Boletim de Ciências Geodésicas
ISSN: 1413-4853
bcg_editor@ufpr.br
Universidade Federal do Paraná
Brasil

Hekimoglu, Serif; Erdogan, Bahattin
APPLICATION OF MEDIAN-EQUATION APPROACH FOR OUTLIER DETECTION IN GEODETIC NETWORKS
Boletim de Ciências Geodésicas, vol. 19, núm. 4, octubre-diciembre, 2013, pp. 548-557
Universidade Federal do Paraná
Curitiba, Brasil

BCG - Boletim de Ciências Geodésicas - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702013000400002

APPLICATION OF MEDIAN-EQUATION APPROACH FOR OUTLIER DETECTION IN GEODETIC NETWORKS

Aplicação da equação mediana aproximada para a detecção de "outliers" nas redes geodésicas.

SERIF HEKIMOGLU
BAHATTIN ERDOGAN

Department of Geomatic Engineering
Yildiz Technical University, Istanbul, Turkey
E-mail: hekim@yildiz.edu.tr ; berdogan@yildiz.edu.tr

ABSTRACT

In geodetic measurements some outliers may occur sometimes in data sets, depending on different reasons. There are two main approaches to detect outliers as Tests for outliers (Baarda's and Pope's Tests) and robust methods (Danish method, Huber method etc.). These methods use the Least Squares Estimation (LSE). The outliers affect the LSE results, especially it smears the effects of the outliers on the good observations and sometimes wrong results may be obtained. To avoid these effects, a method that does not use LSE should be preferred. The median is a high breakdown point estimator and if it is applied for the outlier detection, reliable results can be obtained. In this study, a robust method which uses median with 3σ or $3\sigma_{med}$ as a treshould value on median residuals that are obtained from median equations is proposed. If the a priori variance of the observations is known, the reliability of the new approch is greater than the one in the case where the a priori variance is unknown.

Keywords: Median; Median Absolute Deviation; Outlier; Median Equations; Decision Matrix; Leveling Network.

RESUMO

Em medidas geodésicas, alguns erros grosseiros (*outliers*) podem ocorrer em conjuntos de dados, devido a diferents rezões. Existem dois principais métodos para a detecção de erros grosseiros como os testes de Baarda e Pope's para outliers e os métodos robustos Danish e Huber. Estes métodos, usam estimativas de minimos quadrados (LSE). Nestes casos, as observações com erros grosseiros podem afetar

os resultados por contaminarem observações que são boas pelo efeito de espalhamewnto imposto pela LSE e às vezes resultados errados podem ser obtidos. Para evitar estes efeitos, um método que não usa LSE deve ser escolhido. A mediana é um estimados de pontos de grande valia e se é aplicado a um detector de erros grosseiros, resultados confiáveis podem ser obtidos. Nesta pesquisa, um método robusto que usa a mediana com 3σ ou $3\sigma_{\text{med}}$ é um valor que limita os resíduos medianos que podem ser obtidos a partir de equações medianas propostas. Se uma variância a priori das observações é conhecida, a confiabilidade do novo método é maior do que aquele, uja variância a prior é desconhecida.
Palavras-chave: Mediana; Derivação da Mediana Absoluta; Outlier; Equação Mediana; Matriz de decisão; Rede de Nivelamento.

1. INTRODUCTION

The least squares estimation (LSE) is very sensitive against deviations of the model assumptions (HAMPEL et al. 1986). The LSE spreads the effect of the outliers on the residuals of the good observations which do not have any outlier (HEKIMOGLU et al. 2011a). There are two main reasons for the wrong results of outlier detection methods as spreading effect of the LSE and weakness of configuration of the given geodetic network (HEKIMOGLU et al. 2011b). It is showed that an outlier in the observations of a geodetic network can not be identified reliably by using any method due to the configuration weakness in the network. The outlier affects badly the residual from LSE of another observation that lies close to this bad observation, due to the deficiency of the configuration of the network.

Generally, statistical procedures for detecting outliers work well in practice only in case of one single outlier, but can fail in case of multiple outliers (BAARDA 1968, POPE 1976, HEKIMOGLU and KOCH 1999 and 2000, XU 2005). In addition, Baselga (2007) showed that only one outlier in a geodetic network may be detected by using Test for outliers (Pope's test) when the a priori variance of the observations is not known. In case of more than one gross error, the test becomes inefficient. Moreover, even the sample includes single outlier; the test may fail when observations are correlated. Tests for outliers are based on the assumption of a (possible) single outlier, and frequently with unjustified hope that they are supposed to be successful if multiple outliers appear. It is impossible to detect multiple outliers without additional hypotheses. Also, the single outlier hypothsis is also proven as being sufficient except to the case where the degree of freedom is one. These results of Baselga (2007 and 2011) verify the ones of the mentioned above properties.

The median has a highest breakdown point such as 50%. It means that for one dimensional case (i.e. the population may be defined by a random variable) the median can isolate the good observations with the rate of 51% from the bad observations of any kind with the rate of 49%. Median is used for estimating the

location parameter (μ). However, the median is not as efficient as the mean, i.e. the standard deviation of the median is greater than the one of the mean at Gaussian distribution (MARONNA et al. 2006, p.20). Youcai (1995) applied the median and MAD (Median Absolute Deviation) to the triangulation network by identifying outliers under some criterions such as $|r_i|>2\sigma_{med}$ or $3\sigma_{med}$ where $\sigma_{med}=1.4826$ MAD and r_i are the differences between the coordinates and the median of them. The coordinates of the new points are calculated by taking all the possible combinations of observations (angles), and then the median applied to these coordinates' values. Duchnowski (2010) applied R-Estimator to idendify unstable reference marks where the median and MAD play a main role.

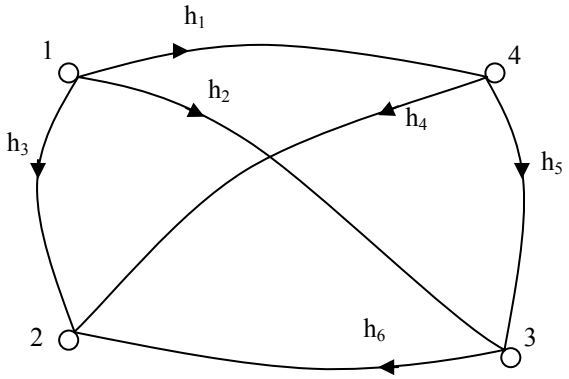
Hekimoglu et al. (2011b) proved that the reasons for the failures in the outlier detection depend on not only due to the ability of the outlier detection method, but also mostly due to the weakness in the configuration of the networks. To detect the probable configuration weakness the median equations were used. In this paper, a new robust approach based on the median and MAD on the observations of the geodetic networks is introduced. To apply the median estimator for outlier detection the median equations are used. The main idea is to do outlier detection in geodetic networks that is based on observations without using LSE.

2. MEDIAN EQUATION APPROACH FOR LEVELING NETWORK

The geodetic networks are established to realize two main topics: estimating the coordinates of the new points optimally based on the coordinates of datum points, and controlling the reliability of the network whether the observations include outliers or not.

The height differences are measured for the leveling network. A minumum configuration that resists against one outlier is given in Fig. 1 (HEKIMOGLU et al. 2011b).

Figure 1 - A leveling network that has a minumum configuration that resists against one outlier.



The following median equations can be written where each observation must appear once in these equations (HEKIMOGLU et al. 2011b):

$$h_1^{(1)} = h_1$$

$$h_1^{(2)} = h_3 - h_4 \tag{1a}$$

$$h_1^{(3)} = h_2 - h_5$$

To clear this, an extra equation can be written such as:

$$h_1^{(4)} = h_3 + h_5 + h_6$$

h_3 and h_5 apper two times in these four equations. If h_3 or h_5 has an outlier, the median can not separate them from the good observations. Therefore, the last equation can not be considered as a median equation. The similar median equations for the other observations can be written as follows:

$$\begin{array}{ccccc} h_2^{(1)} = h_2 & h_3^{(1)} = h_3 & h_4^{(1)} = h_4 & h_5^{(1)} = h_5 & h_6^{(1)} = h_6\,, \\ h_2^{(2)} = h_1 + h_5 & h_3^{(2)} = h_2 + h_6 & h_4^{(2)} = h_3 - h_1 & h_5^{(2)} = h_4 - h_6 & h_6^{(2)} = h_4 - h_5 \\ h_2^{(3)} = h_3 - h_6 & h_3^{(3)} = h_1 + h_4 & h_4^{(3)} = h_5 + h_6 & h_5^{(3)} = h_2 - h_1 & h_6^{(3)} = h_3 - h_2 \end{array} \tag{1b}$$

The median of these equations can be estimated such as:

$$Med_i = median(h_i^{(1)}, h_i^{(2)}, h_i^{(3)}), \quad \text{i=1,2,..,6} \tag{2}$$

The median can separate only one outlier in these three median equations. For example, let h_1 include an outlier. Med_1 can separate it from two other median equations. It can not affect Med_1 and also Med_2, Med_3,…, Med_6. Consequently, the median can separate only one outlier in the observations of this leveling network. For example, if h_1 and h_5 were contaminated, the median can not separate these two outliers.

Now, the question is arised how can this outlier be detected when the median is used as an estimator. If the variance σ^2 of the population is known before, then the outlier may be detected by using the 3σ-rule (KUTTERER et al. 2003, LOON 2008).

Let h_2 contaminated by outlier Δ. $h_1^{(3)}$, $h_2^{(1)}$, $h_3^{(2)}$, $h_5^{(3)}$ and $h_6^{(3)}$ are damaged. They include the outlier Δ and the random error ε due to the second observation together except $h_2^{(1)}$that includes only Δ. $\Delta+\varepsilon > 3\sigma$ when both Δ and ε have the

same sign (i.e. both + or both -) or $\Delta+\varepsilon < 3\sigma$ when both of them have the opposite sign (i.e. + and – or – and +). Therefore, more flagging values than the one may exceed the threshold value of 3σ. How can we decide which one of them is the true outlier? If the median equations of $h_1^{(3)}$, $h_2^{(1)}$, $h_3^{(2)}$, $h_5^{(3)}$ and $h_6^{(3)}$ are considered, it is seen that they all include the common value of h_2. Hence, the true bad observation must be the value of h_2. Thus, if observations have only one outlier and more candidates of outlier are detected, true outlier can be found.

Let's look at the leveling network given in Fig.1. We can obtain r_{ij} instead of observations by using the Eq.(3):

$$r_{ij} = Med_i - h_i^j, i = 1,2,\dots,6 \; and \; j = 1,2,3, \tag{3}$$

where r_{ij} are defined here as "median residuals". They are considered here as measuruments.

When h_2 is contaminated by an outlier $h_1^{(3)}$, $h_2^{(1)}$, $h_3^{(2)}$, $h_5^{(3)}$ and $h_6^{(3)}$ are contaminated. As a result, 5 median equations of 18 are ruined by h_2. The median of r_{ij} is not affected by these five contaminated values. The main question is that how we can detect the outlier in h_2. According to the median equations we can form a matrix which is called here "decision matrix" given Table 1. In the decision matrix, "0" means that there is no relation between observations and "1" means that these equations form one median equation (For example, in the second line of the matrix, r_{12} is formed by h_3 and h_4). When h_2 is contaminated five median residuals (r_{13}, r_{21}, r_{32}, r_{53} and r_{63}) are contaminated, too. According to the contaminated median equations, for r_{13} h_2 and h_5; for r_{21} h_2; for r_{32} h_2 and h_6; for r_{53} h_2 and h_1; for r_{63} h_2 and h_3 can be flagged as outliers. The flagged means that this observation is set candidate for contaminated observation. If the total flagged numbers is estimated, the number of h_2 is five times, and other observations are only once, so that the outlier can be detected considering decision matrix and total flagged number (k). If the total flagged number is bigger than one (k>1) this observation includes an outlier.

If the a priori variance σ^2 is not known, σ_{med} is proposed instead of MAD because σ_{med} is more efficient than MAD (HAMPEL et al. 1986, ROUSSEEUW and LEROY 1987, MARONNA et al. 2006):

$$\sigma_{med} = 1.4826 median\{|med_i - h_i^j|\}, i = 1,2,\dots,6 \; and \; j = 1,2,3 \tag{4}$$

Table 1 - The decision matrix.

r_{ij}	h_1	h_2	h_3	h_4	h_5	h_6
r_{11}	1	0	0	0	0	0
r_{12}	0	0	1	1	0	0
r_{13}	0	1	0	0	1	0
r_{21}	0	1	0	0	0	0
r_{22}	1	0	0	0	1	0
r_{23}	0	0	1	0	0	1
r_{31}	0	0	1	0	0	0
r_{32}	0	1	0	0	0	1
r_{33}	1	0	0	1	0	0
r_{41}	0	0	0	1	0	0
r_{42}	1	0	1	0	0	0
r_{43}	0	0	0	0	1	1
r_{51}	0	0	0	0	1	0
r_{52}	0	0	0	1	0	1
r_{53}	1	1	0	0	0	0
r_{61}	0	0	0	0	0	1
r_{62}	0	0	0	1	1	0
r_{63}	0	1	1	0	0	0

3. SIMULATION OF THE LEVELING NETWORK

To apply the new median approch, the network given in Fig.1 is considered. The heights of four points are H_1=100.000 m, H_2=105.276 m, H_3=104.388 m and H_4=103.055 m respectively. They are not affected from random errors. The height differences (h_{oi}, i=1,2,..,6) are computed. To obtain the measurements of the height differences we assume that the random measurement errors have the same variance σ^2 (i.e. σ=1 mm). Thus, the measurements of the height differences h_i are computed as

$$h_i = h_i + e_i, i = 1,2, \dots ,6 \tag{6}$$

where $e_i \sim N(\mu=0, \sigma^2=1)$, it is assumed that the measurement errors are Gaussian with the expected value which is equal to zero and the variance which is equal to 1. Here e_i are 0.90, -0.52, -0.50, -0.97, 0.00, 1.39 mm

To generate one contaminated height value $\bar{h}_i$, the random error e_i is replaced by the outlier dh_i as follows:

$$\bar{h}_i = h_i + dh_i, i = 1,2, \dots ,6 \tag{7}$$

In this section we have tested the following cases:

I. The observations do not include any outlier.
II. The observation (h_1) is contaminated with +5 mm magnitude.
III. The observation (h_1) is contaminated with +10 mm magnitude.
IV. The observation (h_1) is contaminated with +1000 mm magnitude which is called as a wild observation.

For the first case: The median equations for each height difference h_i are constituted and their medians are taken according to the equations of (1a), (1b) and (2). Then, the differences (r_{ij}) according to (3) are computed. σ_{med} of them is found as 1.0 mm. The method did not detect any outlier which is greater than 3σ for the case when the a priori variance is known, and also $3\sigma_{med}$ for the case when the a priori variance is unknown.

For the second case: The median residuals (r_{ij}) according to Eq. (3) are given as [**-4.5**, 0.0, 0.9, 0.0, **-5.5**, 1.4, 1.4, 0.0, **-3.2**, 0.0, **4.5**, -2.3, -2.4, 0.0, **3.2**, -1.4, 0.9, 0.0] mm. If we look at r_{ij}-values, we can see that the outlier is not spreaded on the adjacent observations as in LSE. The σ_{med} of them is 2.0 mm. The threshold value for r_{ij} is 3 mm for 3σ. We see that there are five values (r_{11}, r_{22}, r_{33}, r_{42} and r_{53}) greater than 3σ. These median equations are contaminated by h_1. The height difference h_1 is the joint value among these five contaminated values. In the decision matrix h_1 is flagged five times and h_2, h_3, h_4, h_5 and h_6 are flagged only once, since the flagged number of the h_1 is greater than one, h_1 is the outlier. If the a priori variance is unknown, $3\sigma_{med}$ is used as a threshold value. Since none of the median residuals exceed the $3\sigma_{med}$ the method can not detect the outlier.

For the third case: The differences (r_{ij}) according to Eq. (3) are given as [**-9.5**, 0.0, 1.0, 0.0, **-10.5**, 1.4, 1.4, 0.0, **-8.2**, 0.0, **9.5**, -2.4, -2.4, 0.0, **8.2**, -1.4, 1.0, 0.0] mm. The σ_{med} of them is 2.0 mm again. If the a priori variance σ^2 of the height differences is known, 3σ is used as threshold value that is the same as in second case. We see that there are five values (r_{11}, r_{22}, r_{33}, r_{42} and r_{53}) that are greater than 3σ. h_1 is the joint observation. In the decision matrix h_1 is flagged five times and h_2, h_3, h_4, h_5 and h_6 are flagged only once, since the flagged number of the h_1 is greater than one h_1 is the outlier. If the variance σ^2 of the height differences is unknown, $3\sigma_{med}$ is used as a threshold values that are the same as in second case. We see that there are five values (r_{11}, r_{22}, r_{33}, r_{42} and r_{53}) that are greater than $3\sigma_{med}$. If we look at these median equations, there is one joint value i.e. h_1 among them. Therefore, h_1 must be contaminated.

For the fourth case: The differences (r_{ij}) according to (3) are given as [**-999.5**, 0.0, 1.0, 0.0, **-1000.5**, 1.4, 1.4, 0.0, **-998.2**, 0.0, **999.5**, -2.4, -2.4, 0.0, **998.2**, -1.3, 1.0, 0.0] mm. If we look at r_{ij} values, we can see that the outlier is not spreaded on the adjacent observations as in LSE. The σ_{med} of them is 2.0 mm again. There are five values (r_{11}, r_{22}, r_{33}, r_{42} and r_{53}) that are greater than 3σ. If we see these median equations, there is one joint value h_1. Thus, we can detect this outlier in h_1. If the a priori variance σ^2 of the height differences is unknown, $3\sigma_{med}$ is used as a threshold value where they are the same as in the second case. We see that there are five

values (r_{11}, r_{22}, r_{33}, r_{42} and r_{53}) that are greater than $3\sigma_{med}$. If we see these median equations, there is one joint value h_1. Therefore, h_1 must be contaminated.

We see that the σ_{med} value for the cases II, III and IV does not change as the magnitude of outlier changes. This is the proof for the property of robustness.

4. SIMULATION RESULTS

We know that the success of the robust methods and Tests for outliers are changed from one sample to the other one where the random errors are different (HEKIMOGLU and KOCH 1999, HEKIMOGLU and KOCH 2000). Therefore, the success of a method used cannot be evaluated by the result from one sample which may be chosen subjectively.

For simulation the network given in Fig.1 was considered. The random errors, 6 mesurements and outlier are generated as done in above section. They come from a Gausian distribution such as N(μ, σ^2=1mm^2). A hundered random error vectors **e** and also a hundered good sample are generated. In addition, each sample is contaminated only by one outlier 100 times. Thus, we have obtained 10 000 contaminated samples.

r_{ij} are analysed by using median and threshold value as 3σ or $3\sigma_{med}$. The results are given in Table 2. To measure the capacity of a method, the mean success rate (MSR) (HEKIMOGLU and KOCH 1999, HEKIMOGLU and KOCH 2000, HEKIMOGLU and ERENOGLU 2007, ERENOGLU and HEKIMOGLU 2010) is used.

Table 2 - MSRs and standard deviations of the new median equation approach.

Method	Median with 3σ		Median with $3\sigma_{med}$	
Number of outlier	3σ-6σ (%)	6σ-12σ (%)	3σ-6σ (%)	6σ-12σ (%)
0	90.0		68.0	
1	67.0±32.1	76.7±32.2	44.5±33.9	84.3±23.9

Considering the algorithm of forming the median equations, i.e. the decision matrix which gives us how the height differences in the median equations are connected, we can detect outlier when the repeat number k of the flagged outliers is greater than 1. If the median and median equations are used for outlier detection the reliability of the method in condition that a priori variance is known is 67% where the magnitude of an outlier lies between 3σ and 6σ. If the a priori variance is unknown the reliability of the method decreases to 44.5%. The method may detect the good observations as outliers for some cases.

5. CONCLUSION

In this study, it is investigated that whether median (with 3σ or $3\sigma_{med}$) may be used as an estimator on the median residuals r_{ij} or not to detect outliers by

considering the algorithm of forming the median equatios, i.e. the decision matrix which gives us how the height differences in the median equations are connected. The flagged number of the observation is obtained from the decision matrix. It can be seen how many times a measurement is flagged in this decision matrix. The outlier can be detected when the flagged number is greater than 1. According the results of the leveling network, the median method can be used especially when the a priori variance σ^2 of the observations is known. If it is not known, σ_{med} should be used. If σ_{med} is used the MSRs decreases by comparing the ones of the case where the a priori variance σ^2 is known.

REFERENCES

BAARDA W. (1968), "A testing procedure for use in geodetic Networks", Publication on Geodesy, *New Series* 2, no.5., Netherlands Geodetic Commission, Delft.

BASELGA S. (2007), "Critical limitation in use of τ test for gross error detection". *J. Surv. Eng*. 133(2): 52–55

BASELGA S. (2011), "Non Existence of Rigorous Tests for Multiple Outlier Detection in Least Squares Adjustment", *J. Surv. Eng*., 137(3): 109-112

DUCHNOWSKI R. (2010), "Median-Based Estimates and Their Application in Controlling Reference Mark Stability", *J. of Surv. Engn*.-ASCE, Vol.136(2):47-52

ERENOGLU R. C., HEKIMOGLU S., (2010), "Efficiency of robust methods and tests for outliers for geodetic adjustment models, *Acta.Geod. Geoph. Hung*. Vol. 45(4): 426-439

HAMPEL F., RONCHETTI E., ROUSSEEUW P., STAHEL W. (1986), *"Robust statistics: the approach based on influence functions*", John Wiley and Sons, New York

HEKIMOGLU S., KOCH K. R. (1999), "How can reliability of the robust methods be measured?", *in Third Turkish German Joint Geodetic Days, Altan and Gründing* (Eds), 1-4 June,Istanbul, Turkey, 179-196.

HEKIMOGLU S., KOCH K. R. (2000) "How can reliability of the test for outliers be measured?, *Allg. Vermes. Nachr*., 107(7): 247-254,

HEKIMOGLU S., ERENOGLU R. C. (2007), "Effect of heteroscedasticity and heterogeousness on outlier detection for geodetic networks, *J.Geodesy*, 81(2): 137-148

HEKIMOGLU S., ERDOGAN B., ERENOGLU R. C., HOSBAS R. G. (2011a), "Increasing the reliability of the tests for outliers for geodetic netwowks", *Survey Review* ,Vol. 46(3):291-308

HEKIMOGLU S., ERENOGLU R. C., SANLI D. U., ERDOGAN B., (2011b), "Detecting Configuration Weaknesses in Geodetic Networks", *Survey Review*, 43 (323): 713-730.

KUTTERER H., HEINKELMANN R., AND TESMER V. (2003), "Robust Outlier Detection in VLBI Data Analysis" *Proceedings of the 16th Working Meeting*

on European VLBI for Geodesy and Astrometry, W. Schwegmann and V. Thorandt, eds., Bundesamt für Kartographie und Geodäsie, Leipzig/Frankfurt, 247-255.

LOON J. P. (2008), “Functional and stochastic modelling of satellite gravity data, *Publications on Geodesy 67*, Netherlands Geodetic Commission, Delft.

MARONNA R., MARTIN D., YOHAI V. (2006), "*Robust Statistics*", Wiley, New York.

POPE A. J. (1976), "The statistics of residuals and the outlier detection of outliers" *NOAA Technical Report*, NOS 65, NGS 1, Rockville, MD.

ROUSSEEUW P. J., LEROY A. M. (1987), "*Robust regression and outlier detection"*, John Wiley and Sons, Inc., New York.

XU P. L. (2005), “Sign-constrained Robust Least Squares, Subjective Breakdown Point and the Effect of Weights of Observations on Robustness”, *Journal of Geodesy* 79:146-159

YOUCAI H. (1995), "On the design of estimators with high breakdown points for outlier identification in triangulation Networks", *Bull. Geod*. 69:292 – 299

(Recebido em maio de 2013. Aceito em agosto de 2013).