A review of outlier detection procedures used in Surveying Engineering

Mevlut Yetkin[1]

[1]*Department of Geomatics Engineering, Izmir Katip Celebi University, Turkey*

Abstract

The method of least squares is the most widely used parameter estimation tool in surveying engineering. It is implemented by minimizing the sum of squares of weighted residuals. The good attribute of the method of least squares is that it can give an unbiased and minimum variance estimate. Moreover, if the observation errors are normally distributed identical results to the maximum likelihood method can be obtained. However, the method of least squares requires gross error and systematic bias free observations to provide optimal results. Unfortunately, these undesired errors are often encountered in practice. Therefore, outlier diagnosis is an important issue in spatial data analysis. There are two different approaches to deal with outliers: statistical outlier test methods and robust estimation. Baarda and Pope methods are well known hypothetical testing methods. On the other hand, there are numerous robust methods to eliminate or reduce disruptive effects of outliers, such as M-estimation method, L1 norm minimization, the least median squares and the least trimmed squares. Robust methods are useful to locate multiple outliers. Yet, statistical testing approach can also be generalized to multiple outliers. Furthermore, reliability measures and robustness analysis enable us to assess the quality of our networks in terms of gross error detection and the effect of undetected errors. In this study, a review of outlier detection procedures is given. The main features of the methods are summarized. Finally, statistical test for multiple outliers is applied to a GPS network.

Introduction

Surveying networks are used in many geomatics engineering projects to provide coordinate and/or height information. In a surveying network, geodetic observations (height differences, distances, angles, directions and GPS baseline components) are made and then parameter estimation is realized using the method of least squares. The method of least squares needs blunder free observations to be able to produce identical results with the maximum likelihood estimate [11-15]. However, observations are often burdened with systematic or gross errors in addition to inevitable random errors. Therefore, errors must be carefully analyzed before and after least squares adjustment calculus [8].

Outlying observations can be determined by statistically examining of residuals. Baarda's method that is applied at two stages (global test and data snooping) is extensively used by geomatics society for this purpose [8]. Outlier testing can be applied for both single outlier case and multiple outliers case [6]. On the other hand, robust statistics includes numerous procedures such as iteratively reweighted least squares [5], L1 norm minimization [11-15] and sign constrained robust least squares [12-16].

In addition to outlier detection and elimination techniques, the effects of undetected errors on estimated quantities such as point positions should also be investigated. Traditionally, reliability measures are computed for this purpose [1]. Additionally, robustness analysis that is a combination of reliability and strain can be used to portray the effects of undetected errors (blunders or systematic errors) [2-17].

Parameter estimation in surveying networks

In surveying networks such as GPS networks, parameter estimation is based on Gauss-Markov model. Outlier diagnosis (for example statistical testing procedure) is also performed inspecting residuals obtained from this adjustment model [8].

Parameter estimation is an optimization problem and generally includes minimization of a particular mathematical function of residuals. The function to be minimized is also called objective function. This function defines the properties of the estimation process. Table 1 shows some widely used estimation methods. The last four methods are known as robust statistical techniques and can be useful in surveying practices.

Table 3 Parameter Estimation Methods

Method	Objective Function
Least Squares [4]	$\sum_{i=1}^{n} v_i^2 \rightarrow min.$
Least Trimmed Squares [10]	$\sum_{i=1}^{u} P_i v_i^2 \rightarrow min.$
Least Median of Squares [9-10]	median(v_i^2) → *min.*
L1 Norm Minimization [11-15]	$p^T \lvert v \rvert \rightarrow min.$
Sign-Constrained Robust Least Squares [12-16]	$v^T \bar{P} v / r_{\bar{P}} \rightarrow min.$ subject to $\sum_{i=1}^{n} sign(v_i) = 0$

In Table 1, v denotes the observational residual, n is the number of observation, u is the number of residuals included in the summation $(u \leq n)$, P_i is the weight of ith observation, **p** is a vector that includes the diagonal elements of the **P** weight matrix, $\bar{\mathbf{P}}$ is the equivalent weight matrix that is computed using the weight function of a certain robust method such as Huber's M-Estimate (see [13]),$r_{\bar{\mathbf{P}}}$ is the rank of the $\bar{\mathbf{P}}$ matrix.

Hypothesis tests for outlier detection

Hypothetical testing procedure is carried out at two phases: the global model test and the outlier test. The global model test is applied to detect whether the observations and the functional and stochastic models are consistent. On the other hand, the outlier test is used to identify outlying observations. It can be said that there are two main statistical testing method used in surveying networks: Baarda and Pope. The a priori variance factor should be known in order to be able to apply the Baarda method. However, Pope method can be used using the posterior variance factor that is obtained from the least squares adjustment [8]. On the other hand, [7] generalized the Baarda method to multiple outliers.

In this paper, only the computation of the outlier test statistic (both single and multiple outlier case) will be elaborated. The outlier test statistic that follows a non-central chi-squared distribution is given as

$$w^2 = \frac{\boldsymbol{l}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\mathrm{v}}\mathbf{P}\mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\mathrm{v}}\mathbf{P}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\mathrm{v}}\mathbf{P}\boldsymbol{l}}{\sigma_0^2} \sim \chi^2_{1-\alpha_{w^2},\theta} \qquad (1)$$

where $\boldsymbol{l}$ is the vector of n observations, $\mathbf{P}$ is the weight matrix, $\mathbf{Q}_{\mathrm{v}}$ is the cofactor matrix of the residuals. It can be computed as $\mathbf{Q}_{\mathrm{v}} = \mathbf{P}^{-1} - \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$. θ is the number of outliers, $\mathbf{H}$ is an $n \times \theta$ matrix containing zeros with a one in each column corresponding to an outlier, σ_0^2 is the a-priori variance factor, α_{w^2} is the significance level for the outlier test. If we assumed that there are θ outliers, we can form $\binom{n}{\theta}$ combinations of the **H** matrix. Thus, $\binom{n}{\theta}$ w^2 test statistics can be computed in a surveying network. If the test statistic of a certain observation group exceeds the critical value at a selected significance level, it is suspected that relevant observations may be burdened with blunders. Naturally, the test statistic corresponding to outlying observations takes greater values [6].

Reliability and robustness

Outliers cannot always be determined at all times using mentioned methods in this paper. Therefore, a surveyor should be aware of the effects of undetected blunders. The effects of undetected blunders on the network results may be important. At this point, the reliability and robustness concepts are considered. Reliability is defined as the ability of a network to sense and control the blunders [1]. However, robustness means to insensitivity to outliers. It is a fusion of reliability and strain. It uses an analogy, i.e. some disturbing effects such as earthquakes or wind cause a deformation on structures or earth crust, similarly the undetected blunders are disturbing effects for a surveying network and the (virtual) deformation caused by them can be interpreted using strain concept as done in a building or earth crust. Robustness analysis method has been studied in [2-14-17].

Numerical results

In order to test the efficiency of the multiple outlier tests, a small GPS network was used. The design matrix **A**, the weight matrix **P** and the observation vector **l** of the network can be found in [4]. In the computations, the datum of the network was provided by minimum constraints, i.e. Point A was selected as a fixed station.

Some observations were contaminated by blunders. The **H** matrix was constructed according to $\theta = 2$ and $\theta = 3$. Then, the outlier test statistics in Eq. 1 were computed. Some of the obtained outlier test statistics were indicated in Table 2. $\binom{39}{2} = 741$ and $\binom{39}{3} = 9139$ w^2

statistics were computed for $\theta = 2$ and for $\theta = 3$, respectively. Since, there were numerous numerical values, only some selected values were indicated.

Table 4 Outlier testing for multiple outliers

Observation No	[illegible]	Observation No	[illegible]
1-2	5813.81	1-2-3	116.30
2-10	3782.59	1-2-10	45.95
1-3	1978.63	1-3-8	76.60
1-4	1981.85	2-3-5	108.96
1-5	1997.07	1-3-20	79.31
3-6	3.05	4-5-6	6.98
4-22	88.05	7-8-9	2.94
3-7	65.23	37-38-39	0.94
38-39	66.23	10-11-12	0.54
20-30	110.32	24-25-35	2.97

[a] Two outliers, of 1.5 m in observation 1 and -2 m in observation 2

[b] Three outliers, of 0.1 m in observation 1, -0.2 m in observation 2 and 0.3 m in observation 3

As seen from Table 2, the test statistics corresponding to contaminated observations are greater than those corresponding to good observations (not contaminated by blunders). The observation combinations that have greater outlier test statistics contain outlying observations with a high probability. Therefore, one should suspect them firstly. The numerical example presented in this paper shows that the multiple outlier testing procedure can be used in GPS networks for a reliable outlier diagnosis.

Conclusion

The classical least squares adjustment method is vulnerable against blunders. Statistical tests or robust estimation methods may be utilized in order to prevent the disruptive effects of blunders on network results. However, using the statistical tests each observation is tested one by one. On the other hand, surveying networks may contain multiple outliers and as the number of outliers increases the success of these methods decreases. This fact is well known by geomatics community. In this paper, we applied the outlier test for multiple outliers to a GPS network. Our results showed that this approach can be used in daily GPS surveying practice as a reliable outlier diagnosis technique.

Acknowledgments

The author wishes to acknowledge the support by the Scientific Research Projects Coordinator ship of Izmir Katip Celebi University.

References

[120] Baarda, W. (1968) A Testing Procedure for Use in Geodetic Networks. *Publications on Geodesy, New Series, Vol. 2, No. 5*, Delft, Netherlands Geodetic Commission.

[121] Berber, M. (2006) Robustness Analysis of Geodetic Networks. Technical Report No.242, *Department of Geodesy and Geomatics Engineering, UNB*, Fredericton, NB, Canada.

[122] Berber, M. (2008) Error Analysis of Geodetic Networks. *Shaker-Publishing BV*, Maastricht, Netherlands.

[123] Ghilani, D.G. and Wolf, P.R. (2006) Adjustment Computations-Spatial Data Analysis, Fourth Edition. *Wiley*, New Jersey.

[124] Hekimoğlu, S. and Berber, M. (2003) Effectiveness of robust methods in heterogeneous linear models. *Journal of Geodesy*, 76, 706-713.

[125] Knight, N.L. Wang, J. and Rizos, C. (2010) Generalized measures of reliability for multiple outliers. *Journal of Geodesy*, 84, 625-635.

[126] Kok, J.J. (1984) On data snooping and multiple outlier testing. *NOAA Technical Report, NOS NGS.30*, US Department of Commerce, Rockville, Maryland.

[127] Kuang, S.L. (1996) Geodetic Network Analysis and Optimal Design. *Ann Arbor Press*, Ann Arbor, MI.

[128] Rousseeuw, P.J. (1984) Least Median of Squares Regression. *J. of the American Statistical Association*, 79(388), 871-880.

[129] Rousseeuw, P.J. and Leroy, A.M. (2003) Robust Regression and Outlier Detection. *Wiley*, New Jersey.

[130] Simkooei, A.A. (2003) Formulation of L1 norm minimization in Gauss-Markov models. *J. Surveying Engrg., ASCE*, 129(1), 37-43.

[131] Xu, P. (2005) Sign-constrained robust least squares, subjective breakdown point and the effects of weights of observations on robustness. *Journal of Geodesy*, 79(1-3), 146-159.

[132] Yang, Y., Song, L. and Xu, T. (2002) Robust estimator for the correlated observations based on bifactor equivalent weights. *Journal of Geodesy*, 76, 353-358.

[133] Yetkin, M. (2012) GNSS Gözlemlerinin Robust Kestirim ve Robustluk Analizi Yöntemleriyle Değerlendirilmesi üzerine Bir İnceleme. PhD Dissertation, *Department of Geomatics Engineering, The Graduate School of Natural and Applied Science of Selcuk University*, Konya, Turkey. (in Turkish)

[134] Yetkin, M. and Inal, C. (2011) L1 norm minimization of GPS networks. *Survey Review*, 43(323), 523-532.

[135] Yetkin, M. and Berber, M. (2013a) Application of the sign-constrained robust least squares method to surveying networks. *J. Surveying Engrg., ASCE*, 139(1), 59-65.

[136] Yetkin, M. and Berber, M. (2013b) Robustness analysis using the measure of external reliability for multiple outliers. *Survey Review*, DOI 10.1179/1752270612Y.0000000026.