II Simpósio Regional de Agrimensura e Cartografia

"Ampliando os horizontes e discutindo o futuro da geoinformação e do cadastro territorial do Brasil"
Universidade Federal de Uberlândia – UFU / Campus Monte Carmelo
22 a 24 de novembro de 2021

A NEURAL-NETWORK-BASED CRITICAL VALUES FOR THE CASE OF MULTIPLE HYPOTHESIS TESTING

Maria Luísa Silva Bonimani [1]; Caique Eduardo de Jesus Nascimento Simionato [2]; Caio Cesar de Campos[3]; Vinicius Francisco Rofatto[4]; Ivandro Klein[5]; Marcelo Tomio Matsuoka[6]

ABSTRACT

Outlier is an observation that has moved away from most likely value to the point of not belonging to the mathematical model (functional and stochastic) stipulated. Failure to identify an outlier can jeopardise the reliability level of a system.Outliers must be appropriately treated to ensure the quality of data analysis. Data snooping outlier statistical testing procedure has been applied in Geodesy.The test procedure is liable to decision errors, such as Type I error (saying that a measure is an outlier when in fact it is not).It has been demonstrated that to effectively user-control the type I error rate, critical values must be computed numerically by means of Monte Carlo. We provide a model based on an artificial neural network. The results prove that the proposed model can be used to compute the critical values and, therefore, it is no longer necessary to run the Monte Carlo-based critical value every time the quality control is performed by means of data snooping.Details of that work can be found in Rofatto et al. (2021).

Keywords: Neural Network. Reliability. Monte Carlo. Quality Control. Hypothesis Testing.

1 INTRODUCTION

Data Snooping is one of the most best-established methods for processing observations contaminated by outliers. It has also become very popular and is routinely used in adjustment computations (GHILANI, 2017). This testing procedure consists of screening each individual observation for the presence of an outlier. The test statistic employed in the data snooping is given by a normalised least-squares residual, and it is well-known as w-$test$ (BAARDA, 1968). The w-$test$, which is based on a linear mean-shift model, can also be derived as a particular case of the generalised likelihood ratio test (TEUNISSEN, 2006).

[1] Universidade Federal de Uberlândia, malubonimani@hotmail.com
[2] Universidade Federal de Uberlândia, caiquesimionato@hotmail.com
[3] Universidade Federal de Uberlândia, caio.campos@ufu.br
[4] Universidade Federal de Uberlândia, vfrofatto@gmail.com
[5] Instituto Federal de Santa Catarina, ivandroklein@gmail.com
[6] Universidade Federal de Uberlândia, tomiomatsuoka@gmail.com

In principle, $w\text{-}test$ only makes a decision between the null, denoted by $\mathcal{H}_0$, and a single alternative hypothesis, say $\mathcal{H}_a$. The model under the null hypothesis is often formulated under the condition of absence of outliers, whereas the alternative model is proposed when there is an outlier in the dataset. In order to verify whether the alternative hypothesis is significant or not (i.e., whether we accept or reject the null hypothesis), the $w\text{-}test$ statistic is compared with its critical values (i.e., the percentile of its probability distribution), which can be taken from well-known standard normal distribution. In that case, rejection of null hypothesis $\mathcal{H}_0$ automatically implies acceptance of the alternative hypothesis $\mathcal{H}_a$, and vice versa (IMPARATO; TEUNISSEN; TIBERIUS, 2019).

In that case, the probability level associated with data snooping is restrict to Type I error (rejection of a true null hypothesis) and Type II error (acceptance of a false null hypothesis). The probability of committing the type I error is well-known significance level or type I error rate (denoted by α_0), which is controlled by the user. However, data snooping is by nature a procedure that involves multiple hypothesis testing (TEUNISSEN; IMPARATO; TIBERIUS, 2017; ZAMINPARDAZ; TEUNISSEN, 2019). The multiple hypothesis testing problem occurs when a number of individual hypothesis tests are considered simultaneously. This means testing $\mathcal{H}_0$ against $\mathcal{H}_a^{(1)}$, $\mathcal{H}_a^{(2)}$,…, $\mathcal{H}_a^{(n)}$ (LEHMANN, 2012).

To make it clearer, let's start by assuming that the individual tests are independent and the significance level for each test (α_i) be α_0; then the probability that one of the tests is rejecting the null hypothesis is given by (LEHMANN; LÖSLER, 2016):

$$\alpha' = 1 - \prod_{i=1}^{n}(1 - \alpha_i) = 1 - (1 - \alpha_0)^n \tag{1}$$

which α' is the the probability of making one or more false positives, or type I errors, when performing multiple hypotheses tests. It is known as family-wise error rate (FWER).

In this case, the significance level or the type I error rate of individual tests (α_0) no longer represents the error rate of the combined set of tests (denoted by α'). Therefore, we are interested in controlling α'. In this context, methods that deal with multiple alternative hypotheses are referred to as multiple comparison methods. Investigations on such methods are rare in geodetic applications, which, in a way, opens the way for research and applications. More details can be seen, for example, in (MILLER, 1981; SIMES, 1986; WRIGHT, 1992; SARKAR; CHANG, 1997; LEHMANN; ROMANO, 2005; ROM, 2013).

One of the simplest methods employed to control the FWER is known as Šidák correction Šidák (1967). The Šidák correction is derived by assuming that the individual tests are independent, as follows:

$$\alpha_0 = 1 - (1 - \alpha')^{\frac{1}{n}} \ (2) \tag{2}$$

The goal of the $\check{S}$idák correction in Equation (2) is to adjust α_0 so that the significance level for the entire series of tests α' is warranted. $\check{S}$idák correction produces a family-wise error rate of exactly α' when the tests are independent from each other and all null hypotheses are true. In that case, the critical values for $w\text{-}test$ statistic can be computed as follows:

$$k_{sid} = \Phi_N^{-1}\left[\left(1 - \frac{\alpha'}{2}\right)^{\frac{1}{n}}\right] \tag{3}$$

with Φ_N^{-1} being the inverse of the standard normal cumulative distribution. The number two in the denominator is because the $w\text{-}test$ is two-sided test.

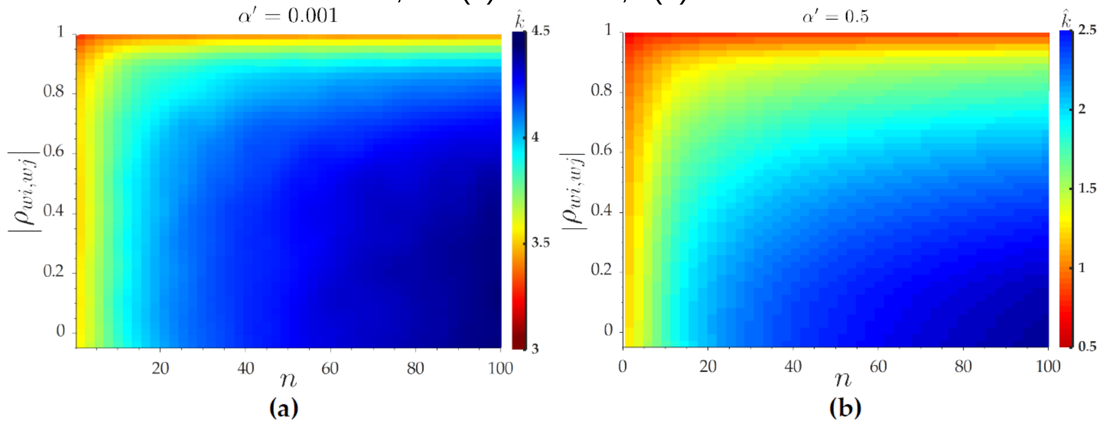
In fact, however, the mathematical model promotes correlation between $w\text{-}test$ test statistics. This means that we will always have some degree of correlation between the tests. If we neglect the correlation between the tests, we overestimate the critical values computed from $\check{S}$idák correction in Equation (3) (LEHMANN, 2013).

In this contribution, we apply a procedure based on Monte Carlo simulation in order to obtain the critical values that considers the correlation between the $w\text{-}test$ statistics. The drawback is that every time data snooping is run, Monte Carlo method is used to compute the critical values (ROFATTO et al., 2020b). To overcome this issue, here, on the other hand, we first compute a series of critical values for a fixed number of observations with pre-fixed correlation between the $w\text{-}test$ statistics by using Monte Carlo. Then, a Supervised Back Propagation Neural Network (SBPNN) architecture was trained and tested using such critical values databases. The purpose of the SBPNN model is to suppress the use of Monte Carlo method when the data snooping is in play. Furthermore, we provide a MATLAB's flexible network object type (called SBPNN.mat) which allows anyone to obtain the critical values quickly and easily, in addition to ensuring good control of the type I error rate when applying Data Snooping.

2 MATERIAL AND METHODS

In total 1.200 critical values for Data Snooping test were computed and stored. For instance, Figure (1) shows an example of the behavior of the critical values for the extreme cases of having α'=0,001 and α'=0,5 in function of the number of observations n and $\rho_{wi,wj}$.

Figure 1 – Critical values for Data Snooping computed by Monte Carlo method for α′=0,001 (a) and α′=0,5 (b).



Fonte: Rofatto et al. (2021).

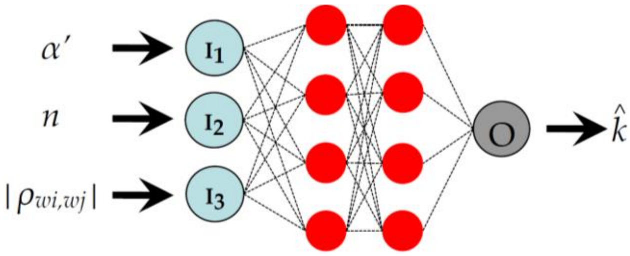
In general, we can observe the following relationship Equation (4) e (5) (↑ means increase and ↓ decrease):

$$\uparrow \rho_{wi,wj} \rightarrow \downarrow \hat{k} \ and \ \uparrow n \rightarrow \uparrow \hat{k} \ or \ \downarrow \rho_{wi,wj} \rightarrow \uparrow \hat{k} \ and \ \downarrow n \rightarrow \downarrow \hat{k} \quad (4)$$

$$\downarrow \rho_{wi,wj} \ and \ \uparrow n \rightarrow \uparrow \hat{k} \ or \ \uparrow \rho_{wi,wj} \ and \ \downarrow n \rightarrow \downarrow \hat{k} \quad (5)$$

From Figure (1), we also observe that the critical values do not depend on of the size of samples n when $\rho_{wi,wj}$>0,99. This fact was also presented by (YANG et al., 2013). We also note that for n>30 and for a given $\rho_{wi,wj}$, the critical values are virtually constant. These datasets were used to build a neural network for computation the critical values for any correlation matrix R_w(Figure 2). Details of the neural network can be found in Rofatto et al. (2021).

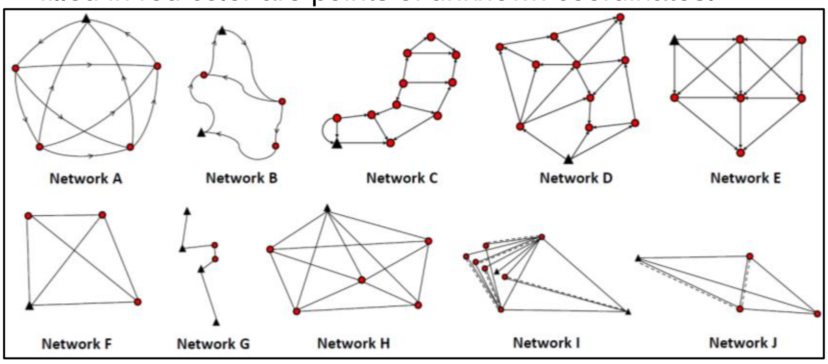
Figure 2 – Topology of the feed-forward neural network with Levenberg-Marquardt back-propagation algorithm for prediction of critical values.



Fonte: Rofatto et al. (2021).

We considered ten geodetic networks for the experiments Figure (3), namely: five levelling networks (A, B, C, D and E), two horizontal networks (F and G) and three GNSS (Global Navigation Satellite System) baseline networks by static relative positioning method (H, I and J).

Figure 3 - Geodetic Networks: observations are displayed in solid black line; repeated observations in dash line; the black triangles represent the control points, whereas the circles filled in red color are points of unknown coordinates.



Fonte: Rofatto et al. (2021).

The SBPNN-based critical values were obtained according to the following Equations (6) (7) (8) (9) (10):

$$k_{SBPNN\ (mean)} = k_{SBPNN\ (mean|\rho_{wi,wj}|)} \quad (6)$$

$$k_{SBPNN\ (max/mean)} = \frac{k_{SBPNN\ (max|\rho_{wi,wj}|)} + k_{SBPNN\ (mean|\rho_{wi,wj}|)}}{2} \quad (7)$$

$$k_{SBPNN\ (max/min)} = \frac{k_{SBPNN\ (max|\rho_{wi,wj}|)} + k_{SBPNN\ (min|\rho_{wi,wj}|)}}{2} \quad (8)$$

$$k_{SBPNN(max_{eig})} = k_{SBPNN\ (pc)} \quad (9)$$

$$pc = \sqrt{\frac{max_{eig}(R_w)}{n}} \quad (10)$$

where $K_{SBPNN(.)}$ is the overall neural-network-based critical value. The subscript within the parenthesis corresponds to each criterion, as follows: (max/mean) corresponds to the mean of the critical values computed by the SBPNN-based critical value for the maximum $max|\rho_{wi,wj}|$ and mean $mean|\rho_{wi,wj}|$ absolute values of the correlation between the w-$test$ statistics; (max/min) the mean from the maximum $max|\rho_{wi,wj}|$ and minimum $min|\rho_{wi,wj}|$; (mean) corresponds to the the SBPNN-based critical value for the mean $mean|\rho_{wi,wj}|$; (max_{eig}) the SBPNN-based critical value obtained from the square root of the ration between the largest eigenvalue of the correlation matrix R_w (signs of correlations are considered) and number of observations n. The largest eigenvalue represents the maximum amount of information of the R_w and it can be easily obtained using Matlab's "eigs" command. In the context of principal component analysis (PCA), the maximum eigenvalue represents the largest possible variance.

Obviously, the expressions above are computed for a given significance level (α') and number of observations (n), as can be seen in Figure (2). In the next section, we evaluate each of these criteria. The goal is to investigate the extent to which the critical values computed by the Supervised Back-Propagation Neural Network (k_{SBPNN}) deviate from the critical values computed by Monte Carlo method.

3 RESULTS AND DISCUSSION

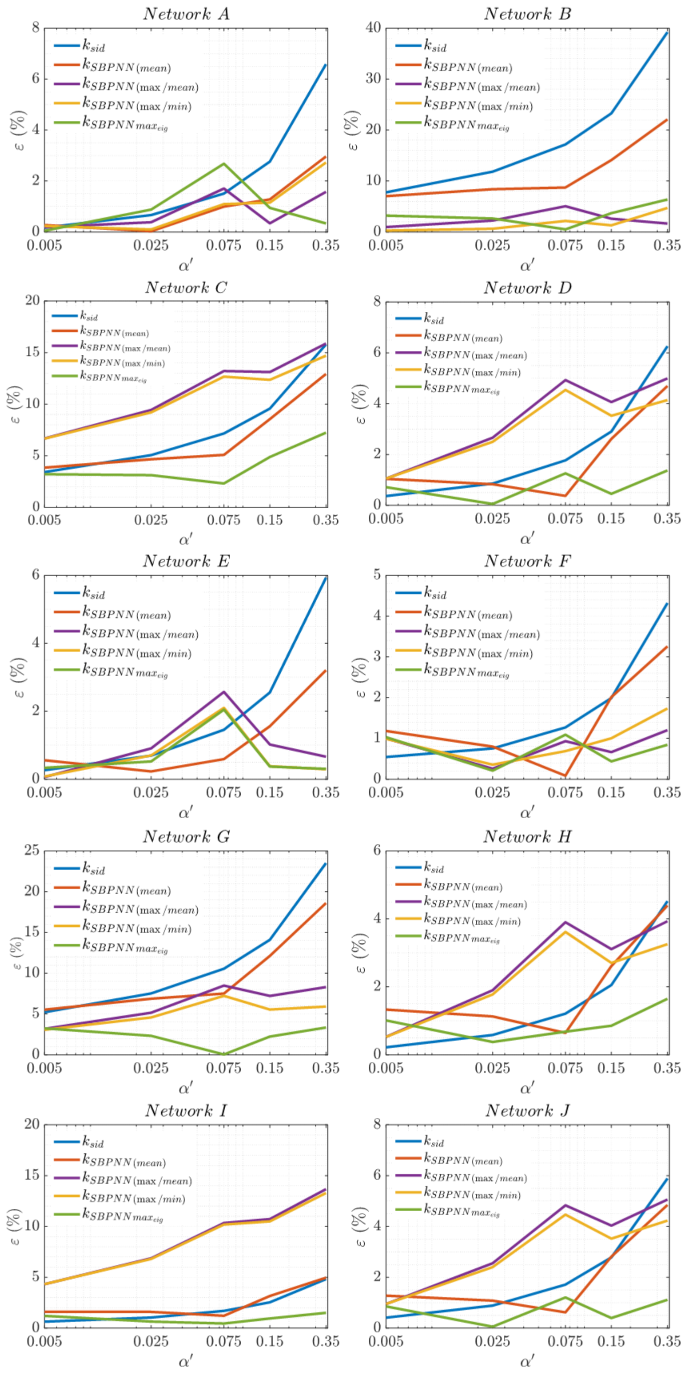
Figure (4) shows that the higher the significance level (α'), the larger the error when applying the $\check{S}$idák correction k_{sid}This is due to the fact that the correlation between the w-$test$ statistics is neglected by the $\check{S}$idák correction and, therefore, it serves as an upper bound for analyzing the error of the SBPNN method.

In the case of the Network (A) and for significance levels lower than 10% (α'<0.1), (k_{SBPNN}) for all criteria and $\check{S}$idák correction (k_{sid}) provided critical values practically equal to those obtained by the Monte Carlo method, with error less than 3% (Figure 9). However, the increase in the significance level (α') has enlarged the error in the case of $\check{S}$idák correction, while the error for all criteria in the SPBNN method remained less than 3%. Similar results can be found for networks (E) and (F).

For the case of Network (B), the errors related to the SBPNN-based critical values for all criteria were less than $\check{S}$idák correction (k_{sid}). In that case, SBPNN for $k_{SBPNN\,(max/mean)}$, $k_{SBPNN\,(max/minn)}$ and $k_{SBPNN\,(max_{eig})}$ had errors less than 6,5%, while k_{sid} reached 39% (α′=0,35). Among the criteria considered, the one based on the mean value $k_{SBPNN\,(mean)}$ presented the highest error (22% for α′=0,35). Similar results can be found for network (G).

For the case of Network (C), SBPNN for $k_{SBPNN\,(max/mean)}$ and $k_{SBPNN\,(max/minn)}$ and for α′<0.15 had larger errors than $\check{S}$idák correction (k_{sid}), with maximum error of the order of ~16% (α′=0,35) and ~14\% (α′=0,35), respectively. On the other hand, $k_{SBPNN\,(max_{eig})}$ had error less than 7,5%. A similar relationship can be found for network (I), but both $k_{SBPNN\,(max/mean)}$ and $k_{SBPNN\,(max/minn)}$ with errors larger than the k_{sid} for all significance levels α′. In latter case, $k_{SBPNN\,(max_{eig})}$ had error less than 1,5%. Actually, there is the same behavior for networks (D), (H) and (J), but the magnitude of the errors for both $k_{SBPNN\,(max/mean)}$ and $k_{SBPNN\,(max/minn)}$ are much smaller, with the maximum of the order of ~5%.

Figure 4 – Relative error (ε) of the critical values for $\check{S}$idák correction (sid) and SBPNN model based on each criterion of the correlation



Fonte: Rofatto et al. (2021).

In general, therefore, the criterion "max_{eig}" based on the expression in Equation (10) provides a better balance between the correlation and the number of observations for the computation of the critical values.

4 CONCLUSIONS

Considering the critical values obtained by Monte Carlo as references, SBPNN method presented a mean relative error of ~2% ($\pm$2%) and a maximum of 7%, whereas the $\check{S}$idák correction about ~9%($\pm$9%) and maximum of 54%. Therefore, we observe that SBPNN is able to capture the dependence of the test statistics, and therefore it can be considered as good approximations for the control of the false positive rates (α'). Since $\check{S}$idák correction does not take into account the correlation between the test statistics, we reinforce that its use for controlling the type I error rate should only be used for systems with high redundancy (r_i>0.5), low correlation between w-$test$ statistics ($|\rho_{wi,wj}| \leq$ 0,05) and for low rates of individual false positives $\alpha_0 <$ 0,01 (1%).

Finally, we provide a MATLAB's flexible network object type (called SBPNN.mat) that allows anyone to obtain the desired critical value with good control of type I error for the case where the random errors follow the normal distribution. We also provide a MATLAB's function (called "kNN.m") which uses {SBPNN.mat} and the criterion max_{eig} to compute a desired critical value for the case where Gauss-Markov model is in play. Those interested in dataset of this work, {SBPNN.mat and "kNN.m" please send their request to the authors by e-mail or access https://data.mendeley.com/datasets/77sfpx9b74/3.

REFERENCES

ANDERSON, T. An introduction to multivariate statistical analysis. 2nd ed. **Wiley, New York**, 1984.

BAARDA, W. A. Testing Procedure for Use in Geodetic Networks. **Publications on Geodesy**, v. 2, n.5, 97 p., 1968.

GHILANI, C.D. Adjustment computations: Spatial data analysis. 6 ed. **John Wiley Sons**, Ltd, 2017.

IMPARATO, D., TEUNISSEN, P.,TIBERIUS, C. Minimal detectable and identifiable biases for quality control. **Survey Review**, v. 51, n. 367, p. 289–299. 2019. DOI: 10.1080/00396265.2018.1437947.

LEHMANN, E.L., ROMANO, J.P.Testing statistical hypotheses. 3 ed. **Springer-Verlag New York**, 2005.

LEHMANN, R. Improved critical values for extreme normalized and studentized residuals in Gauss–Markov models. **Journal of Geodetic Science**, v. 86, n. 12, p. 1137–1146, 2012. DOI. 10.1007/s00190-012-0569-0.

LEHMANN, R. On the formulation of the alternative hypothesis for geodetic outlier detection. **Journal of geodesy**, v. 87, n. 4, p. 373–386, 2013.

LEHMANN, R., LÖSLER, M. Multiple outlier detection: Hypothesis tests versus model selection by information criteria. **Journal of surveying engineering**, v. 42, n. 4, 2016. DOI: 10.1061/(ASCE)SU.1943-5428.0000189.

MILLER, R.G.J. Simultaneous statistical inference. 2 ed. **Springer-Verlag New York**, 1981.

MITTELHAMMER, R.C., JUDGE, G.G., MILLER, D.J. Econometric foundations. 1st ed. **Cambridge University Press**, 2000.

ROFATTO, V., MATSUOKA, M., KLEIN, I. Design of geodetic networks based on outlier identification criteria: An example applied to the leveling network. **Boletim de Ciências Geodésicas**, v. 24, n. 2, p. 152–170, 2018. DOI: 10.1590/s1982-21702018000200011.

ROFATTO, V. F.; MATSUOKA, M. T.; KLEIN, I.; VERONEZ, M. R., DA SILVEIRA, L. G. A Monte Carlo Based outlier diagnosis method for sensitivity analysis. **Remote Sensing**, v. 12, n. 5, p. 1-41, 2020. DOI: 10.3390/rs12050860.

ROFATTO, V. F.; MATSUOKA, M. T., KLEIN, I., BONIMANI, M. L. S., RODRIGUES, B. P., CAMPOS, C. C., VERONEZ, M. R., SILVEIRA JR, L. G. An Artificial Neural Network-based Critical Values for Multiple Hypothesis Testing: Data Snooping case. **Survey Review**, v. 53, p. 1-16, 2021. DOI:10.1080/00396265.2021.1968176

ROM, D.M. An Improved Hochberg Procedure for Multiple Tests of Significance. **The british journal of mathematical and statistical psychology**, v. 66, n. 1, p. 189-196, 2013. DOI: 10.1111/j.2044-8317.2012.02042.x.

SARKAR, S.K., CHANG, C.K. The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics. **Journal of the american statistical association**, v. 92 n. 440, p.1601–1608, 1997.

ŠIDÁK, Z. Rectangular confidence regions for the means of multivariate normal distribuitions. **Journal of the american statistical association**, v. 62, n. 318, p. 626–633, 1967.

SIMES, R.J. An Improved Bonferroni Procedure for Multiple Tests of Significance. **Biometrika**, v. 73, n. 3, p. 751–754, 1986. DOI: 10.1093/biomet/73.3.751

TEUNISSEN, P. Testing Theory: an introduction. 2º ed. **Delft University Press**, 2006.

TEUNISSEN, P.J.G., IMPARATO, D.,TIBERIUS, C.C.J.M. Does RAIM with correct exclusion produce unbiased positions?. **Sensors**, v.17, n. 7, 2017. DOI: 10.3390/s17071508

WRIGHT, S.P. Adjusted p-values for simultaneous inference. **Biometrics**, v. 48, n. 4, p. 1005– 1013, 1992. DOI: http://www.jstor.org/stable/2532694.

YANG, L.; WANG, J.; KNIGHT, N. L.; SHEN, Y. Outlier separability analysis with a multiple alternative hypotheses test. **Journal of Geodesy Science**, v. 87, n. 6, p. 591–604, 2013. DOI.10.1007/s00190-013-0629-0.

ZAMINPARDAZ, S., TEUNISSEN, P. DIA-datasnooping and identifiability. **Journal of Geodesy**, v. 93, n. 1, p. 85–101, 2019. DOI: 10.1007/s00190-018-1141-3