal{P}_{WE}$. Moreover, the larger the outlier magnitude, the smaller the $\mathcal{P}_{WE}$. We also observe from Figure 13 that the larger the ρ_{w_i,w_j}, the larger the weighting factor $p_{i(\mathcal{P}_{WE})}$. The weighting factors $p_{i(\mathcal{P}_{WE})}$ for the highest correlations (i.e., $\rho_{w_i,w_j} = 0.415$ for $r_i = 0.519$ and $\rho_{w_i,w_j} = 0.346$ for $r_i = 0.681$) increase as the outlier magnitude increases. However, this is not significant. While the weighting factor $p_{i(\mathcal{P}_{WE})}$ for the highest correlation coefficient increases by around 1%, the overall $\mathcal{P}_{WE}$ decreases by around 20%. In general, the weighting factor $p_{i(\mathcal{P}_{WE})}$ is relatively constant. The weighting factor $p_{i(\mathcal{P}_{WE})}$ was obtained by Equation (41).

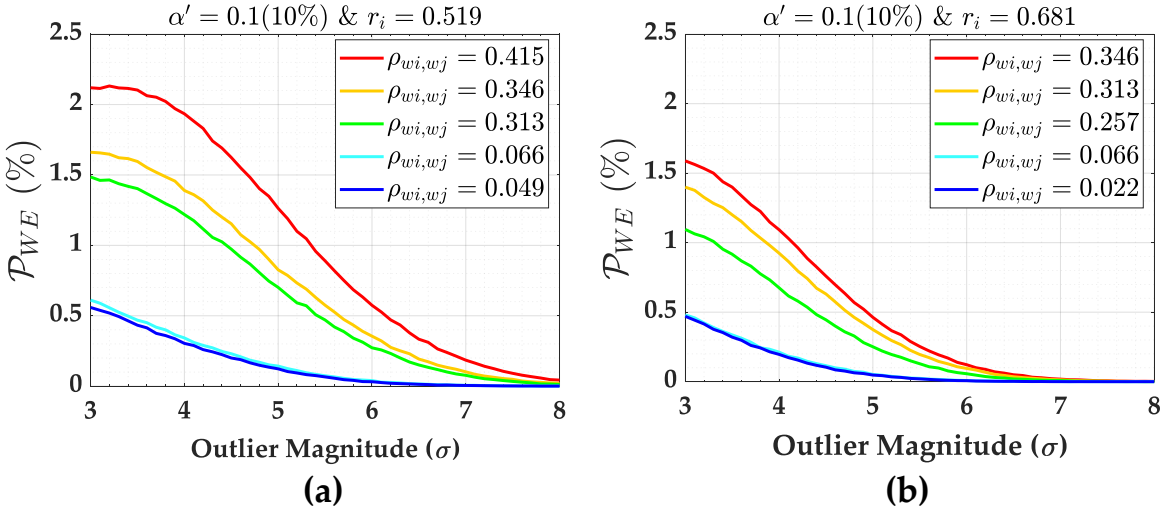

Figure 11. Individual contribution of external and internal connections to the overall $\mathcal{P}_{WE}$ and their relationship with the correlation coefficient ρ_{w_i,w_j} for network (**a**). (**a**) Individual contribution to $\mathcal{P}_{WE}$ by external connections. (**b**) Individual contribution to $\mathcal{P}_{WE}$ by internal connections.

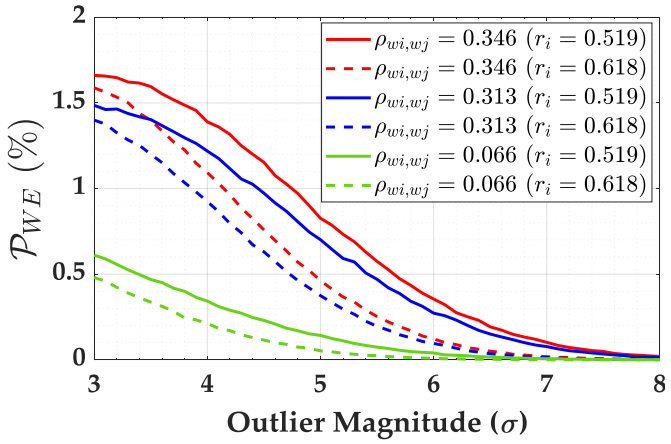

Figure 12. Individual contributions to the overall $\mathcal{P}_{WE}$ for network (**a**) and their relationship with the redundancy number r_i for residuals with the same correlation coefficient ρ_{w_i,w_j} and $\alpha' = 0.1$.

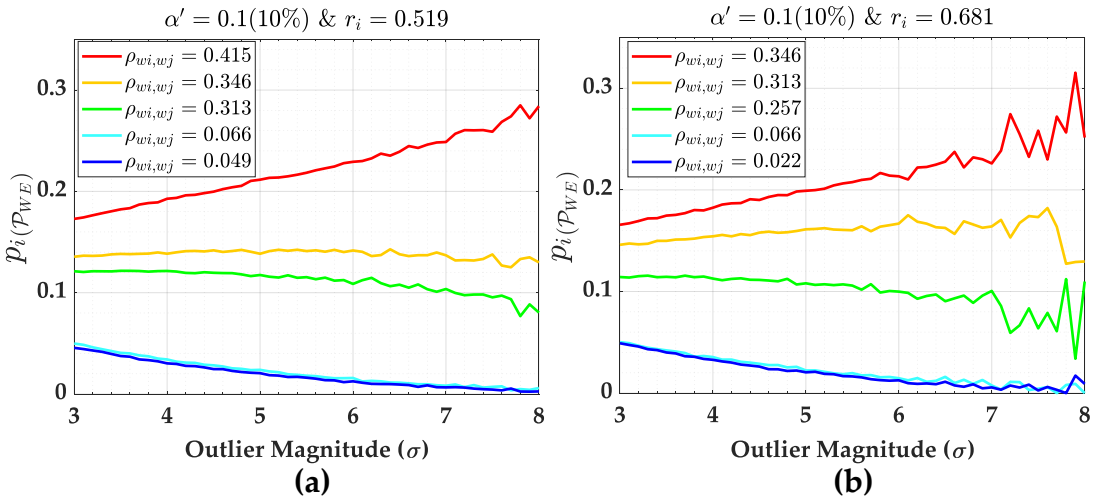

Figure 13. Weighting factors $p_{i(\mathcal{P}_{WE})}$ for network (**a**) and for $\alpha' = 0.1$. (**a**) Weighting factors $p_{i(\mathcal{P}_{WE})}$ for external connections. (**b**) Weighting factors $p_{i(\mathcal{P}_{WE})}$ for internal connections.

The over-identification cases $\mathcal{P}_{over+}$ and $\mathcal{P}_{over-}$ are presented in Figure 14. In general, the larger the Type I decision error α', the larger the over-identification cases for that network. The larger the magnitude of the outlier, the larger the $\mathcal{P}_{over+}$ and smaller the $\mathcal{P}_{over-}$. For small α', we observe that $\mathcal{P}_{over-}$ and $\mathcal{P}_{over+}$ are practically null (see, e.g., for $\alpha' = 0.001$ and $\alpha' = 0.0027$). In general, the larger the correlation coefficient ρ_{w_i,w_j}, the smaller the $\mathcal{P}_{over+}$ and the larger the $\mathcal{P}_{over-}$. Moreover, we also observe that the larger the redundancy number r_i, the larger the $\mathcal{P}_{over+}$ and the smaller the $\mathcal{P}_{over-}$.

The probability of statistical overlap $\mathcal{P}_{ol}$ is practically null for this network. This is because each point of network (**a**) has at least four connections. This means that even with an exclusion, there are still three measurement levels per point (i.e., three connections per point), which guarantees the minimum redundancy necessary for the second round of *IDS*. The very low residual correlation of this network also contributes to the non-occurrence of statistical overlap.

The results presented so far are valid for the case of a system with high redundancy and low residual correlation. In the next section, we present the results for a system with low redundancy and high residual correlation.

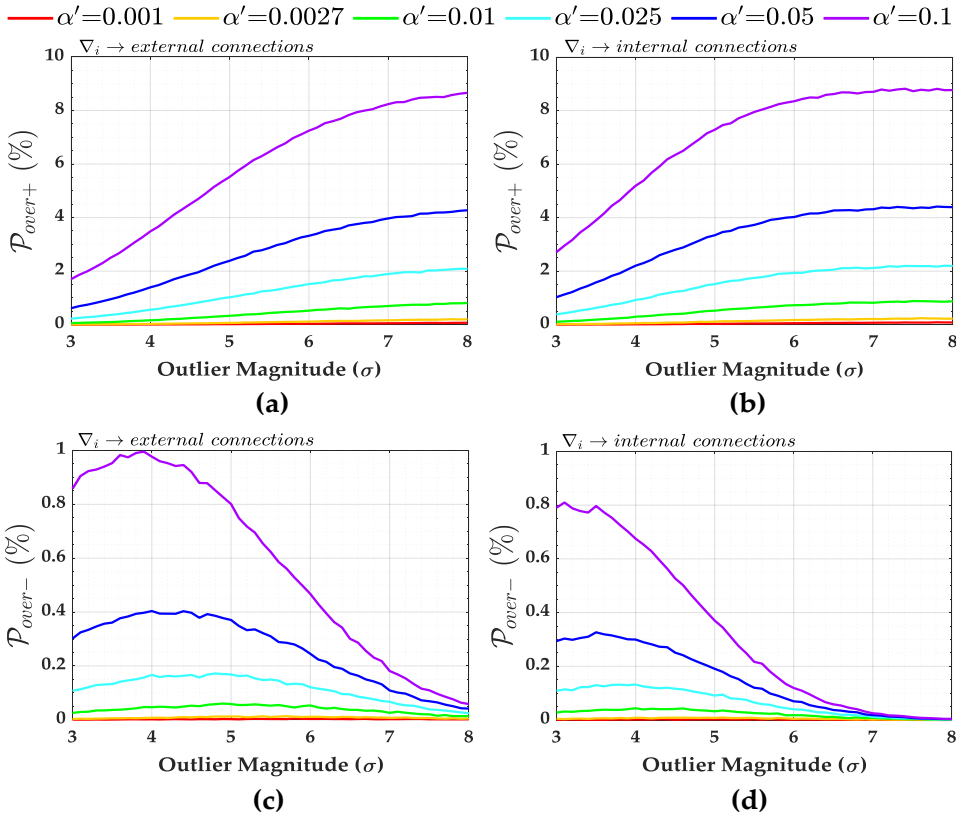

Figure 14. Over-identification cases for network (**a**). (**a**) $\mathcal{P}_{over+}$ for external connections. (**b**) $\mathcal{P}_{over+}$ for internal connections. (**c**) $\mathcal{P}_{over-}$ for external connections. (**d**) $\mathcal{P}_{over-}$ for internal connections.

6.2.2. Geodetic Network with High Correlation Between Residuals

Now, the correlation between residuals is very high. This is the case for network (**b**) detailed in Figure 2. Since the measurements are correlated for network (**b**), instead of redundancy numbers, reliability numbers $(\bar{r}_i)$ should be given as an internal reliability measure, as follows [43]:

$$\bar{r}_i = \boldsymbol{c}_i^T \boldsymbol{Q}_e \boldsymbol{c}_i \boldsymbol{c}_i^T \boldsymbol{W} \boldsymbol{Q}_{\hat{e}} \boldsymbol{W} \boldsymbol{c}_i, \forall i = 1, \cdots, n \tag{61}$$

The reliability numbers $(\bar{r}_i)$ in Equation (61) are equivalent to redundancy numbers when it is assumed that the measurements are uncorrelated. Table 7 gives the reliability numbers $(\bar{r}_i)$, the standard deviation of each measurement $\sigma_{\Delta h_{i-j}}$ and the standard deviation of each estimated outlier σ_{∇_i} for network (**b**).

Table 7. Reliability numbers $(\bar{r}_i)$, standard deviation $\sigma_{\Delta h_{i-j}}$ and standard deviation of estimated outlier σ_{∇_i} for network (**b**).

Δh_{i-j}	$\bar{r}_i$	$\sigma_{\Delta h_{i-j}}$ (m)	σ_{∇_i} (m)
Δh_1	10.58	2.35	0.72
Δh_2	0.62	1.97	2.50
Δh_3	0.13	0.89	2.50
Δh_4	13.68	2.32	0.63
Δh_5	1.95	0.45	0.32
Δh_6	3.56	1.18	0.63

The probabilities of correct identification $(\mathcal{P}_{CI})$ for this network are displayed in Figure 15. The critical values $(\hat{k})$ for network (**b**) are those given in Table 4. The probability levels of correct detection $(\mathcal{P}_{CD})$ are provided in Figure 16.

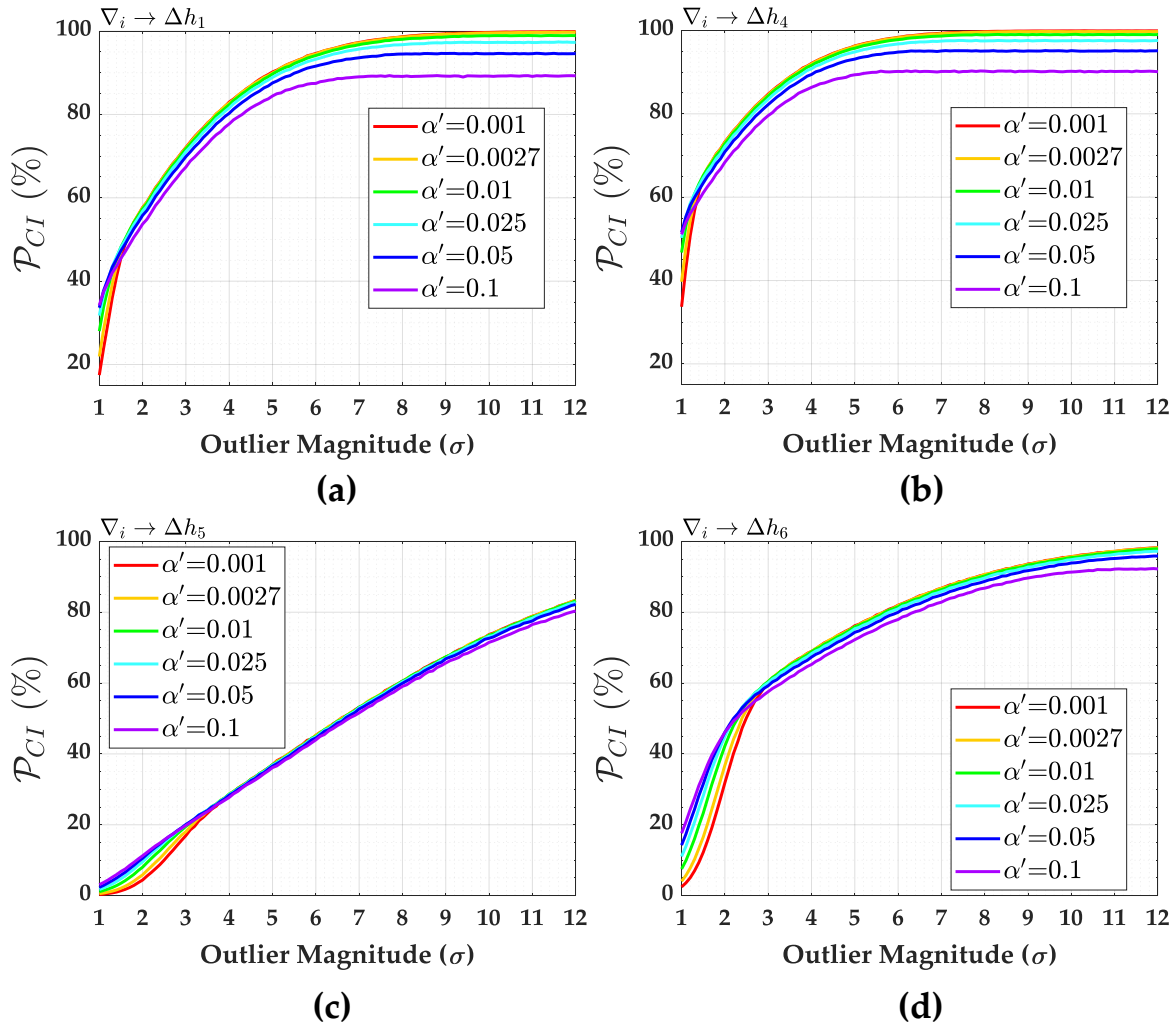

Figure 15. Probability of correct identification ($\mathcal{P}_{CI}$) for network (**b**). (**a**) $\mathcal{P}_{CI}$ for Δh_1. (**b**) $\mathcal{P}_{CI}$ for Δh_4. (**c**) $\mathcal{P}_{CI}$ for Δh_5. (**d**) $\mathcal{P}_{CI}$ for Δh_6.

In contrast to network (**a**), the probability of correct identification ($\mathcal{P}_{CI}$) for network (**b**) is different for each measurement. It is also found that the larger the Type I decision error α', the higher the probability of correct identification ($\mathcal{P}_{CI}$). However, it is only true up to a certain level of outlier magnitude. After this magnitude level, the larger the Type I decision error α', the lower the probability of correct identification ($\mathcal{P}_{CI}$).

The user-defined Type I error α' has indeed become less significant at a certain outlier magnitude. Note, for example, that the probability of correct identification for measurement Δh_1 for $\alpha' = 0.1$ is higher than that for $\alpha' = 0.001$ when the outlier magnitude is between 1σ and 1.5σ. For a magnitude greater than 1.5σ, we note that the larger the Type I decision error α', the lower the probability of correct identification $\mathcal{P}_{CI}$. The choice of Type I error α', however, has no significant effect on the probability of correct identification $\mathcal{P}_{CI}$ for an outlier magnitude greater than 1.5σ. This analysis can also be done with Δh_4, Δh_5 and Δh_6.

There is no probability of identification for both measurements Δh_2 and Δh_3. This is because the residual correlation of these measurements is equal to exactly one (i.e., $\rho_{w_i,w_j} = 1.00$). Furthermore, the reliability numbers ($\bar{r}_i$) in Table 7 for those measurements are close to zero. However, if one of those measurements were affected by a single outlier, then *IDS* would have the ability to detect it. In other words, there is reliability in terms of outlier detection for Δh_2 and Δh_3.

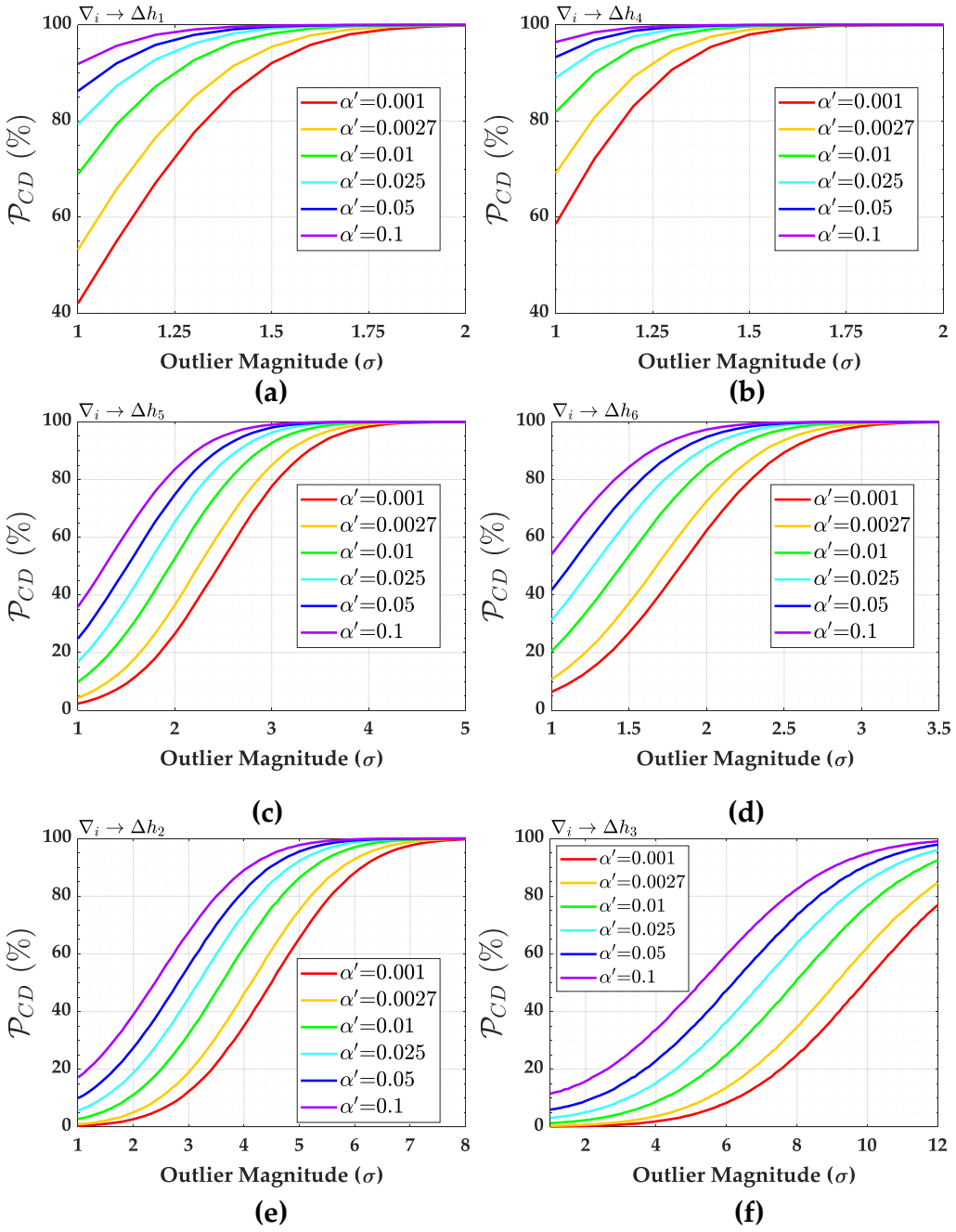

Figure 16. Probability of correct detection ($\mathcal{P}_{CD}$) for network (**b**). (**a**) $\mathcal{P}_{CD}$ for Δh_1. (**b**) $\mathcal{P}_{CD}$ for Δh_4. (**c**) $\mathcal{P}_{CD}$ for Δh_5. (**d**) $\mathcal{P}_{CD}$ for Δh_6. (**e**) $\mathcal{P}_{CD}$ for Δh_2. (**f**) $\mathcal{P}_{CD}$ for Δh_3.

We observe that the higher the reliability numbers in Table 3, the higher the power of detection $\mathcal{P}_{CD}$ and identification $\mathcal{P}_{CI}$. In general, the larger the Type I decision error α', the lower the probability of missed detection $\mathcal{P}_{MD}$ and, therefore, the higher the probability of correct detection $\mathcal{P}_{CD}$.

The sensitivity indicators MDB and MIB for Δh_1, Δh_4, Δh_5 and Δh_6 are shown in Tables 8–11, respectively. Both MIBs and MDBs were computed for each α' and for a success rate of 0.8 (80%) for both outlier detection and identification, i.e., $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 80\%$. The non-centrality parameters for outlier detection and identification were computed according to Equations (45) and (46), respectively. In general, the larger the Type I decision error α', the larger the MIB and the smaller the MDB. In other words, the larger the Type I decision error α', the greater the chances of outlier detection but the lower

the chances of outlier identification. In that case, the larger the Type I Error α', the larger the MIB/MDB ratio. Therefore, an outlier with a size of the MDB should be enlarged in order to identify it [1,2,12,46].

Table 8. Relationship between the MDB and MIB for Δh_1 and for $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 80\%$.

α'	MIB	MDB	$\lambda_{q=1}^{(MIB)}$	$\lambda_{q=1}^{(MDB)}$	MIB/MDB
0.001	3.700σ	1.327σ	145.839	18.759	2.788
0.0027	3.700σ	1.240σ	145.839	16.380	2.984
0.010	3.750σ	1.109σ	149.807	13.102	3.381
0.025	3.840σ	1.009σ	157.084	10.846	3.806
0.050	3.980σ	0.930σ	168.747	9.214	4.280
0.100	4.320σ	0.830σ	198.810	7.339	5.205

Table 9. Relationship between the MDB and MIB for Δh_4 and for $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 80\%$.

α'	MIB	MDB	$\lambda_{q=1}^{(MIB)}$	$\lambda_{q=1}^{(MDB)}$	MIB/MDB
0.001	2.558σ	1.170σ	88.735	18.564	2.186
0.0027	2.566σ	1.093σ	89.291	16.201	2.348
0.010	2.598σ	0.982σ	91.532	13.077	2.646
0.025	2.659σ	0.895σ	95.902	10.863	2.971
0.050	2.784σ	0.820σ	105.107	9.118	3.395
0.100	3.082σ	0.738σ	128.771	7.390	4.174

Table 10. Relationship between the MDB and MIB for Δh_5 and for $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 80\%$.

α'	MIB	MDB	$\lambda_{q=1}^{(MIB)}$	$\lambda_{q=1}^{(MDB)}$	MIB/MDB
0.001	11.290σ	3.065σ	252.065	18.577	3.684
0.0027	11.260σ	2.863σ	250.727	16.209	3.933
0.010	11.315σ	2.565σ	253.183	13.011	4.411
0.025	11.360σ	2.328σ	255.201	10.717	4.880
0.050	11.530σ	2.127σ	262.896	8.947	5.421
0.100	11.940σ	1.906σ	281.925	7.184	6.264

Table 11. Relationship between the MDB and MIB for Δh_6 and for $\tilde{\mathcal{P}}_{CD} = \tilde{\mathcal{P}}_{CI} = 80\%$.

α'	MIB	MDB	$\lambda_{q=1}^{(MIB)}$	$\lambda_{q=1}^{(MDB)}$	MIB/MDB
0.001	5.680σ	2.289σ	113.183	18.375	2.482
0.0027	5.700σ	2.134σ	113.981	15.976	2.671
0.010	5.695σ	1.908σ	113.781	12.769	2.985
0.025	5.825σ	1.729σ	119.035	10.492	3.368
0.050	6.021σ	1.579σ	127.180	8.747	3.813
0.100	6.394σ	1.409σ	143.426	6.965	4.538

The overall probabilities of wrong exclusion ($\mathcal{P}_{WE}$) for network (**b**) are provided in Figure 17. In general, we observe that the wrong exclusion rate ($\mathcal{P}_{WE}$) increases up to a certain outlier magnitude and, from this point on, the wrong exclusion rate ($\mathcal{P}_{WE}$) starts to decline, and the effect of the user-defined Type 1 decision error (α') on $\mathcal{P}_{WE}$ becomes neutral in practical terms. This effect is due to the residuals' correlation. To see this effect more clearly, we also computed the individual contribution of each measurement to the overall wrong exclusion $\mathcal{P}_{WE}$ and their corresponding weighting factors given by Equations (40) and (41), respectively.

The individual contributions to the overall $\mathcal{P}_{WE}$ and their weighting factors for $\alpha' = 0.1$ are displayed in Figures 18 and 19, respectively. It is important to mention that the behaviour shown in Figures 18 and 19 is similar to that for other α' values. We observe that the correlation coefficient (ρ_{w_i,w_j}) only has a direct relationship with $\mathcal{P}_{WE}$ for a certain outlier magnitude. Let us consider the

case in which Δh_6 is set up as an outlier. In that case, the larger the correlation coefficient (ρ_{w_i,w_j}), the higher the individual contribution to $\mathcal{P}_{WE}$. Of course, this only holds true if the outlier magnitude is larger than 3.2σ. This is also evident from the results of the weighting factors in Figure 19.

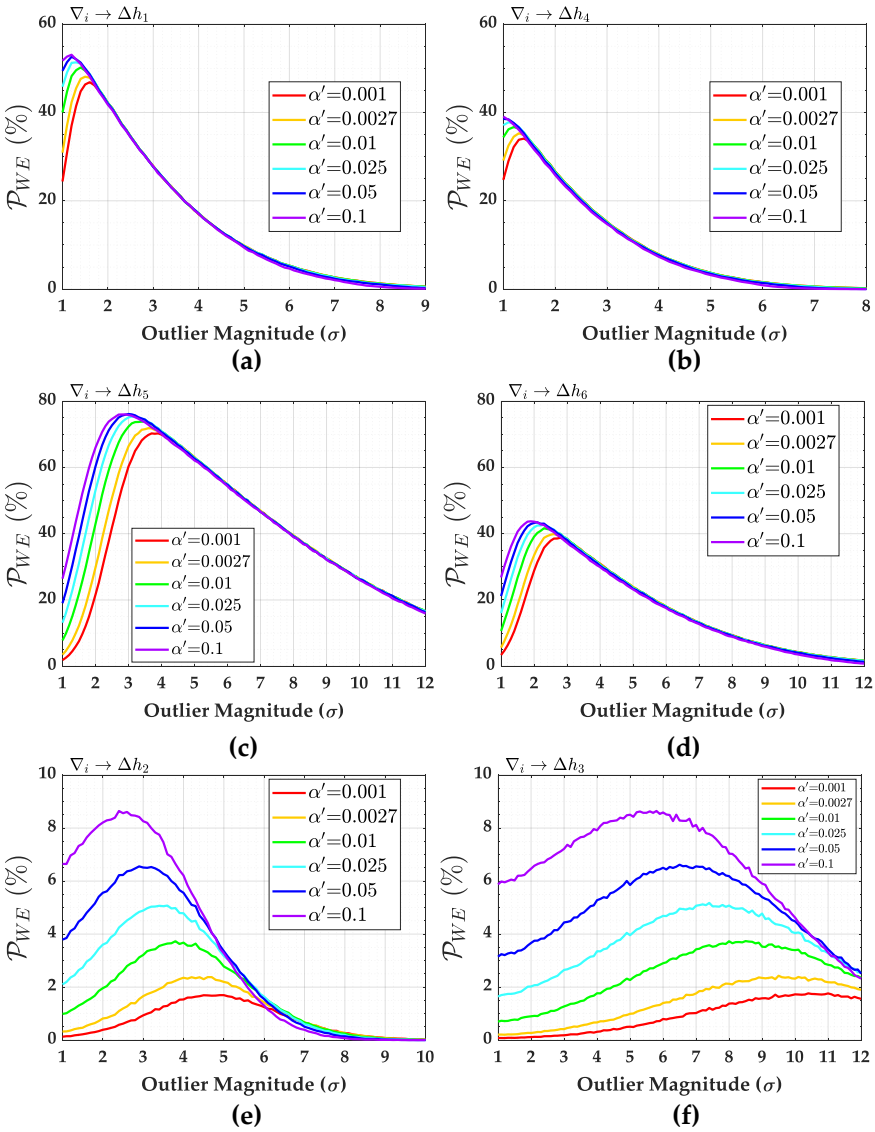

Figure 17. Probability of wrong exclusion ($\mathcal{P}_{WE}$) for network (**b**). (**a**) $\mathcal{P}_{WE}$ for Δh_1. (**b**) $\mathcal{P}_{WE}$ for Δh_4. (**c**) $\mathcal{P}_{WE}$ for Δh_5. (**d**) $\mathcal{P}_{WE}$ for Δh_6. (**e**) $\mathcal{P}_{WE}$ for Δh_2. (**f**) $\mathcal{P}_{WE}$ for Δh_3.

An important highlight is the association between the MIB and the contribution of each measurement to the probability of wrong exclusion $\mathcal{P}_{WE}$ in Figure 18.

We observe that it is possible to find the value of the MIB at high success rates when the individual contributions to the overall wrong exclusion $\mathcal{P}_{WE}$ of a given outlier start to decrease simultaneously. It is important to mention that this simultaneous decay occurs when there is a direct relationship between the correlation coefficient (ρ_{w_i,w_j}) and the wrong exclusion rate $\mathcal{P}_{WE}$. In that case, the identifiability of a given outlier can be verified for a given significance level α' and probability of correct identification $\mathcal{P}_{CI}$.

Figure 20 illustrates an example for measurements Δh_1 and Δh_4. The black dashed line corresponds to the probability of correct identification $\mathcal{P}_{CI}$ and the respective MIB for $\alpha' = 0.001$. Note that when the effect of all measurements on $\mathcal{P}_{WE}$ decreases, it is possible to find an outlier

magnitude that can be identified. In other words, the effect of the correlation between residuals (ρ_{w_i,w_j}) becomes insignificant at a certain outlier magnitude, which increases the probability of identification.

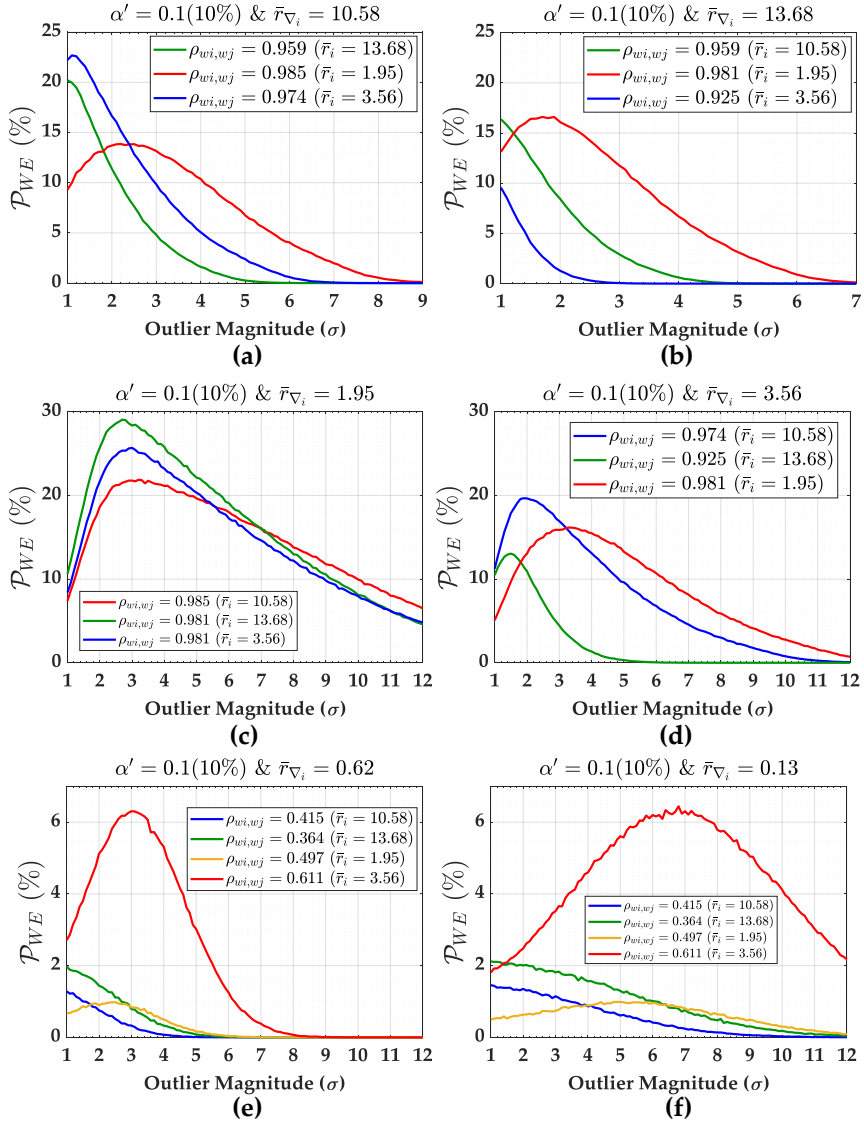

Figure 18. Individual contribution of each measurement to the overall wrong exclusion probability ($\mathcal{P}_{WE}$) for network (**b**) and for $\alpha' = 0.1$. (**a**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_1. (**b**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_4. (**c**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_5. (**d**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_6. (**e**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_2. (**f**) Individual contribution to $\mathcal{P}_{WE}$ for Δh_3.

The probabilities of wrong exclusion for both Δh_2 and Δh_3 are smaller than those for the other cases. This is because of the correlation between residuals (ρ_{w_i,w_j}). In fact, we also note that although there is no reliability in terms of outlier identification for cases in which the correlation is $\rho_{w_i,w_j} = 1.00$ (i.e., 100%), there is reliability for outlier detection. In this case, outlier detection is caused by overlapping w-test statistics. The result for statistical overlap ($\mathcal{P}_{ol}$) is displayed in Figure 21. In general, the larger the Type 1 decision error α', the larger the statistical overlap ($\mathcal{P}_{ol}$).

The over-identification cases ($\mathcal{P}_{over+}$ and $\mathcal{P}_{over-}$) are displayed in Figures 22 and 23, respectively. We observe that the larger the Type I decision error (α'), the larger the over-identification cases. It should be noted that $\mathcal{P}_{over+}$ is always larger than $\mathcal{P}_{over-}$. Over-identification $\mathcal{P}_{over-}$ is practically null. The over-identification cases $\mathcal{P}_{over+}$ for Δh_2, Δh_3 and Δh_6 and $\mathcal{P}_{over-}$ for Δh_1 are exactly null. The over-identifications $\mathcal{P}_{over-}$ for Δh_2 and Δh_3 are less than 0.2% and are therefore not shown here.

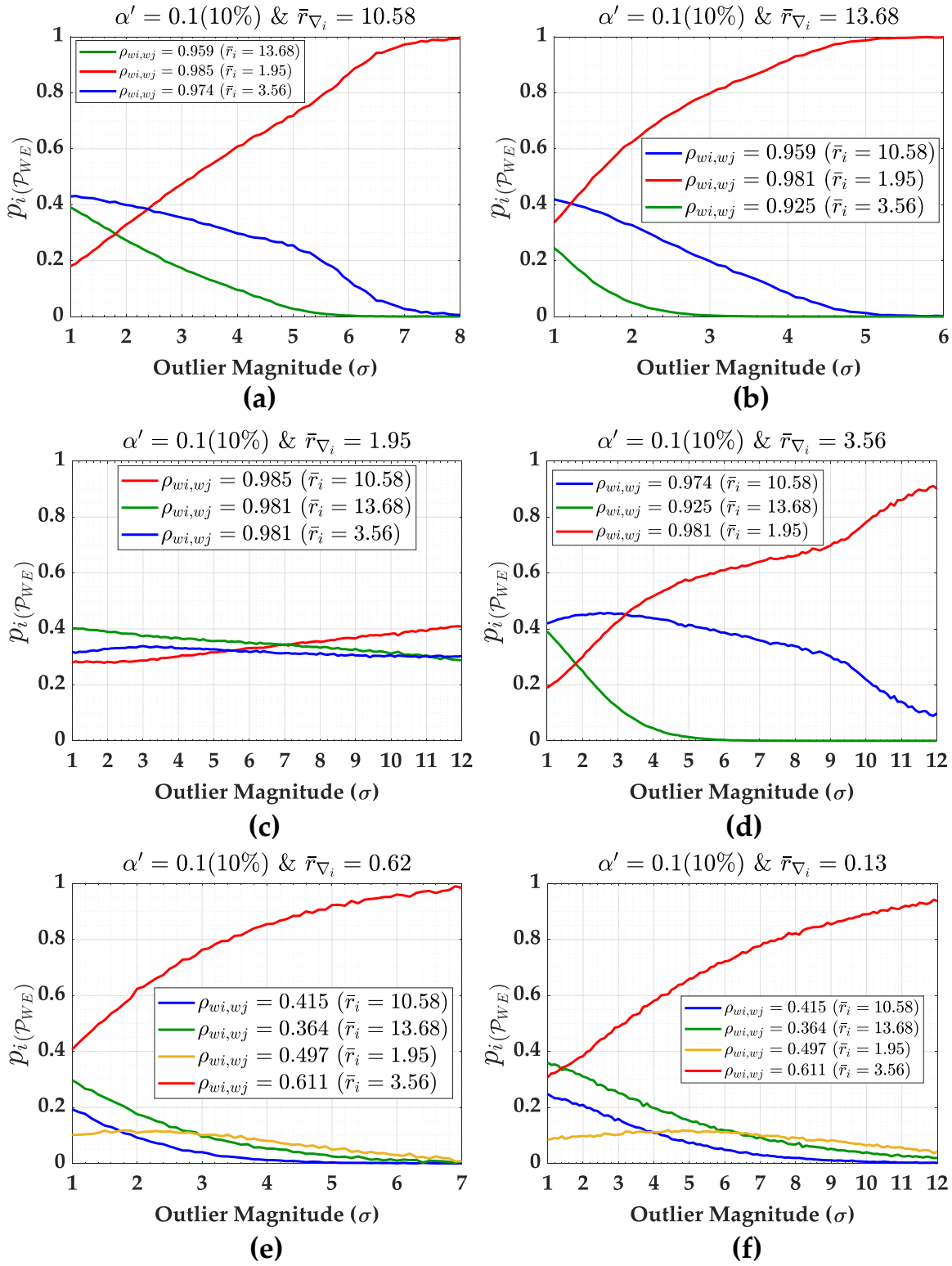

Figure 19. Weighting factors ($p_{i(\mathcal{P}_{WE})}$) of each measurement's contribution to the overall wrong exclusion probability ($\mathcal{P}_{WE}$) for network (**b**) and for $\alpha' = 0.1$. (**a**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_1. (**b**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_4. (**c**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_5. (**d**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_6. (**e**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_2. (**f**) Weighting factors of contributions to $\mathcal{P}_{WE}$ for Δh_3.

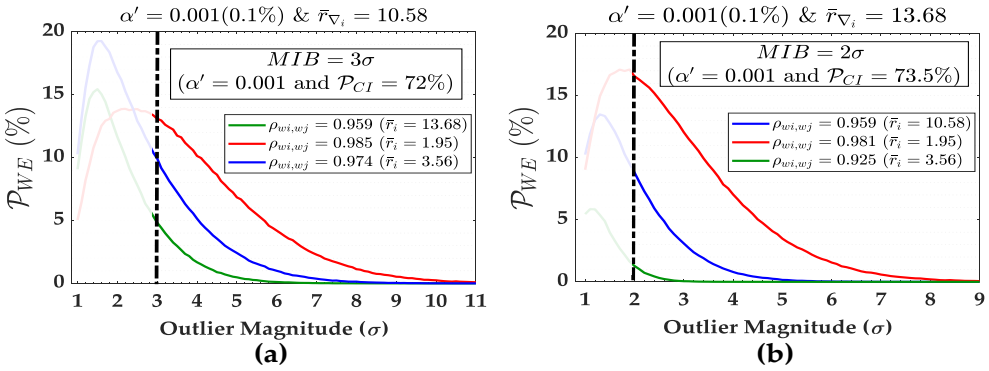

Figure 20. Relationship of the individual contributions to the overall probability of wrong exclusion ($\mathcal{P}_{WE}$) with correct identification rate $\mathcal{P}_{CI}$ and MIB for $\alpha' = 0.001$. (**a**) Example for measurement Δh_1. (**b**) Example for measurement Δh_4.

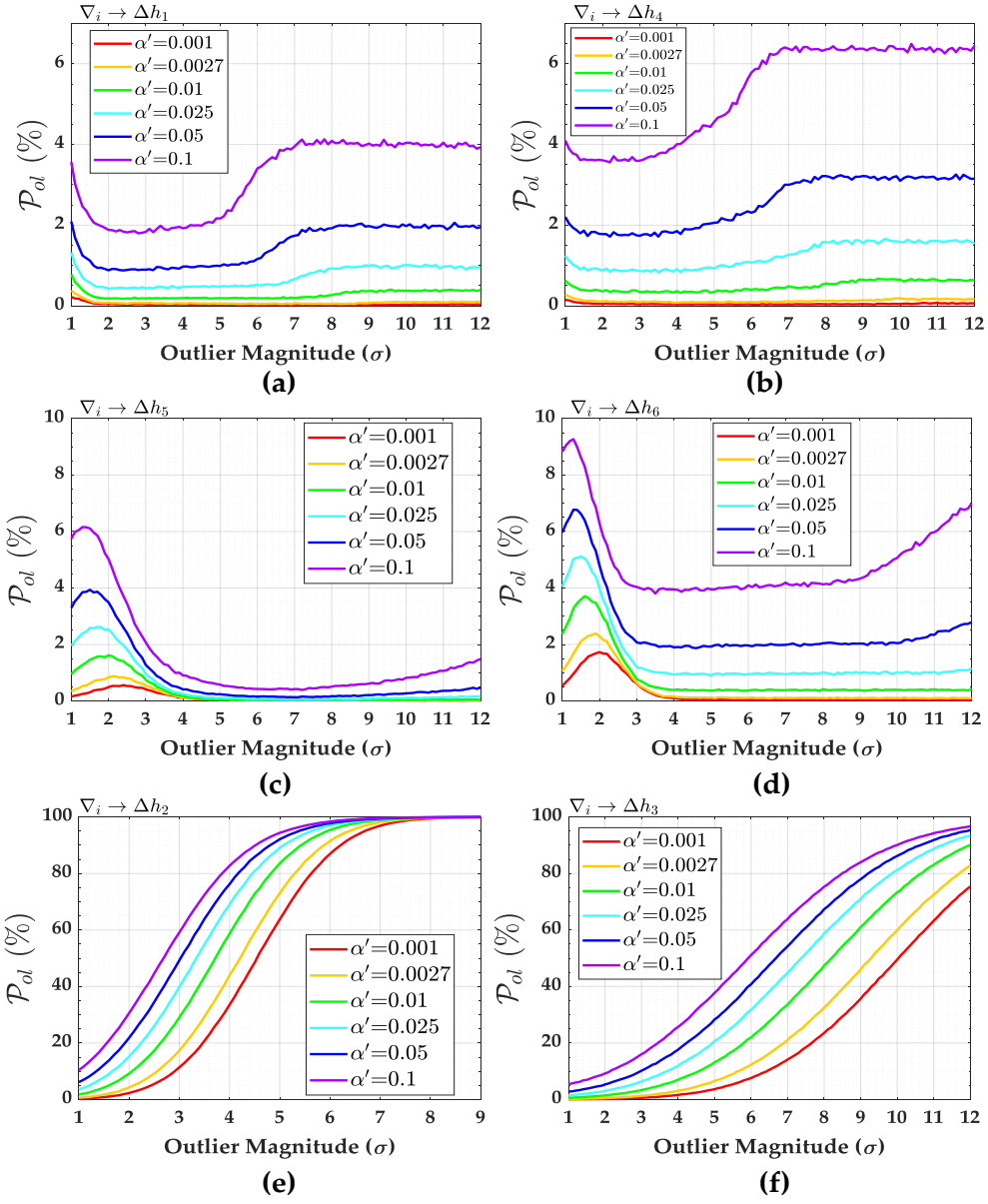

Figure 21. Probability of statistical overlap ($\mathcal{P}_{ol}$) for network (**b**). (**a**) $\mathcal{P}_{ol}$ for Δh_1. (**b**) $\mathcal{P}_{ol}$ for Δh_4. (**c**) $\mathcal{P}_{ol}$ for Δh_5. (**d**) $\mathcal{P}_{ol}$ for Δh_6. (**e**) $\mathcal{P}_{ol}$ for Δh_2. (**f**) $\mathcal{P}_{ol}$ for Δh_3.

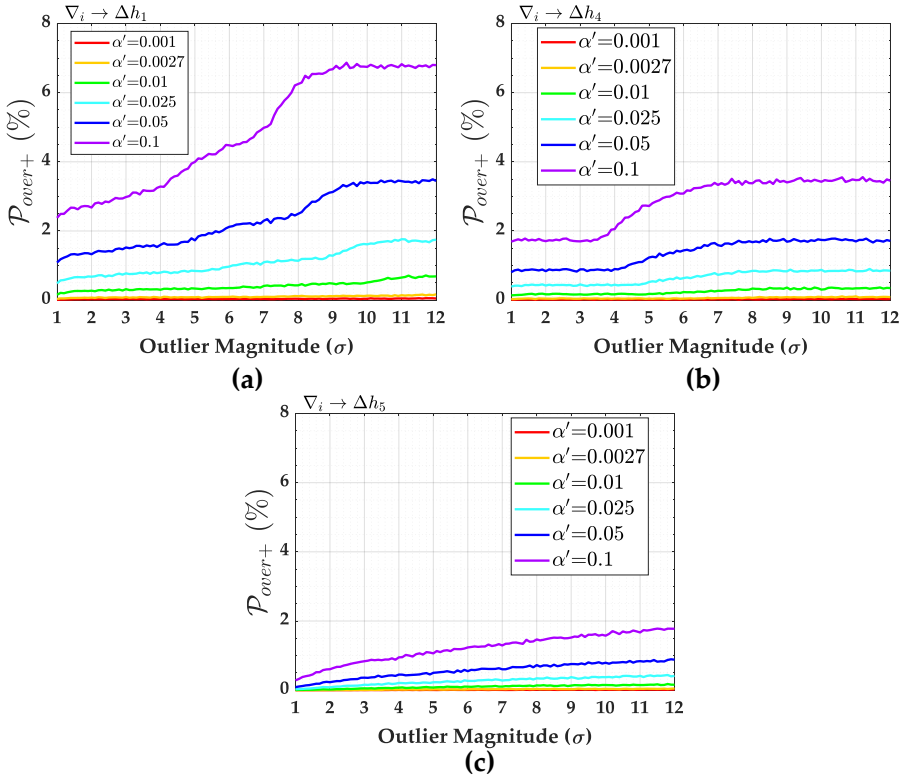

Figure 22. Probability of over-identification ($\mathcal{P}_{over+}$) for network (**b**). (**a**) $\mathcal{P}_{over+}$ for Δh_1. (**b**) $\mathcal{P}_{over+}$ for Δh_4. (**c**) $\mathcal{P}_{over+}$ for Δh_5.

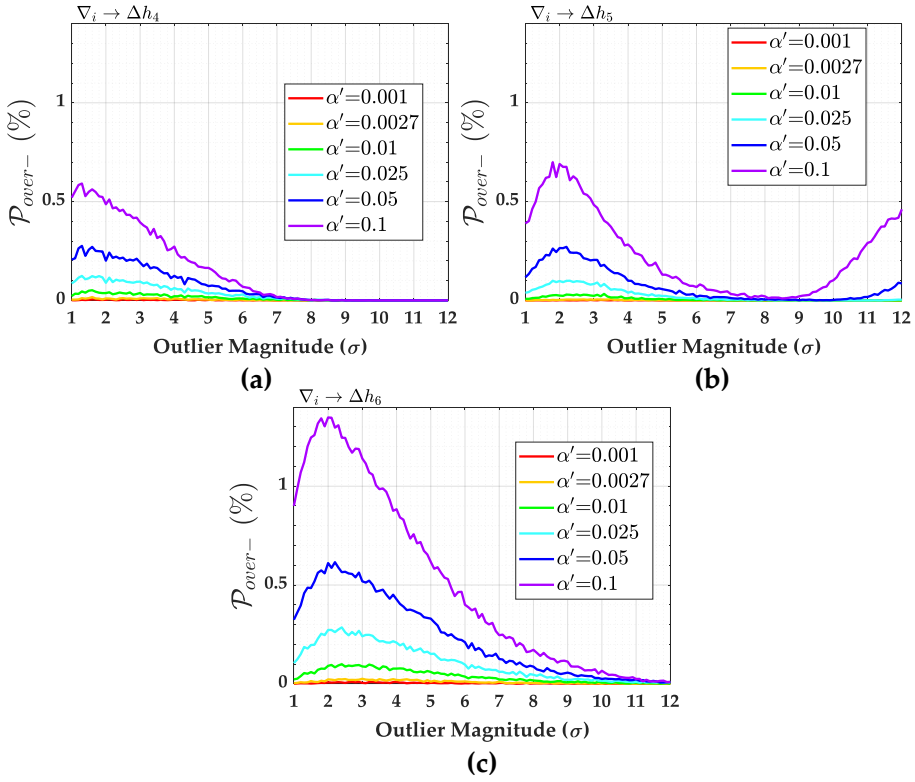

Figure 23. Probability of over-identification ($\mathcal{P}_{over-}$) for network (**b**). (**a**) $\mathcal{P}_{over-}$ for Δh_4. (**b**) $\mathcal{P}_{over-}$ for Δh_5. (**c**) $\mathcal{P}_{over-}$ for Δh_6.

7. Conclusions and Outlook

In this paper, we propose a procedure to compute the probability levels associated with an iterative outlier elimination procedure. This iterative outlier elimination procedure is known among geodesists as iterative data snooping *IDS*. On the basis of the probability levels of *IDS*, the sensitivity indicators—the Minimal Detectable Bias (MDB) and Minimal Identifiable Bias (MIB)—can also be determined for a given measurement system.

We emphasize that the probability levels associated with *IDS* in the presence of an outlier were analysed as a function of the user-defined Type I decision error (α'), outlier magnitude (∇_i), correlation between test statistics (ρ_{w_i,w_j}) and reliability indicators (i.e., redundancy number r_i and reliability number $\bar{r}_i$). It is important to highlight that these probability levels are based on critical values that were optimized via Monte Carlo.

We highlight the main findings of the paper below:

1. If one adopts the Bonferroni correction to compute the critical value of the test statistic associated with *IDS*, one does not have control over Type I decision errors. This is only true for small α' values and for a measurement system with high redundancy and a low correlation between test statistics.
2. If one maintains the condition of a measurement system with a low correlation between test statistics, the probability of wrong exclusion $\mathcal{P}_{WE}$ is too low. In that case, one should opt for a larger α' so that the probability of missed detection $\mathcal{P}_{MD}$ is as small as possible. Thus, it is possible to guarantee a high outlier identification rate. However, we verify that, under certain circumstances, the larger the Type I decision error α', the higher the probability of correct detection $\mathcal{P}_{CD}$ but the lower the probability of correct identification $\mathcal{P}_{CI}$. In that case, the larger the Type I Error α', the larger the ratio between the sensitivity indicators MIB/MDB.
3. The larger the Type I error (α'), the higher the probability of correct outlier identification ($\mathcal{P}_{CI}$). However, it is valid only to a certain limit of outlier magnitude (threshold). There is an inversion when the outlier magnitude is greater than this threshold: i.e., the larger the α', the lower the $\mathcal{P}_{CI}$. This is more critical in the case of a measurement system with a high correlation between test statistics. Moreover, the Type I decision error α' restricts the maximum rate of $\mathcal{P}_{CI}$.
4. We also observe that it is possible to find the value of the MIB when the contributions of each measurement to the probability of wrong exclusion $\mathcal{P}_{WE}$ start to decline simultaneously. In that case, the identifiability of a given outlier can be verified for a given α' and $\mathcal{P}_{CI}$. In other words, for a certain outlier magnitude, the effect of the correlation between test statistics becomes insignificant, which increases the probability of identification. Moreover, if a small outlier magnitude (outlier with a magnitude close to the measurement uncertainty) were to arise for a measurement system with a high correlation between test statistics, the alternative hypotheses would not be distinguished; i.e., this outlier would never be identified.
5. The larger the Type I decision error α', the larger the over-identification cases. The over-identification case $\mathcal{P}_{over+}$ is always larger than $\mathcal{P}_{over-}$. We also note that the lower the correlation between test statistics, the higher the probability of over-identification positive $\mathcal{P}_{over+}$. For small α' (close to $\alpha' = 0.001$), $\mathcal{P}_{over-}$ is practically null.
6. When the correlation between two test statistics is equal to exactly 1.00, $\mathcal{P}_{CI}$ does not exist, but there is $\mathcal{P}_{CD}$, which is mainly caused by $\mathcal{P}_{ol}$.

The computation procedure presented in this paper was successfully applied to a practical example of geodetic networks. Although the procedure was applied to geodetic networks, it is a generally applicable method. The authors have been working on solutions to find a relationship between the variables computed deterministically (e.g., local redundancy, residuals' correlation) with the probability levels computed by Monte Carlo. The use of Monte Carlo will no longer be needed to find the MIB if a model is found. Moreover, further investigation is required to apply this analysis to general problems with multiple outliers.

Author Contributions: Conceptualization, V.F.R., M.T.M. and I.K.; methodology, V.F.R.; software, V.F.R. And L.G.d.S.J.; validation, V.F.R., M.T.M. and I.K.; formal analysis, V.F.R., M.T.M. and I.K.; investigation, V.F.R.; data curation, V.F.R., I.K. and M.T.M.; writing—original draft preparation, V.F.R.; writing—review and editing, V.F.R., M.T.M., I.K. and M.R.V.; supervision, M.T.M., I.K., M.R.V. and L.G.d.S.J.; project administration, M.T.M. and I.K.; funding acquisition, M.T.M., M.R.V. and L.G.d.S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the CNPq—Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (proc. nº 103587/2019-5). This research and the APC were also funded by PETROBRAS (Grant Number 2018/00545-0).

Acknowledgments: The authors would like to thank the three anonymous reviewers who contributed to the improvement of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, J.; Knight, N.L. New Outlier Separability Test and Its Application in GNSS Positioning. *J. Glob. Position. Syst.* **2012**, *11*, 46–57. [CrossRef]
2. Yang, L.; Wang, J.; Knight, N.L.; Shen, Y. Outlier separability analysis with a multiple alternative hypotheses test. *J. Geod.* **2013**, *87*, 591–604. [CrossRef]
3. Teunissen, P.J.G.; Imparato, D.; Tiberius, C.C.J.M. Does RAIM with Correct Exclusion Produce Unbiased Positions? *Sensors* **2017**, *17*, 1508. [CrossRef] [PubMed]
4. Na, W.; Park, C.; Lee, S.; Yu, S.; Lee, H. Sensitivity-Based Fault Detection and Isolation Algorithm for Road Vehicle Chassis Sensors. *Sensors* **2018**, *18*, 2720. [CrossRef] [PubMed]
5. Crispoltoni, M.; Fravolini, M.L.; Balzano, F.; D'Urso, S.; Napolitano, M.R. Interval Fuzzy Model for Robust Aircraft IMU Sensors Fault Detection. *Sensors* **2018**, *18*, 2488. [CrossRef]
6. Nguyen, V.K.; Renault, E.; Milocco, R. Environment Monitoring for Anomaly Detection System Using Smartphones. *Sensors* **2019**, *19*, 3834. [CrossRef]
7. Mei, X.; Wu, H.; Xian, J.; Chen, B.; Zhang, H.; Liu, X. A Robust, Non-Cooperative Localization Algorithm in the Presence of Outlier Measurements in Ocean Sensor Networks. *Sensors* **2019**, *19*, 2708. [CrossRef]
8. Nie, Y.; Yang, L.; Shen, Y. Specific Direction-Based Outlier Detection Approach for GNSS Vector Networks. *Sensors* **2019**, *19*, 1836. [CrossRef]
9. Leslar, M.; Wang, J.G.; Hu, B. Comprehensive Utilization of Temporal and Spatial Domain Outlier Detection Methods for Mobile Terrestrial LiDAR Data. *Remote. Sens.* **2011**, *3*, 1724–1742. [CrossRef]
10. Rofatto, V.F.; Matsuoka, M.T.; Klein, I.; Veronez, M.R. Monte-Carlo-based uncertainty propagation in the context of Gauss–Markov model: a case study in coordinate transformation. *Sci. Plena* **2019**, *15*, 1–17. [CrossRef]
11. Lehmann, R. Observation error model selection by information criteria vs. normality testing. *Stud. Geophys. Geod.* **2015**, *59*, 489–504. [CrossRef]
12. Rofatto, V.F.; Matsuoka, M.T.; Klein, I.; Veronez, M.R.; Bonimani, M.L.; Lehmann, R. A half-century of Baarda's concept of reliability: A review, new perspectives, and applications. *Surv. Rev.* **2018**, 1–17. [CrossRef]
13. Lehmann, R. On the formulation of the alternative hypothesis for geodetic outlier detection. *J. Geod.* **2013**, *87*, 373–386. [CrossRef]
14. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*, 1st ed.; Wiley-Interscience: Hoboken, NJ, USA, 2003.
15. Yang, Y. Robust estimation of geodetic datum transformation. *J. Geod.* **1999**, *73*, 268–274. [CrossRef]
16. Wilcox, R. *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2013. [CrossRef]
17. Duchnowski, R. Hodges–Lehmann estimates in deformation analyses. *J. Geod.* **2013**, *87*, 873–884. [CrossRef]
18. Klein, I.; Matsuoka, M.T.; Guzatto, M.P.; de Souza, S.F.; Veronez, M.R. On evaluation of different methods for quality control of correlated observations. *Surv. Rev.* **2015**, *47*, 28–35. [CrossRef]
19. Baarda, W. Statistical Concepts in Geodesy. *Publ. Geod. Neth. Geod. Comm.* **1967**, *2*, 1–74.
20. Baarda, W. A Testing Procedure for Use in Geodetic Networks. *Publ. Geod. Neth. Geod. Comm.* **1968**, *2*, 1–97.

21. Förstner, W. Reliability and discernability of extended Gauss-Markov models. In *Seminar on Mathematical Models to Outliers and Systematic Errors*; No. 98; German Geodetic Commission (DGK): Munich, Germany, 1983; Volume A, pp. 79–103.
22. Lehmann, R. Improved critical values for extreme normalized and studentized residuals in Gauss–Markov models. *J. Geod.* **2012**, *86*, 1137–1146. [CrossRef]
23. Prószyński, W. Revisiting Baarda's concept of minimal detectable bias with regard to outlier identifiability. *J. Geod.* **2015**, *89*, 993–1003. [CrossRef]
24. Marshall, J. L1-norm pre-analysis measures for geodetic networks. *J. Geod.* **2002**, *76*, 334–344. [CrossRef]
25. Huber, P.J. Robust Statistics. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1248–1251. [CrossRef]
26. WiÅniewski, Z. Estimation of parameters in a split functional model of geodetic observations (M split estimation). *J. Geod.* **2008**, *82*. [CrossRef]
27. WiÅniewski, Z.; Duchnowski, R.; Dumalski, A. Efficacy of Msplit Estimation in Displacement Analysis. *Sensors* **2019**, *19*, 47. [CrossRef] [PubMed]
28. Hodges, J.L.; Lehmann, E.L. Estimates of Location Based on Rank Tests. *Ann. Math. Stat.* **1963**, *34*, 598–611. [CrossRef]
29. Duchnowski, R. Robustness of Strategy for Testing Levelling Mark Stability Based on Rank Tests. *Surv. Rev.* **2011**, *43*, 687–699. [CrossRef]
30. Wyszkowska, P.; Duchnowski, R. Subjective breakdown points of R-estimators applied in deformation analysis. In Proceedings of the International Conference on Environmental Engineering, Vilnius, Lithuania, 27–28 April 2017; pp. 1–6. [CrossRef]
31. Koch, I.É.; Klein, I.; Gonzaga, L.; Matsuoka, M.T.; Rofatto, V.F.; Veronez, M.R. Robust Estimators in Geodetic Networks Based on a New Metaheuristic: Independent Vortices Search. *Sensors* **2019**, *19*, 4535. [CrossRef]
32. Lehmann, R. 3σ-Rule for Outlier Detection from the Viewpoint of Geodetic Adjustment. *J. Surv. Eng.* **2013**, *139*, 157–165. [CrossRef]
33. Lehmann, R.; Scheffler, T. Monte Carlo based data snooping with application to a geodetic network. *J. Appl. Geod.* **2011**, *5*, 123–134. [CrossRef]
34. Koch, K.R. *Parameter Estimation and Hypothesis Testing in Linear Models*, 2nd ed.; Springer: Berlin, Germany, 1999.
35. Teunissen, P. *Testing Theory: An Introduction*, 2nd ed.; Delft University Press: Delft, The Netherlands, 2006.
36. Ghilani, C.D. *Adjustment Computations: Spatial Data Analysis*, 6th ed.; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2017.
37. Zaminpardaz, S.; Teunissen, P. DIA-datasnooping and identifiability. *J. Geod.* **2019**, *93*, 85–101. [CrossRef]
38. Zhao, Y.; Sun, R.; Ni, Z. Identification of Natural and Anthropogenic Drivers of Vegetation Change in the Beijing-Tianjin-Hebei Megacity Region. *Remote Sens.* **2019**, *11*, 1224. [CrossRef]
39. Wang, K.N.; Ao, C.O.; Juárez, M.D. GNSS-RO Refractivity Bias Correction Under Ducting Layer Using Surface-Reflection Signal. *Remote Sens.* **2020**, *12*, 359. [CrossRef]
40. Lee, G. An Efficient Compressive Hyperspectral Imaging Algorithm Based on Sequential Computations of Alternating Least Squares. *Remote Sens.* **2019**, *11*, 2932. [CrossRef]
41. Zhang, Y.; Wang, X.; Balzter, H.; Qiu, B.; Cheng, J. Directional and Zonal Analysis of Urban Thermal Environmental Change in Fuzhou as an Indicator of Urban Landscape Transformation. *Remote Sens.* **2019**, *11*, 2810. [CrossRef]
42. Kok, J.J.; States, U. *On Data Snooping and Multiple Outlier Testing*, NOAA Technical Report NOS, NGS, 30; U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, Charting and Geodetic Services: Rockville, MD, USA, 1984.
43. Knight, N.L.; Wang, J.; Rizos, C. Generalised measures of reliability for multiple outliers. *J. Geod.* **2010**, *84*, 625–635. [CrossRef]
44. Gui, Q.; Li, X.; Gong, Y.; Li, B.; Li, G. A Bayesian unmasking method for locating multiple gross errors based on posterior probabilities of classification variables. *J. Geod.* **2011**, *85*, 191–203. [CrossRef]
45. Klein, I.; Matsuoka, M.T.; Guzatto, M.P.; Nievinski, F.G. An approach to identify multiple outliers based on sequential likelihood ratio tests. *Surv. Rev.* **2017**, *49*, 449–457. [CrossRef]
46. Imparato, D.; Teunissen, P.; Tiberius, C. Minimal Detectable and Identifiable Biases for quality control. *Surv. Rev.* **2019**, *51*, 289–299. [CrossRef]

47. Teunissen, P.J.G. Distributional theory for the DIA method. *J. Geod.* **2018**, *92*, 59–80. [CrossRef]
48. Hekimoglu, S.; Koch, K.R. How can reliability of the robust methods be measured? In *Third Turkish-German Joint Geodetic Days: Towards a Digital Age*; Altan, M.O., Gründig, L., Eds.; Istanbul Technical University: Istanbul, Turkey, 1999; Volume 1, pp. 179–196.
49. Aydin, C. Power of Global Test in Deformation Analysis. *J. Surv. Eng.* **2012**, *138*, 51–56. [CrossRef]
50. Nowel, K. Application of Monte Carlo method to statistical testing in deformation analysis based on robust M-estimation. *Surv. Rev.* **2016**, *48*, 212–223. [CrossRef]
51. Klein, I.; Matsuoka, M.T.; Guzatto, M.P.; Nievinski, F.G.; Veronez, M.R.; Rofatto, V.F. A new relationship between the quality criteria for geodetic networks. *J. Geod.* **2019**, *93*, 529–544. [CrossRef]
52. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*, 2 ed.; Springer: New York, NY, USA, 2004.
53. Gamerman, D.; Lopes, H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed.; Taylor & Francis: London, UK, 2006.
54. Koch, K. Bayesian statistics and Monte Carlo methods. *J. Geod. Sci.* **2018**, *8*, 18–29. [CrossRef]
55. Rofatto, V.; Matsuoka, M.; Klein, I. An Attempt to Analyse Baarda's Iterative Data Snooping Procedure based on Monte Carlo Simulation. *S. Afr. J. Geomat.* **2017**, *6*, 416–435. [CrossRef]
56. Bonferroni, C. Teoria Statistica Delle Classi E Calcolo Delle Probabilità. *Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze* **1936**, *8*, 1–62.
57. Velsink, H. On the deformation analysis of point fields. *J. Geod.* **2015**, *89*, 1071–1087. [CrossRef]
58. Lehmann, R.; Lösler, M. Multiple Outlier Detection: Hypothesis Tests versus Model Selection by Information Criteria. *J. Surv. Eng.* **2016**, *142*, 04016017. [CrossRef]
59. Lehmann, R.; Lösler, M. Congruence analysis of geodetic networks—Hypothesis tests versus model selection by information criteria. *J. Appl. Geod.* **2017**, *11*, 271–283. [CrossRef]
60. Rofatto, V.; Matsuoka, M.; Klein, I. Design of geodetic networks based on outlier identification criteria: An example applied to the leveling network. *Bull. Geod. Sci.* **2018**, *24*, 152–170. [CrossRef]
61. Matsuoka, M.T.; Rofatto, V.F.; Klein, I.; Roberto Veronez, M.; da Silveira, L.G.; Neto, J.B.S.; Alves, A.C.R. Control Points Selection Based on Maximum External Reliability for Designing Geodetic Networks. *Appl. Sci.* **2020**, *10*, 687. [CrossRef]
62. Koch, K.R. Expectation Maximization algorithm and its minimal detectable outliers. *Stud. Geophys. Geod.* **2017**, *61*, 1–18. [CrossRef]
63. Arnold, S. *The Theory of Linear Models and Multivariate Analysis*, 1st ed.; Wiley: Hoboken, NJ, USA, 1981.
64. Teunissen, P.J.G. An Integrity and Quality Control Procedure for Use in Multi Sensor Integration In Proceedings of the 3rd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1990), Colorado Spring, CO, USA, 19–21 September 1990; pp. 513–522.
65. Aydin, C.; Demirel, H. Computation of Baarda's lower bound of the non-centrality parameter. *J. Geod.* **2004**, *78*, 437–441. [CrossRef]
66. Mierlo, J.V. Statistical Analysis of Geodetic Measurements for the Investigation of Crustal Movements. In *Recent Crustal Movements, 1977*; Whitten, C., Green, R., Meade, B., Eds.; Elsevier: Amsterdam, The Netherlands, 1979; Volume 13, pp. 457–467. [CrossRef]
67. Hawkins, D.M. *Identification of Outliers*, 1st ed.; Springer: Amsterdam, The Netherlands, 1980. [CrossRef]
68. Van der Marel, H.; Rösters, A.J.M. Statistical Testing and Quality Analysis in 3-D Networks (part II) Application to GPS. In *Global Positioning System: An Overview*; Bock, Y., Leppard, N., Eds.; Springer: New York, NY, USA, 1990; pp. 290–297.
69. Romano, J.P.; Wolf, M. Multiple Testing of One-Sided Hypotheses: Combining Bonferroni and the Bootstrap. In *Predictive Econometrics and Big Data*; Kreinovich, V., Sriboonchitta, S., Chakpitak, N., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 78–94.
70. Bonimani, M.; Rofatto, V.; Matsuoka, M.; Klein, I. Application of artificial random numbers and Monte Carlo method in the reliability analysis of geodetic networks. *Rev. Bras. Comp. Apl.* **2019**, *11*, 74–85. [CrossRef]
71. Altiok, T.; Melamed, B. *Simulation Modeling and Analysis with Arena*, 1st ed.; Academic Press: Cambridge, MA, USA, 2007.
72. Matsumoto, M.; Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8*, 3–30. [CrossRef]
73. Box, G.E.P.; Muller, M.E. A Note on the Generation of Random Normal Deviates. *Ann. Math. Stat.* **1958**, *29*, 610–611. [CrossRef]

74. Lemeshko, B.Y.; Lemeshko, S.B. Extending the Application of Grubbs-Type Tests in Rejecting Anomalous Measurements. *Meas. Tech.* **2005**, *48*, 536–547. [CrossRef]
75. Algarni, D.A.; Ali, A.E. Heighting and Distance Accuracy with Electronic Digital Levels. *J. King Saud Univ. Eng. Sci.* **1998**, *10*, 229–239. [CrossRef]
76. Gemin, A.R.; Matos, É.S.; Faggion, P.L. Application of calibration certificate of digital leveling systems in the monitoring of structures: A case study at the governador josã richa hydroelectric power plant-pr. *Boletim CiÃ GeodÃ* **2018**, *24*, 235–249. [CrossRef]
77. Takalo, M.; Rouhiainen, P. Development of a System Calibration Comparator for Digital Levels in Finland. *Nord. J. Surv. Real Estate Res.* **2004**, *1*, 119–130.