# Empirical estimation of the power of test in outlier detection problem

BAHATTIN ERDOGAN, SERIF HEKIMOGLU, UTKAN MUSTAFA DURDAG AND TAYLAN OCALAN

Department of Geomatic Engineering, Faculty of Civil Engineering, Yildiz Technical University, Istanbul, Turkey (berdogan@yildiz.edu.tr)

*Received: July 19, 2018; Revised: November 16, 2018; Accepted: November 29, 2018*

## ABSTRACT

*Classical outlier detection test methods such as Baarda test and Pope test are generally preferred in geodetic problems. They depend on the Least Square Estimation (LSE) and LSE is very sensitive to the variations of the model. The capacity of the LSE changes depending on the different significance level, different type of outlier, the number of outlier, magnitude of outlier, number of observations and the number of unknowns. In statistics, the power of test is the probability of rejecting the null hypothesis when the null hypothesis is false. It is a theoretical assumption and depends on the significance level $\alpha$ (Type I error) and $\beta$ (Type II error). The different types of the outliers, such as random or non-random, affect the results of the test methods; but the power of test is the same for all different types of the outliers. In this study, empirical estimation of the power of test is presented as Mean Success Rate (MSR). The theoretical power of test and empirical MSR have been estimated for univariate model and linear model by using Baarda test; according to the obtained results, MSR can be used as empirical value of the power of test and capacity of the test models. Also, MSR reflects more realistic results than the theoretical power of test.*

Keywords: power of test, efficacy, mean success rate, Baarda test, outlier

## 1. INTRODUCTION

Outlier detection problem has a great importance in geodetic applications and Baarda test is often used to detect outliers (*Baarda, 1968*). Least Square Estimation (LSE) is preferred in Baarda test. Generally, it works well in practice when observations include only one single outlier. LSE is very sensitive against the deviations from model assumptions (*Hampel et al., 1986*), and it smears the effect of the outliers not only on the residuals of the corrupted observations, but also the other remained good observations (*Hekimoglu et al., 2011; Hekimoglu and Erdogan, 2012*). Therefore, the efficacies of the test methods depending on the LSE should be investigated to make correct interpretations on the results. In statistics, theoretically power of test is used to measure the capacity of the test method which is efficient for the Baarda test. Moreover, the alternative hypothesis must be formed for a special value to compute the power of test. It measures the reliability

of a test for the given fixed level of significance $\alpha$ and the greater the power of test, the greater the efficacy of a test.

In practice, expectation value of normally distributed random observations $\mu$ and standard deviation $\sigma$ are not known before the analysis. Therefore, it is not possible to compute the power of test. On the other hand, it is mostly sensed that the power of test is only a design parameter used for optimizing deformation networks or control networks (*Kuang, 1991*; *Wang and Chen, 1999*; *Knight et al., 2010*; *Aydin, 2012*; *Wang and Knight, 2012*). If the covariance matrix of observations is given realistically beforehand, the power of test gives a realistic result about how the considered changes will be discriminated against the random errors (*Aydin, 2012*).

Theoretical power of test involves some risks. If $\alpha$ is chosen large, the small outliers may be detected by the test; but some good observations may be detected as outlier (*Hekimoglu, 1997*; *Lehman, 2010, 2012a,b*). If $\alpha$ is chosen small, the small outliers may not be detected. But, the smaller the significance level $\alpha$ is chosen, the more frequently we are inclined to accept null hypothesis and the more frequently we will accept it when it is actually false (*Lehman, 2012a,b*). For these different cases Type I error or Type II error may be arised. The main goal is to minimize the probabilities of these decision errors. In addition, *Lehman (2012b)* proposed to improve the critical values of the tests. He suggests the way to improve by using a procedure based on the Monte Carlo method for the numerical computation of such critical values. Monte Carlo methods are able to compute statistical quantities numerically. He argued that the improved critical values detect more outliers than those computed by Baarda test when the same level of significance $\alpha$ is chosen.

The capacity of the test method also can be measured by the Mean Success Rate (*MSR*), which is the success rate divided by total number of experiments by using Monte Carlo method (*Hekimoglu and Koch, 1999, 2000*). *Aydin (2012)* showed that the *MSR* is the estimated value of the power of the global test in deformation analysis.

In geodetic applications the observation vector $\boldsymbol{L}$ is used for the estimation of the unknown parameters such as point coordinates. Before the adjustment step, outlier detection step is applied and the power of test is very important for outlier detection process; because undetectable outliers affect the results of LSE. If the numbers of the observations and unknown parameters change, the results of the outlier detection test methods change, too. However, the theoretical power of test does not depend on the number of the observations and the number of unknowns, it is assumed that the number of observations goes to the infinity and the number of unknowns parameters are two ($\mu$, $\sigma$).

In this study, a simulation study is carried out to estimate the empirical power of test by using *MSR*. The *MSR* and the values of the power of test have been estimated in both univariate model and linear model for one and more outliers by using Baarda test. It has been seen that power of test does not change depending on the different type of outlier, number of observations and the number of unknowns; but it is expected that the capacities of the method should change when these parameters differ. Eventually, it is shown that the estimated MSRs are more realistic than the theoretical power of test and it can be considered as empirical power of test.

## 2. TEST OF STATISTICAL HYPOTHESIS

In order to test a hypothesis, two cases arise: the first hypothesis to be tested is called as "the null hypothesis" denoted by $H_0$, the second one is called as "alternative hypothesis" denoted by $H_A$. If the null hypothesis $H_0$ is false in reality, the alternative hypothesis $H_A$ is true, and vice versa. Generally it is said that $H_0$ is tested against or versus $H_A$.

In geodesy a scalar random variable $L$ is measured more than once and the values form the observation vector $\boldsymbol{L}$. It is assumed that $\boldsymbol{L}$ obeys the normal distribution. The expectation value $E(\boldsymbol{L})$ of the random values (observations) is $\mu$. It is assumed that in the null hypothesis $\mu$ is equal to $\mu_0$ and in the alternative hypothesis $\mu_A$. This assumption is the statistical hypothesis to be tested. The null hypothesis $H_0$ is formulated as:

$$H_0 : E(\boldsymbol{L}) = \mu_0 , \tag{1}$$

and the alternative hypothesis $H_A$ as

$$H_A : E(\boldsymbol{L}) = \mu_A \neq \mu_0 . \tag{2}$$

The problem is testing of the null hypothesis $H_0$ against the alternative hypothesis $H_A$. During the testing $H_0$, firstly a single observation in the observation vector $\boldsymbol{L}$ is considered; then the question is to obtain what is the critical value for accepting $H_0$ and rejecting $H_0$. The critical region $K$ of a test, which is at the right tail or the left tail of the distribution, are constituted by the rejecting values of $\boldsymbol{L}$.

## 3. TWO TYPES OF ERRORS

In order to specify a statistical test, first the significance level $\alpha$ is given. The null hypothesis $H_0$ is rejected if the sample value or the observations of $\boldsymbol{L}$ is in the tailed area. Otherwise, the null hypothesis $H_0$ is accepted. After this decision two types of errors may be formed:

**Type I error**: Rejection of $H_0$ when in reality $H_0$ is true.

**Type II error**: Acceptance of $H_0$ when in reality $H_0$ is false.

The size of a Type I error is defined as the probability $P$ that a sample value of $\boldsymbol{L}$ is in the tailed area, which can be called as critical region $K$, when $H_0$ is true in reality. This probability is denoted by $\alpha$.

$$\alpha = P\left(\boldsymbol{L} \in K \middle| H_0\right) = \int_K f\left(L \middle| H_0\right) dL . \tag{3}$$

If the probability density function $f(\boldsymbol{L})$, which shows mostly normal distribution, and the critical region $K$ are given, the significance level of the test can be computed. The size of a Type II error is defined as the probability that a sample value of observation vector $\boldsymbol{L}$ falls outside the critical region $K$ when $H_0$ is false in fact. This probability is denoted by $\beta$:

$$\beta = P\left(\boldsymbol{L} \notin K \middle| \mathrm{H_A}\right) = 1 - \int_K f\left(L \middle| \mathrm{H_A}\right) dL\,. \tag{4}$$

If the probability density function $f\left(\boldsymbol{L}\right)$ and the critical region $K$ are given under $\mathrm{H_A}$, the size of Type II error $\beta$ can be computed (*Teunissen, 2000*). In addition, the expression $1-\beta$ defines the power of a test $\gamma$:

$$\gamma = 1 - \beta = \int_K f\left(L \middle| \mathrm{H_A}\right) dL\,. \tag{5}$$

To define the best test, we might resonably use the size of two type errors, $\alpha$ and $\beta$. A good test should be a test where $\alpha$ is small (ideally zero) and $\beta$ is small (ideally zero). Therefore, it would be nice if the test minimizes both $\alpha$ and $\beta$, simultaneously. Unfortunately this is not possible. If we decrease $\alpha$, we tend to increase $\beta$, and vice versa (*Teunissen, 2000*). To solve this problem, Neyman-Person principle says that we should fix the size of Type I error $\alpha$, and minimize the size of Type II error $\beta$ (Neyman and Pearson, 1933). Four factors affect the power of a hypothesis test:

1. Sample size ($n$): The greater the sample size, the greater the power of test;
2. Significance level ($\alpha$): The higher the significance level $\alpha$, the higher the power of test;
3. The expected value of the parameter to be tested: The greater is the difference between the expected value of a parameter and the specified value in the alternative null hypothesis, the greater the power of a test.
4. In geodetic application the different networks are established and measured. The configuration of geodetic networks affects also the power of a hypothesis test (*Teunissen, 2000*).

The ideal hypothesis test should reject the null hypothesis when it is false in reality. The power of test is the measure of how good a test is. Also, there is a probability that the test will reject $\mathrm{H_0}$ when it is false in fact.

## 4. AN EXAMPLE FOR THE NORMALLY DISTRIBUTED OBSERVATIONS

In geodetic practice, the observations are mostly normally distributed. The scalar variables are measured and they can be modelled as a random variable vector $\boldsymbol{L}$ with density function $f\left(\boldsymbol{L}\right)$ as follows:

$$f\left(\boldsymbol{L}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(L-\mu_0\right)^2\right). \tag{6}$$

Now we can compute the level of significance $\alpha$ and $\beta$:

$$\alpha = \int_{K_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(L-\mu_0\right)^2\right) dL\,, \tag{7}$$

$$\beta = \int_{-\infty}^{K_\alpha} \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{1}{2\sigma^2} \left( L - \mu_{\mathrm{A}} \right)^2 \right) dL \,, \tag{8}$$

where $K_\alpha$ is the critical value for the chosen significance level $\alpha$. The level of significance $\alpha$ can usually be computed by using the tables given in statistical handbooks for standard normal distribution. The normal distribution is transformed to standard normal distribution as follows:

$$z = \frac{L - \mu_0}{\sigma} \,, \tag{9}$$

where $L \sim N\left( \mu_0 , \sigma \right)$, $N$ denotes the gaussian normal distribution, and $z$ is the random variable which is standard normal distributed under $\mathrm{H}_0$. The statistical table for the normal distribution is computed by using the standard normal distribution. Levels of significance $\alpha$ and $\beta$ and power of test $\gamma$ can be determined as follows (*Teunissen, 2000*):

$$\alpha = P\left( L > K_\alpha \left| \mathrm{H}_0 \right. \right) = P\left( z > \frac{K_\alpha - \mu_0}{\sigma} \middle| \mathrm{H}_0 \right) \tag{10}$$

or

$$\alpha = \int_{(K_\alpha - \mu_0)/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} z^2 \right) dz \,, \tag{11}$$

$$\beta = \int_{-\infty}^{(K_\alpha - \mu_{\mathrm{A}})/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} z^2 \right) dz \,, \tag{12}$$

$$\gamma = 1 - \beta = \int_{(K_\alpha - \mu_{\mathrm{A}})/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} z^2 \right) dz \,. \tag{13}$$

In geodetic applications the different observation vector $\boldsymbol{L}$ is used for the estimation of the unknown parameters such as point coordinates. Outlier detection is very important for the optimum estimation of the parameters. If the numbers of the observations and unknown parameters change, the results of the outlier detection test method and parameter estimation change, too. However, the theoretical power of test does not include the number of the observations and the number of unknowns, it is assumed that the number of observations goes to the infinity and the number of unknowns parameters are two, i.e. $\mu$ and $\sigma$). The location of the critical region $K$ (Type I Error) is determined by the value chosen for $K_\alpha$ that is the critical value of the test. Here, the problem is which value should be chosen for $K_\alpha$. After the test has been explicitly formulated, we can obtain whether the sample or observation falls into the critical region $K$ or not. We can never claim that the hypotheses have been false or true before testing (*Teunissen, 2000*).

## 5. BAARDA TEST

A test procedure can be presented in the context of Detection, Identification and Adaptation (DIA) (*Teunissen, 2018*). The observations may be corrupted by blunders (outliers), or the chosen model may fail to give an enough description of physical reality. These mistakes may be occasionally happened and they must be detected.

$$\boldsymbol{L}+\boldsymbol{v}=\mathbf{A}\boldsymbol{x}, \quad \mathbf{C}_{ll}=\sigma_0^2\mathbf{P}^{-1}, \tag{14}$$

$$\hat{\boldsymbol{x}}=\left(\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{P}\boldsymbol{L}, \tag{15}$$

$$\mathbf{Q}_{\hat{x}\hat{x}}=\left(\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A}\right)^{-1}, \tag{16}$$

$$\mathbf{Q}_{ll}=\mathbf{A}\mathbf{Q}_{\hat{x}\hat{x}}\mathbf{A}^{\mathrm{T}}, \tag{17}$$

$$\mathbf{Q}_{vv}=\mathbf{P}^{-1}-\mathbf{Q}_{ll}, \tag{18}$$

where $\boldsymbol{L}$ is the $n \times 1$ vector of observations, $\boldsymbol{v}$ is the $n \times 1$ residual vector, $\mathbf{A}$ is the $n \times u$ design matrix, $\mathbf{C}_{ll}$ is the $n \times n$ covariance matrix of the observations, $\mathbf{P}$ is the diagonal $n \times n$ weight matrix, $\hat{\boldsymbol{x}}$ is the $u \times 1$ vector of unknown parameters, $\sigma_0^2$ is the variance with unit weight, $\mathbf{Q}_{\hat{x}\hat{x}}$ is the cofactor matrix of the unknown parameters, $\mathbf{Q}_{ll}$ is the cofactor matrix of the observations, $\mathbf{Q}_{vv}$ is the cofactor matrix of the residuals, $n$ is the number of observations and $u$ is the number of unknown parameters.

The null hypothesis $\mathrm{H}_0$ is defined as:

$$\mathrm{H}_0 : \boldsymbol{L} \sim N\left(\mathbf{A}\boldsymbol{x}, \sigma_0^2\mathbf{Q}_{ll}\right). \tag{19}$$

If we want to find out whether $i$-th observation is erroneous or not, we can specify the model in the alternative hypothesis as:

$$\mathrm{H}_\mathrm{A} : \boldsymbol{L}_i \sim N\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{c}_{yi}, \sigma_0^2\mathbf{Q}_{ll}\right), \tag{20}$$

with $\boldsymbol{c}_y$ for $i$-th element is:

$$\boldsymbol{c}_{yi}=\left[0,\ldots,0,1,0,\ldots 0\right]^{\mathrm{T}}, \tag{21}$$

where $\Delta$ is a scalar (outlier) non-random error in the observation. Depending on the DIA method, at the detection step, an overall model test on $\mathrm{H}_0$ is performed to diagnose whether an unspecified model error has occurred or not and then identification step is carried out (*Teunissen, 2018*). At the identification step, the normalized residuals can be estimated as (*Baarda, 1967*; *Teunissen, 2000*; *Zaminpardaz and Teunissen, 2018*):

$$w=\frac{\boldsymbol{c}_y^{\mathrm{T}}\mathbf{Q}_{ll}^{-1}\boldsymbol{v}}{\sigma_0\sqrt{\boldsymbol{c}_y^{\mathrm{T}}\mathbf{Q}_{ll}^{-1}\mathbf{Q}_{vv}\mathbf{Q}_{ll}^{-1}\boldsymbol{c}_y}}. \tag{22}$$

$\mathrm{H}_0$ is rejected if $w<\sqrt{-K_\alpha}$ or $w>\sqrt{K_\alpha}$ (*Teunissen, 2000*).

If the a priori variance of unit weight $\sigma_0^2$ is known, the normalized residual vector $\boldsymbol{w}$ is estimated to obtain test statistic as follows:

$$w_i = \frac{|v_i|}{\sigma_0 \sqrt{(q_{vv})_{ii}}} \sim N(0,1). \tag{23}$$

If $\max|w_i| > z_{1-(\alpha/2)}$, then the *i*-th observation is considered as an outlier. Here, *Baarda (1968)* had chosen $\alpha = 0.001$. *Teunissen (2000)* calls Eq. (23) as *w*-statistic and showed that this *w*-statistic is a simple likelihood ratio test of size $\alpha$ and the vector $\boldsymbol{w}$. For the *i*-th observation Eq. (22) is equivalent to Eq. (23). Generally if the observation with $\max|w_i| > z_{1-(\alpha/2)}$, the null hypothesis should be rejected.

If the observations contain two outliers, then the vector $\boldsymbol{c}_y$ becomes the $n \times n$ matrix. The $n \times 1$ vector $\Delta$ is given as:

$$\Delta = \left[0, \ldots, 0, \delta_i, \ldots, 0, \delta_j, \ldots, 0\right]^{\mathrm{T}}. \tag{24}$$

In this case, Baarda test performs iteratively. For each step, only one outlier is identified and then the Baarda test is applied again. This procedure, also called as "data snooping", is repeated until there is no outlier in the observations.

## 6. COMPUTING THE THEORETICAL POWER OF TEST

The conventional outlier detecting procedures, widely used in geodetic practice, are based on LSE (*Baarda, 1968*; *Pope, 1976*; *Heck, 1981*; *Kok, 1984, 1999*; *Chen et al., 1987*). Even if the observations have only one single outlier, the results from LSE, such as the residuals, the unknown parameters and their variances, are affected by this outlier because of smearing effect of LSE. Generally after the LSE, the test decision is considered successful when both type errors are small. Usually, the level of significance $\alpha$ is firstly fixed as $\alpha$=0.001 (for Baarda test), then $\beta$ or the power of test, i.e. $1-\beta$, is computed. One requires approximately that the power of test would become 80%. When the $\alpha$ is fixed, the greater the efficiacy of the test, the greater the power of test. The power of test may be interpreted as the reliability of the test.

To compute the power of test, the alternative hypothesis must be given as a special value. Let's consider the alternative hypothesis given in Eq. (20). The power of test $\gamma$ can be given for outlier in geodetic networks as:

$$\begin{aligned} \gamma = P\left(w > K_\alpha \,|\, \mathrm{H_A}\right) &= \int_{K_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}\left(w-\Delta_w\right)^2\right) dw \\ &= \int_{(K_\alpha - \Delta_w)/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) dw. \end{aligned} \tag{25}$$

Here,

$$\Delta_w = \sqrt{\boldsymbol{c}_y^{\mathrm{T}} \mathbf{Q}_{ll}^{-1} \mathbf{Q}_{vv} \mathbf{Q}_{ll}^{-1} \boldsymbol{c}_y}\, \Delta, \quad \Delta = \left[0, \ldots, 0, \delta_i, 0, \ldots, 0\right]^{\mathrm{T}} \neq 0 . \tag{26}$$

If we consider the DIA method, the relation between the probabilities of the missed detection (MD), correct identification (CI) and wrong identification (WI) can be written as (*Teunissen, 2018*)

$$1 = P_{MD} + P_{CI} + P_{WI} . \tag{27}$$

In case of the estimation of the theoretical power of test, since theoretical power of test do not consider the identification step and it considers only the detection step, the probability of wrong identification is identically zero, $P_{WI} = 0$, and we have $P_{CI} = 1 - P_{MD}$ (*Teunissen, 2018*).

Also, there is an other question is that how can we compute the power of test for more than one alternative hypothesis such as $m_{\mathrm{A1}}$ and $m_{\mathrm{A2}}$. We can constitute more than one alternative hypothesis as follows (*Yang et al., 2013*):

$$\mathrm{H}_0 : \mu_0 = 0, \quad \mathrm{H}_{\mathrm{A1}} : \mu_0 \neq m_{\mathrm{A1}}, \quad \mathrm{H}_{\mathrm{A2}} : \mu_0 \neq m_{\mathrm{A2}}, \tag{28}$$

where $m_{\mathrm{A1}} < m_{\mathrm{A2}}$.

The normal distributions of these two alternative hypotheses may be given as:

$$f(x)_1 = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left( -\frac{1}{2\sigma_1^2} \left(x - \mu_{\mathrm{A1}}\right)^2 \right) \quad \text{for } \mathrm{H_{A1}}, \tag{29}$$

$$f(x)_2 = \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp\left( -\frac{1}{2\sigma_2^2} \left(x - \mu_{\mathrm{A2}}\right)^2 \right) \quad \text{for } \mathrm{H_{A2}}. \tag{30}$$

Let us consider a random event $C_1$ which is the area bounded by $K_\alpha$, $\mathrm{H_{A1}}$ and $\infty$ in Fig. 1. The probability of $C_1$ is the power of test depending on the $\mathrm{H_{A1}}$ and $\alpha$. It may be given as:

$$P(C_1) = \int_{(K_\alpha - \mu_{\mathrm{A1}})/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left( -\frac{1}{2} \left( \frac{K_\alpha - \mu_{\mathrm{A1}}}{\sigma} \right)^2 \right) dz . \tag{31}$$

Let us consider another random event $C_2$ which is the area bounded by $K_\alpha$, $\mathrm{H_{A2}}$ and $\infty$. The probability of $C_2$ is the power of test depending on $\mathrm{H_{A2}}$ and $\alpha$, given as:

$$P(C_2) = \int_{(K_\alpha - \mu_{\mathrm{A2}})/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left( -\frac{1}{2} \left( \frac{K_\alpha - \mu_{\mathrm{A2}}}{\sigma} \right)^2 \right) dz . \tag{32}$$

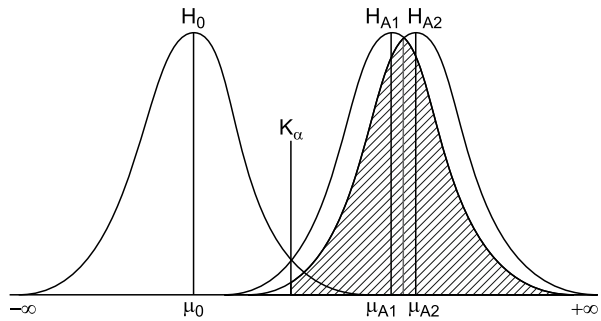


**Fig. 1.** Scheme of the probability of the occurence of the event $B$ (Eq. (34), shaded area). $H_0$ is the null hypothesis, $H_{A1}$ and $H_{A2}$ are two alternative hypotheses, $\mu_0$, $\mu_{A1}$ and $\mu_{A2}$ are the corresponding expected values. $K_\alpha$ is the critical value for the chosen significance level $\alpha$.

Since $C_1$ and $C_2$ are independent random events, the probability of $C_1 \bigcap C_2$ is given as:

$$P\left(C_1 \bigcap C_2\right) = P\left(C_1\right) P\left(C_2\right). \tag{33}$$

The probability of $C_1 \bigcap C_2$ is the power of test for two alternative hypotheses $m_{A1}$ and $m_{A2}$, where the alternative hypotheses $H_{A1}$ and $H_{A2}$ are accepted at the same time. Let $C_1$ event be represented by the $i$-th bad observation (outlier) and $C_2$ event by the $j$-th bad one. When the events $C_1$ and $C_2$ occur at the same time, the $B$ event, which is shaded in Fig. 1, occurs too. We can write the probability of $C_1 \bigcap C_2$ as follows:

$$P\left(B\right) = P\left(C_1 \bigcap C_2\right) < P\left(C_1\right) \quad \text{and} \quad P\left(B\right) = P\left(C_1 \bigcap C_2\right) < P\left(C_2\right). \tag{34}$$

The probability of the occurrence of the event $B$ is always smaller than the one of occurrence of the $i$-th bad observation and the one of occurrence of the $j$-th bad observation at the same time. Therefore, statistical tests generally work well in practice when observations include only one single outlier.

## 7. COMPUTING THE EMPIRICAL POWER OF TEST AS MEAN SUCCESS RATE

*Hekimoglu and Koch (1999, 2000)* used the *MSR* for measuring the capacity of the robust methods and tests for outlier methods, repectively. In practice it is impossible to know which observation contains outliers or not. Therefore, the simulation technique is used to generate the observations and outliers for the estimation of the capacities of the methods. In this case, it is possible to know exactly whether an observation contains outlier or not, before carrying out the analysis. After the Baarda test is applied, if the observations, which are identified as being outliers, are the same as the observations known to be contaminated, the method is considered as successful. If the Baarda test misidentifies observations it is considered unsuccessful. *MSR* contains the results of both

detection and identification steps. Thus, as distinct from theoretical power of test, $P_{CI} = 1 - P_{MD} - P_{WI}$. Since the *MSR* takes into account all possible conditions in practice and $P_{WI}$ is generally is very small comparing with $P_{MD}$, It can be considered as the empirical estimation of the theoretical power of test.

Assume a certain observation vector $\boldsymbol{L}$, for a given number m of outliers, many working samples t can be produced by randomly changing the random error vector ($\boldsymbol{e}$) and the magnitudes of the displacement in a given interval (*int*). The *MSR* is then defined as (*Hekimoglu and Koch, 1999*):

$$MSR\left(\boldsymbol{L}, \boldsymbol{e}, m, n, u, int, a, type\right) = \frac{s}{t}, \tag{35}$$

where $n$ is the number of the observation, $u$ is the number of unknown parameters, $\alpha$ is the significance level, *type* is the type of the outliers (random or non-random), and $s$ is the number of the successful results. According to the Eq. (35), the *MSR* of the methods change depending on the different significance level, different type of outlier, the number of outlier, magnitude of outlier, number of observations and the number of unknowns.

## 8. MODELLING OUTLIERS

The performance of the test can only be measured or optimized with respect to a proper alternative hypothesis. In the literature there are two approaches to model the outlier in LSE: mean shift model and variance inflation model. If the outliers should be treated as non-random quantities, the alternative hypothesis must be based on the common mean shift model. If outliers should be treated as random variables, the variance inflation model is appropriate (*Lehman, 2012a*).

In the mean shift model, to formalize the distribution of the "good" observations it is assumed that they are mostly distributed normally, and the bad observations (outliers) are in one single model:

$$F = \left(1-\varepsilon\right)G + \varepsilon H\,, \tag{36}$$

where $F$ denotes the distribution of the mixture of "contaminated normal distribution", $G$ represents the distribution of good observations ($N\left(\mu, \sigma\right)$), $H$ indicates the distribution of the bad observations whose distribution is not known before, $0 \le \varepsilon < 1$ is a known number (*Hampel et al., 1986*; *Huber, 1981*; *Maronna et al., 2006*). $H$ may be any distribution. In addition, it may be a normal distribution $N\left(\mu_1, \sigma_1\right)$ with $\mu_1 \neq \mu$ and $\sigma_1^2 > \sigma^2$.

Depending on the alternative hypothesis given in Eq. (20), the contaminated functional linear model may be written:

$$\bar{\boldsymbol{L}} = \mathbf{A}\boldsymbol{x} + \boldsymbol{e} + \Delta\,, \tag{37}$$

where $\bar{\boldsymbol{L}}$ denotes the contaminated observation vector, $\boldsymbol{e}$ is the random error vector, and $\Delta$ is a $n \times 1$ vector of bias (outlier) (*Cook and Weisberg, 1982*), which contains the outliers such as $\Delta^{\mathrm{T}} = \left[0, \ldots, \delta_i, \ldots, \delta_j, \ldots 0\right]$. This model is called as "mean shift outlier model" (*Atkinson, 1985*; *Koch, 2013a,b*).

If each outlier $\delta$ in the bias of $\Delta$ has different variances, then this model is called as "variance inflation (increasing) model" (*Lehmann, 2012a*; *Koch, 2013a,b*). It means that the outliers do not come out from common distribution *G*, but they come out from the different distributions which have different expectation values and different variances.

*Donoho and Huber (1983)* say that we corrupt the given good sample in many ways to obtain a corrupted finite sample. We consider here only one way of $\varepsilon$-replacement for Monte-Carlo method. The $\varepsilon$-replacement means that if $\boldsymbol{L}_1 = \left[ L_{11}, L_{12}, \ldots, L_{1n} \right]$ is the good observation vector, *m*-th observation of the vector $\boldsymbol{L}_1$ is replaced by outlier such as $\boldsymbol{L}_2 = \left[ L_{21}, L_{22}, \ldots, L_{2m} \right]$. The bad observations lie in the intervals of $\left( -\infty, \mu - z_{1-\alpha/2}\sigma \right)$ and $\left( \mu + z_{1-\alpha/2}\sigma, +\infty \right)$.

### 8.1. Random and non-random outliers

If the production of outlier mechanism works randomly, the outliers are called as "random outliers". The outlier's magnitude must be greater than the corresponding greatest magnitude of good observation. All the outliers may be drawn from a common distribution such as normal distribution, a uniform distribution, or other distributions. When there are two or more outliers in the observations, sometimes outliers may have the same signs positive or negative. These outliers are called as non-random outliers or influential outliers. These outliers affect the results of LSE more badly than the random outliers (*Hadi and Siminoff, 1993*; *Hekimoglu, 1997*).

## 9. MONTE CARLO METHOD

The purpose of this study is to show the relationship of the *MSR* and the power of test for outlier detection problem. The problem was firstly studied in univariate samples, i.e. the number of unknown is one ($u = 1$). Then, a linear model, which the number of unknowns was two, was considered.

### 9.1. Univariate model

$\boldsymbol{L}$ is a random variable vector with Gaussian distribution with $\mu = L_0$ and $\sigma$. One sample includes $n$ values, i.e. $L_1$, $L_2$, …, $L_n$. The random errors $e_i$ are generated by a random generator, and $L_i$ (the good observation) is obtained as follows:

$$L_i = L_0 + e_i, \quad i = 1, 2, \ldots, n, \tag{38}$$

where number of observations $n = 10$, $L_0 = 0$ m, and $\sigma = 2$ cm was used for genering $\boldsymbol{e}$.

A bad observation $\overline{L}_i$ is generated by adding $\delta_i$ (i.e. an outlier) to $L_0$ instead of $e_i$

$$\overline{L}_i = L_0 + \delta_i . \tag{39}$$

In this study, 10000 contaminated samples was obtained depending on "mean shift model" and the Baarda test were applied to these samples. We chose $\alpha = 0.001$ and $z_{1-\alpha/2} = 3.29$. The *MSR*s were studied for the given outlier magnitude interval such as

**Table 1.** Mean Success Rate (*MSR*) and power of test in univariate samples for significance level $\alpha = 0.001$, one outlier and 10 observations. $\sigma$: standard deviaition.

| Outlier Interval | *MSR* [%] | Power of Test |
|---|---|---|
| $3\sigma$–$4\sigma$ | 57.84 ± 5.15 | 59.71 ± 1.08 |
| $4\sigma$–$5\sigma$ | 100 | 88.52 ± 0.51 |
| $5\sigma$–$6\sigma$ | 100 | 98.46 ± 0.10 |

**Table 2.** Mean Success Rate (*MSR*) and power of test in univariate samples for significance level $\alpha = 0.001$, two outliers and 10 observations

| Outlier Interval | Power of Test | Random Outliers | Non-Random Outliers |
|---|---|---|---|
| | | *MSR* [%] | *MSR* [%] |
| $3\sigma$–$4\sigma$ | 46.76 ± 1.70 | 48.14 ± 4.54 | 38.76 ± 4.30 |
| $4\sigma$–$5\sigma$ | 83.36 ± 0.94 | 99.31 ± 1.20 | 92.26 ± 2.40 |
| $5\sigma$–$6\sigma$ | 97.73 ± 0.18 | 100 | 100 |

$3\sigma$–$4\sigma$, $4\sigma$–$5\sigma$ and $5\sigma$–$6\sigma$. The theoretical power of test was computed by using Eq. (13) with $\alpha = 0.001$. It changed depending on the alternative hypothesis. The different alternative hypotheses were constituted for 10000 different samples in each magnitude interval. The values of the power of test in Tables 1 and 2 were estimated as the mean value of estimated 10 000 values. If two outliers $\delta_i$ and $\delta_j$ ($i, j = 1, 2, \ldots, n$; $i \neq j$) are generated, each of them was formed like one outlier. The power of test and the *MSR* increase depending on the rising of outlier magnitude. Also, non-random outliers affect the capacity of the Baarda test badly. Its *MSR* decrease from 48.14% to 38.76%, despite that the power of test is the same for random and non-random outliers.

The number of the observations affects the results of the Baarda test. To show the effects of this parameter, the number of observation was increased to 11, same analysis were achieved and the results were given Tables 3 and 4. The magnitude of the outliers and other criterias are the same as the analysis that their results are given at Tables 1 and 2. Only the number of the observations was changed.

The *MSR* changes depending on the different criteria. In Tables 3 and 4, the number of the observations was changed and the results of the Baarda test were changed, too; but the theoretical power of test was the same as in Tables 1 and 2. It is not true to say that the results of the outlier test are the same for the different cases such as number of the observations is different. However, the power of test is the same because Eq. (13) does not depend on the number of observations. According to these results, *MSR* is the empirical value of the power of test and it reflects the more reliable results for the efficacy of the Baarda test.

### 9.2. Linear model

A simple regression model can be formed as follows:

$$L_i = a + bx_i + e_i, \quad i = 1, 2, \ldots, n, \tag{40}$$

where the unknown parameters are $a = b = 1$, and $\boldsymbol{x}^{\mathrm{T}} = [1, 2, \ldots, 10]$.

**Table 3.** The same as in Table 1, but for 11 observations.

| Outlier Interval | *MSR* [%] | Power of Test |
|---|---|---|
| $3\sigma$–$4\sigma$ | 57.95 ± 5.10 | 59.71 ± 1.08 |
| $4\sigma$–$5\sigma$ | 100 | 88.52 ± 0.51 |
| $5\sigma$–$6\sigma$ | 100 | 98.46 ± 0.10 |

**Table 4.** The same as in Table 2, but for 11 observations.

| Outlier Interval | Power of Test | Random Outliers | Non-Random Outliers |
|---|---|---|---|
| | | *MSR* [%] | *MSR* [%] |
| $3\sigma$–$4\sigma$ | 46.76 ± 1.70 | 52.58 ± 5.10 | 41.00 ± 4.48 |
| $4\sigma$–$5\sigma$ | 83.36 ± 0.94 | 99.55 ± 0.91 | 94.56 ± 2.32 |
| $5\sigma$–$6\sigma$ | 97.73 ± 0.18 | 100 | 100 |

In this case, the random errors, and the bad observations of the linear model are exactly the same as those of the univariate sample. The power of test and *MSR* are computed for intervals $3\sigma$–$4\sigma$, $4\sigma$–$5\sigma$ and $5\sigma$–$6\sigma$.

The power of test is computed by using Eq. (13) with $\alpha = 0.001$ for linear model. The power of test and *MSR* for one outlier and two outliers in linear model are given in Tables 5 and 6, respectively. The power of test and the *MSR* increase depending on the rising of outlier magnitude. Also, non-random outliers affect the capacity of the Baarda test badly. It decreases from 36.39% to 24.14%, despite that the power of test is the same as for random and non-random outliers.

Although, the mathematical models of the univariate and linear models are different, the values of the power of test of these models are the same. Also, to test the effect of the number of observations for *MSR* and power of test, the number of observations was changed to 11. The magnitude of the outliers and other criterias were the same as the

**Table 5.** The same as in Table 1, but for linear model.

| Outlier Interval | *MSR* [%] | Power of Test |
|---|---|---|
| $3\sigma$–$4\sigma$ | 69.97 ± 14.56 | 59.71 ± 1.08 |
| $4\sigma$–$5\sigma$ | 95.16 ± 6.94 | 88.52 ± 0.51 |
| $5\sigma$–$6\sigma$ | 98.29 ± 2.62 | 98.46 ± 0.10 |

**Table 6.** The same as in Table 2, but for linear model.

| Outlier Interval | Power of Test | Random Outliers | Non-Random Outliers |
|---|---|---|---|
| | | *MSR* [%] | *MSR* [%] |
| $3\sigma$–$4\sigma$ | 46.76 ± 1.70 | 36.39 ± 6.89 | 24.14 ± 4.37 |
| $4\sigma$–$5\sigma$ | 83.36 ± 0.93 | 83.72 ± 9.64 | 53.78 ± 4.18 |
| $5\sigma$–$6\sigma$ | 97.73 ± 0.18 | 89.81 ± 4.70 | 71.68 ± 3.54 |

**Table 7.** The same as in Table 5, but for 11 observations.

| Outlier Interval | *MSR* [%] | Power of Test |
|---|---|---|
| $3\sigma$–$4\sigma$ | 71.55 ± 12.72 | 59.71 ± 1.08 |
| $4\sigma$–$5\sigma$ | 98.04 ± 3.03 | 88.52 ± 0.51 |
| $5\sigma$–$6\sigma$ | 100 | 98.46 ± 0.10 |

**Table 8.** The same as in Table 6, but for 11 observations.

| Outlier Interval | Power of Test | Random Outliers | Non-Random Outliers |
|---|---|---|---|
| | | *MSR* [%] | *MSR* [%] |
| $3\sigma$–$4\sigma$ | 46.76 ± 1.70 | 43.05 ± 7.40 | 33.11 ± 4.60 |
| $4\sigma$–$5\sigma$ | 83.36 ± 0.93 | 92.33 ± 5.11 | 76.97 ± 3.52 |
| $5\sigma$–$6\sigma$ | 97.73 ± 0.18 | 97.36 ± 1.71 | 90.48 ± 2.63 |

analysis that their results were given in Tables 5 and 6, the same analyses were carried out and the results are shown in Tables 7 and 8.

The *MSR* changed depending on the changing of number of observations. At the Tables 7 and 8, the number of the observations was changed and the results of the Baarda test were changed, too; but the power of test was the same as Tables 5 and 6. It is not expected that the results of the outlier test are the same for the different cases for example, different number of observations. However, the power of test is the same. According to these results, *MSR* is the empirical value of the power of test and it reflects the more reliable results for the efficacy of the Baarda test. *MSR* gives us the more information in detail than the theoretical power of test. It is more realistic than the theoretical power of test in applying the different outlier models and multivariate models.

## 10. CONCLUSIONS

The power of test is used to measure the capacity of the test method to detect outlier in statistics. Although, the capacity of the test method depends on the number of outliers, the magnitude of outlier, the number of unknowns, the number of observations and the type of outliers, the values of power of test are the same for all different situations; because Eq. (13) does not have any variable of these factors that can be changeable for any model; but the *MSR* of the methods change for all different cases, which is the expected situation and to be considered situation. If we consider the detection and identification steps of the test methods, the theoretical power of test considers the only probability of the missed detection and accepts that the probability of the wrong identification is zero. However, *MSR* considers the both missed detection and wrong identification steps. Since the probabilitiy of the wrong identification is very small comparing with the probability of the missed detection, there are small differences between theoretical power of test and *MSR*. Since, *MSR* takes into account all different cases, it is the more realistic than theoretical power of test and it can be preferred for the empirical estimation of the power of test in outlier detection problem.

*References*

Atkinson A.C., 1985. *Plots, Transformations and Regression*. Oxford University Press, Oxford, U.K.

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

Baarda W., 1967. Statistical concepts in geodesy. *Publication on Geodesy, New Series*, **2(4)**, Netherlands Geodetic Commission, Delft, The Netherlands.

Baarda W., 1968. A testing procedure for use in geodetic Networks. *Publication on Geodesy, New Series*, **2(5)**, Netherlands Geodetic Commission, Delft, The Netherlands.

Chen Y.Q., Kavouras M. and Chrzanowski A., 1987. A strategy for detection of outlying observations in measurements of high precision. *The Canadian Surveyor*, **41**, 529–540.

Cook R.D. and Weisberg S., 1982. *Residuals and Influence in Regression*. Chapman & Hall, New York.

Donoho D.L. and Huber P.J., 1983. The notion of breakdown point. In: Bickel P.J., Doksum K. and Hodges J.L. Jr. (Eds), *A Festschrift for Erich Lehmann*. Wadsworth, Belmont, CA, 157–184.

Hadi A.S. and Simonoff J.S., 1993. Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc.*, **88**, 1264–1272.

Hampel F., Ronchetti E., Rousseeuw P. and Stahel W., 1986. *Robust Statistics: the Approach Based on Influence Functions.* John Wiley and Sons, New York.

Heck B., 1981. Der Einfluss einzelner Beobachtungen auf das Ergebnis einer und die Suche nach Ausreissern in den beobachtungen. *Allg. Verm. Nachricht.*, **88**, 17–34.

Hekimoglu S., 1997. The finite sample breakdown points of the conventional iterative outlier detection procedures. *J. Surv. Eng.*, **123**, 15–31.

Hekimoglu S. and Koch K.R., 1999. How can reliability of the robust methods be measured? In: Altan M.O. and Gründige L. (Eds), *Third Turkish-German Joint Geodetic Days*, Volume 1. Istanbul Technical University, Istanbul, Turkey, 179–196.

Hekimoglu S. and Koch K.R., 2000. How can reliability of the test for outliers be measured? *Allg. Verm. Nachricht.*, **107**, 247–254.

Hekimoglu S., Erdogan B., Erenoglu R.C. and Hosbas R.G., 2011. Increasing the efficacy of the tests for outliers for geodetic networks. *Acta Geod. Geophys. Hung.*, **46**, 291–308.

Hekimoglu S. and Erdogan B., 2012. New median approach to define configuration weakness of deformation networks. *J. Surv. Eng.*, **138**, 101–108.

Huber P.J., 1981. *Robust Statistics.* John Wiley and Sons., New York.

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

Koch K.R., 1999. *Parameter Estimation and Hypothesis Ttesting in Linear Models, 2nd Edition*. Springer-Verlag, Berlin, Germany.

Koch K.R., 2013a. Robust estimation by expectation maximization algorithm. *J. Geodesy*, **87**, 107–116.

Koch K.R., 2013b. Comparison of two robust estimations by expectation maximization algorithms with Huber's method and outlier tests. *J. Appl. Geodesy*, **7**, 115–123.

Kok J.J., 1984. *On Data Snooping and Multiple Outlier Testing.* NOAA Technical Report, **30**. U.S. Department of Commerce, Rockville, MD.

Kuang S., 1991. *Optimization and Design of Deformation Monitoring Schemes*. Ph.D. Thesis. Report 157. Department of Surveying Engineering, University of New Brunswick, Fredericton, NB, Canada.

Lehmann R., 2010. Normierte Verbesserungen - wie groß ist zu groß? *Allgemeine Vermessungsnachrichten*, **2**, 53–61 (in German.

Lehmann R., 2012a. Geodätische Fehlerrechnung mit der skalenkontaminierten Normalverteilung. *Allgemeine Vermessungsnachrichten*, 143–149 (in German).

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

Maronna R., Martin D. and Yohai V., 2006. *Robust Statistics.* John Wiley and Sons., New York.

Neyman J. and Pearson E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A-Math. Phys. Eng. Sci.*, **231**, 289–337.

Pope A.J., 1976. *The Statistics of Residuals and the Outlier Detection of Outliers.* NOAA Technical Report, **65**. U.S. Department of Commerce, Rockville, MD.

Teunissen P.J.G., 2000. *Testing Theory - an Introduction*. Delft University, Delft, The Netherlands.

Teunissen P.J.G., 2018. Distributional theory for the DIA method. *J. Geodesy*, **92**, 59–80.

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

Wang J. and Knight N.L., 2012. New outlier separability test and its application in GNSS positioning. *J. Glob. Posit. Syst.*, **11**, 46–57.

Yang L., Wang J., Knight N.L. and Shen Y., 2013. Outlier separability analysis with a multiple alternative hypotheses test. *J. Geodesy*, **87**, 591–604.

Zaminpardaz S. and Teunissen P.J.G., 2018. DIA - datasnooping and identifiability. *J. Geodesy*, DOI: 10.1007/s00190-018-1141-3 (in print).