International Journal of Engineering and Geosciences
https://dergipark.org.tr/en/pub/ijeg
e-ISSN 2548-0960

New approaches for outlier detection: The least trimmed squares adjustment

Hasan Dilmaç [*1], **Yasemin Şişman** [1]

[1]*Ondokuz Mayıs University, Department of Geomatics Engineering, Türkiye*

Keywords
The Least Square
Outliers
Robust Estimation
The Least Trimmed Squares

Research Article
DOI: 10.26833/ijeg.996340

Received: 18.09.2021
Accepted: 19.11.2021
Published: 13.04.2022

Abstract

Classical outlier tests based on the least-squares (LS) have significant disadvantages in some situations. The adjustment computation and classical outlier tests deteriorate when observations include outliers. The robust techniques that are not sensitive to outliers have been developed to detect the outliers. Several methods use robust techniques such as M-estimators, L_1- norm, the least trimmed squares etc. The least trimmed squares (LTS) among them have a high-breakdown point. After the theoretical explanation, the adjustment computation has been carried out in this study based on the least squares (LS) and the least trimmed squares (LTS). A certain polynomial with arbitrary values has been used for applications. In this way, the performances of these techniques have been investigated.

1. Introduction

In geodesy, for different purposes, many kinds of observations are done. Physical and geometric quantities such as angles, distances, heights, and gravity are measured and processed. In this case, a great number of data appears [1]. A quantity is always measured differently even though it is measured many times under the same conditions [2]. It is clear that observations are never equal to real value and they always contain an error. Thus, it is preferred that the observation number is bigger than the unknown number. In this case, the optimum solution must be made for a unique solution according to an aim function. This process is named adjustment computation [3-4].

The least-square (LS) is a conventional method for adjusting surveying measurements. It is one of the most adopted methods because of tradition and ease of computation [5-6]. But, outliers (observations with different distribution compared to the distribution of majority) negatively affect the LS method results [7].

Outliers in the observation group are encountered very often in applications [6]. The adjustment results with classical methods, which should meet some conditions like the normal distribution, are deteriorated. So, these outliers must be detected and eliminated from the observation group. Outlier tests are based on classical methods like Data-Snooping, Pope test and t-test [8-10]. These outlier tests are not robust against outliers. Therefore, new statistical methods have been sought instead of LS, which is sensitive to outliers [11].

The robust statistics deals with developing estimators insensitive to inconsistencies from basic assumptions in classical models [12]. Robust methods aim to find the results that would be found without outliers in the LS method to overcome the effects of outliers. Then, outliers can be detected through their residuals [13]. Many robust techniques have been developed. These techniques can be divided into classes by concepts of high-breakdown point, influence function, etc.

There are a lot of studies done on robust statistics. Some of them are assocaited with L_1-norm and M-estimators and some are associated with the other robust techniques [14-27].

The least trimmed squares (LTS) that is going to be emphasized in this study is a high-breakdown point estimator [6]. This study performs adjustment computations and outlier detection according to LS and LTS methods in different scenarios. Then, LS and LTS have been compared to determine the advantages and disadvantages of the methods.

* **Corresponding Author**

*(hasan.dilmac@omu.edu.tr) ORCID ID 0000-0001-6877-8730
(ysisman@omu.edu.tr) ORCID ID 0000-0002-6600-0623

Cite this article

Dilmaç, H., & Şişman, Y. (2023). New approaches for outlier detection: The least trimmed squares adjustment. International Journal of Engineering and Geosciences, 8(1), 26-31

2. Method and Material

2.1. Method

2.1.1. Adjustment Computation

The adjustment computation is performed to obtain unique values for the unknowns when there are redundant observations in a problem [28]. The adjustment is only performed when the observation number is more than the unknown number [4]. An objective function is selected to find the unique optimum solution [29]. Adjustment computation performed according to the objective function is an optimization problem, which minimizes the selected function [30]. A mathematical model representing the mathematical relationship of observations and unknowns is established [31]. The mathematical model accounts for an essential part of adjustment computation, and it is usually composed of two parts: a functional model and a stochastic model. When observations are made, a functional model is typically chosen to represent the physical situation. The stochastic model determines variances and covariances of the observations [3, 4, 28]. In the classical Gauss-Markov model, the functional and stochastic model can be expressed as below:

$$\begin{aligned} v &= Ax - l \\ P &= Q_{ll}^{-1} = \sigma_0^2 C_{ll}^{-1} \end{aligned} \qquad (1)$$

Here $v, A, x, l, P, \sigma_0^2$ and C_{ll} are the residual vector, the coefficient matrix, unknown vector, the observation vector, the weight matrix, a priori variance, and the covariance matrix, respectively.

If the functional and stochastic models are correct, the adjustment computation gives optimal results [3].

2.1.2. The Least Squares Method

The LS is a method used in adjustment computation by minimizing the sum of the squared weighted residuals to get unique values with redundant measurements [4,32]. The objective function of LS can be given as follows:

$$v^T P v = \sum_{i=1}^{n} p_i v_i^2 \rightarrow min \qquad (2)$$

where n is the number of observations. The steps of adjustment computation can be illustrated as in Figure 1.

The main problem of LS is that even one outlier might severely affect the LS method [33]. LS can propagate errors from one observation to another observation. Therefore, masking and swamping effects occur if there is more than one outlier in the data. A bad observation could seem like a good one because of the propagation of errors; this is called a masking effect. On the contrary, the good observation could seem bad; this is called the swamping effect [7]. There are classical outlier tests to detect outliers depending on the LS. Baarda test (Data-Snooping, W-test), Pope test (Tau test) and t-test are most common in geodesy [8,9,10,34,35,36]. Because the classical outlier tests are based on the LS, the results of these tests can be affected by outliers, too.

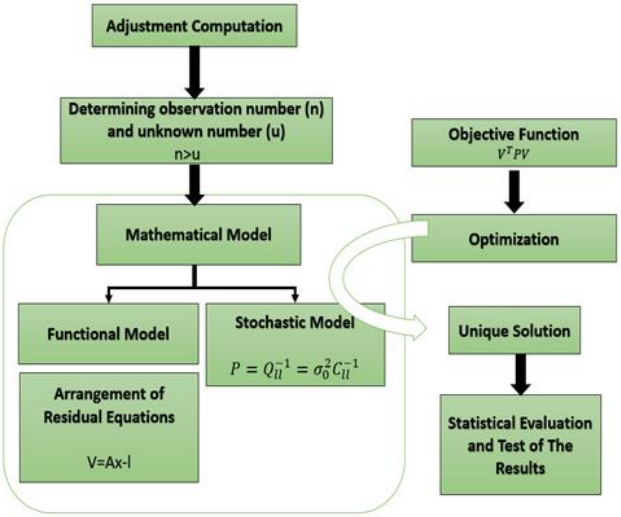

Figure 1. The steps of adjustment computation

2.1.3. Robust Estimation and Outlier Detection

Real data sets may contain outliers [37]. Therefore, robust methods that cannot be affected easily by outliers should be developed. Robustness usually means insensitivity to outliers [38].

There are many robust methods. L_1-norm is the oldest method of these robust methods. Then, M-estimators, R-Estimators, and L-Estimators appeared [6]. To compare the robustness of these methods, the 'breakdown point' concept has been used. The breakdown point means the smallest number of outliers, which may affect an estimator negatively [39]. The robust methods mentioned above don't have a high-breakdown point [6]. Because of this, generalised M-Estimators was developed. Then, Repeated Median, The Least Median of Squares (LMS) [40], S-Estimators, MM-Estimators, and The Least Trimmed Squares (LTS) were developed respectively [19,41,42].

2.1.4. The Least Trimmed Squares

The LTS was presented by Rousseeuw in 1987. This method is quite similar to LS except that the largest squared residuals are removed from the data [21]. The objective function of LTS can be given the following:

$$Min \sum_{i=1}^{h} P_i v_i^2 \qquad (3)$$

Here, h is called as trimming constant and it determines the breakdown point of the LTS [43].

When h is approximately $h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$ (n, number of observations; p, the number of regression parameters), the best robustness is achieved [6]. The LTS problem tries to find which result has the minimum sum of the squared residuals from $S = \binom{n}{h}$ subset of LS solutions [39]. There are two ways to solve the LTS

problem: Exact LTS solution and Approximate LTS solution. The exact LTS solution includes the searching through all subsets of S. But it is hardly possible to solve the Exact LTS solution unless the data size is small enough. On the other hand, the Approximate LTS solution searches through a certain number of subsets of S. [44,45]. In this study, the Exact LTS solution was selected because the data size is small enough.

The step of the Exact LTS solution can be illustrated as in Figure 2.

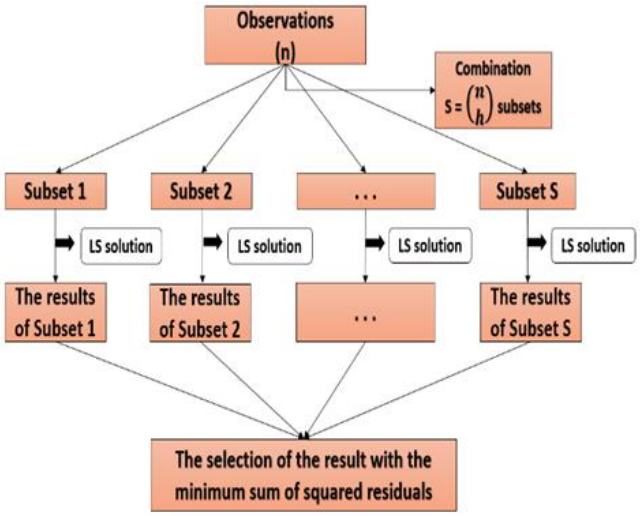

Figure 2. The steps of Exact LTS solution

In the first step, h is determined. Then, $S = \binom{n}{h}$ subsets are obtained. For all subsets, LS method is performed. Finally, the subset with a minimum sum of squared residuals is selected, and this subset's results are used.

2.2. Material

In numerical applications, a regression model such as $y = ax^2 + bx + c$ was used.

Here, a, b and c are regression coefficients. x is the independent variable; y is the dependent variable.

Regression coefficients a, b and c were taken as 2,3 and 5, respectively and y values were calculated according to x values that were chosen arbitrarily for ten observations. Two applications were made by adding random error and gross error to observations. In Application 1, both LS and LTS methods were performed using x and y values with random errors ($\pm 0 - 1$ unit interval). In LTS, h was taken as eight and $\binom{10}{8}$ solutions were made because two observations in the dataset would be simulated as the outlier. In Application 2, gross errors ($+5$ and $+10$ units) were added to y values (observations 3 and 9, respectively) and the outliers were simulated explicitly in this way to compare LS and LTS. Then, LS and LTS methods were performed again using those values. Additionally, classical outlier tests were applied to LS results to detect the outliers. Matlab was used for these solutions.

3. Results

When there is no outlier in observations, the results of LS and LTS are close to the real values and each other. (Table 1). This can be concluded by looking at regression parameters (a, b and c). But if examined more carefully, the regression parameters of the LTS are more accurate except c parameter.

Table 1. The regression results of Application 1

Methods	a	b	c	[VV]
LS	2.03	2.77	5.15	2.13
LTS	2.00	3.09	4.52	0.89

In Application 2, it is clear that the results of the LS are quite contaminated, and the sum of residuals squared increased very much (Table 2). The coefficients b and c of the LS in Application 2 are quite different from expected. The regression parameters of LTS in Application 2 has the results much closer to simulated parameters (2, 3 and 5, respectively) when compared to that of LS in Application 2. Also, the sum of residuals squared is relatively much smaller. It can be seen that the outliers affected the results of the LS regression in Application 2.

Table 2. The regression results of Application 2

Methods	a	b	c	[VV]
LS	2.23	0.28	12.82	89.49
LTS	2.06	2.51	5.44	1.53

Classical outlier tests were applied to detect outliers in the LS method in Application 2. Test values of observations and table values were computed and compared. Classical outlier tests are iterative and detect one outlier (observation with the highest test value) each time. So, two iterations were made in total. Observation 9 was detected as an outlier in Iteration 1 (Table 3).

Table 3. Iteration 1 results of the LS with classical outlier tests in Application 2

Observation	Test Value	Table Value	Result
1	0.1528	2.3646	Observation 9 is an outlier
2	0.0848		
3	0.4089		
4	0.6009		
5	0.1333		
6	2.2771		
7	0.2423		
8	0.1795		
9	12.5992		
10	0.1476		

In Iteration 2, there is no outlier detected. All test values are smaller than the table value (Table 4).

Table 4. Iteration 2 results of the LS with classical outlier tests in Application 2

Observation	Test Value	Table Value	Result
1	0.2875		
2	1.7702		
3	2.3752		
4	0.3710	2.4469	There is no outlier.
5	0.5186		
6	1.7479		
7	1.1552		
8	1.5683		
10	0.1169		

LS method and the classical outlier tests were affected by masking and swamping effects. Although the sum of residuals squared decreased, the regression parameters (especially *b* and *c*) are not close to their real values (Table 5).

Table 5. The regression results of LS results after the classical outlier tests in Application 2

Methods	a	b	c	[VV]
LS	2.15	1.65	7.0355	3.26

LTS method doesn't need classical outlier tests to detect outliers because the trimmed subset with the minimum sum of squared residuals tells us that the removed observations could be an outlier. Observations 3 and 9 were detected as an outlier in LTS method of Application 2. The results are presented in Table 6.

The residuals of the LS regression in Application 1 are small as expected. The effects of outliers on the residuals for the LS in Application 2 can be seen in Figure 3. It is seen that the gross errors added to the Observation 3 were distributed to the other observations in the LS method as a result of distribution of error.

Observations 3 and 9 were simulated as an outlier and classical outlier tests based on the LS method could detect only Observation 9 as outlier correctly. The error in Observation 3 was distributed to the other observations again. Therefore, it could not be detected as outlier (Figure 4).

Table 6. The outlier detection results of the LTS in Application 2

Observation	V (Residuals of Observations)	Result
1	0.1360	Observations 3 and 9 have the largest value of residuals. So, they are removed.
2	-0.6501	
4	-0.0890	
5	-0.6136	
6	0.1842	
7	0.7493	
8	-0.0458	
10	0.3291	

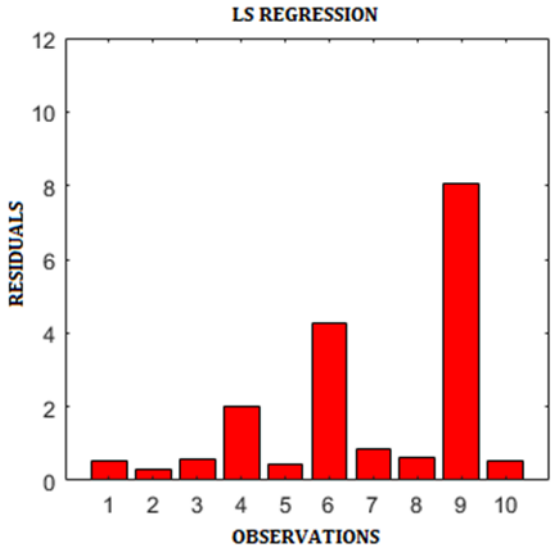

Figure 3. The residuals of the LS in Application 2 with all observations

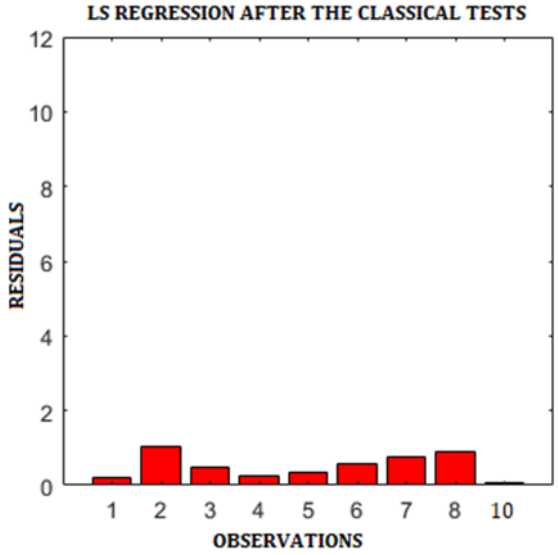

Figure 4. The residuals of the LS in Application 2 after the classical outlier tests

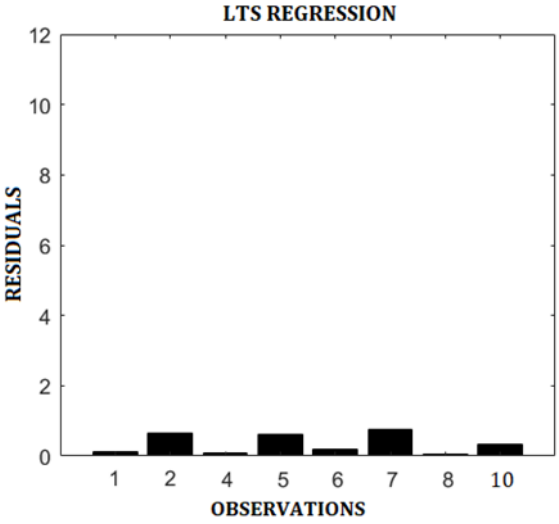

Figure 5. The residuals of LTS in Application 2

The outliers 3 and 9 were detected in the LTS and they were removed from the datasets. (Figure 5) The observations that are removed from the dataset in LTS could be interpreted as an outlier. But, the true selection of the trimming constant h plays a significant role in this. If h is not selected in a way that does not reflect the exact number of outliers, good observations may be removed from the datasets as in Application 1.

4. Discussion and Conclusion

In geodesy, the LS is usually used for adjustment computation. LS results are sufficient when there is no outlier in the observation group. But outliers may occur in observation. Because of this, robust methods have been developed. In this study, the LTS, which is a robust method and the LS were performed.

A regression model was used in this study to analyse the LS and the LTS method results using different scenarios. In Application 1, observations with only random errors were used. The LS and the LTS methods gave good results. But it can be said that the LTS has a little better result when the regression parameters in Table 1 are examined. The parameter a, b except c are closer to their real values and the sum of residuals squared of LTS is relatively smaller. In Application 2 where the simulated outliers were used, while LS results were affected badly from outliers, LTS results gave results close to ones in Application 1.

LTS results might be as good as LS results when observations do not contain any outliers. But it must be noted that the LTS could lead good observations to be eliminated from the dataset simultaneously. Therefore, it is essential to know exactly whether the dataset contains outliers and how many outliers it contains. On the other hand, the LTS can give much better results than the LS when observations have outliers because the LS method can distribute the outlier effect to the other points. It can be said that the LTS method could be more effective in general when compared to the LS. Besides, if the rate of outliers in a dataset is approximately or exactly known, the LTS can detect the outliers correctly without any additional outlier test.

Author contributions

Hasan Dilmaç: Conceptualization, Methodology, Writing-Original draft preparation, Software, Field study, Visualization **Yasemin Şişman:** Validation, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Fan, H. (1997). Theory of errors and least squares adjusment. Royal Institute of Technology, 72, 100-44, Stockholm, Sweden.
2. Ingram, E. L. (1911). Geodetic surveying and the adjustment of observations (methods of least squares). McGraw-Hill Book Company, Inc. 370 Seventh Avenue, New York.
3. Ghilani, C. D. (2017). Adjustment computations: Spatial data analysis (Sixth edition). John Wiley ve Sons, Inc., Hoboken, New Jersey.
4. Mikhail, E. M. & Ackermann, F. E. (1976). Observations and least squares. Thomas Y. Crowell Company, Inc. 666 Fifth Avenue, New York.
5. Čížek, P. & Víšek, J. Á. (2005). Least Trimmed Squares. XploRe®—Application Guide,49-63. Springer, Berlin, Heidelberg.doi: 10.1007/978-3-642-57292-0_2
6. Rousseeuw, J. R. & Leroy, A. M. (1987). Robust Regression and Outlier Detection. John Wiley ve Sons, Inc.
7. Hekimoglu, S. (2005). Do Robust Methods Identify Outliniers More Reliably Than Conventional Tests for Outliniers? Zeitschrift für Vermessungwesen, 3, 174-180.
8. Baarda, W. (1968). A testing procedure for use in geodetic networks. Netherlands Geodetic Com., New Series, Delft, Netherlands, 2(5).
9. Pope, A. J. (1976). The statistics of residuals and the detection of outliers. NOAA Technical Report. NOS 65 NGS 1, U. S. Dept. of Commerce, Rockville, Md.
10. Koch, K. R. (1999). Parameter Estimation and Hypothesis Testing in Linear Models. 2nd Ed. Springer-Verlag, Berlin-Heidelberg, New York.
11. Yetkin, M. & Berber, M. (2013). Application of the sign-constrained robust least-squares method to surveying networks. Journal of Surveying Engineering, 139:1, 59-65. http://doi.org/10.1061/(ASCE)SU.1943-5428.0000088
12. Fabozzi, F. J., Focardi, S. M., Rachev, S. T. & Arshanapalli, B. G. (2014). *The Basics of Financial Econometric: Tools, Concepts and Asset Management Applications*. John Wiley ve Sons, Inc.
13. Rousseeuw, P. J. & Hubert, M. (2018). Anomaly detection by robust statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8:2, e1236. http://doi.org/10.1002/widm.1236
14. Bektas, S. & Sisman, Y. (2010). The comparison of L1 and L2-norm minimization methods. International Journal of the Physical Sciences, 5:11, 1721-1727. http://doi.org/10.5897/IJPS
15. Erdogan, B. (2014). An outlier detection method in geodetic networks based on the original observation. Boletim de Ciencias Geodesicas, 20:3, 578-589, http://doi.org/10.1590/S1982-21702014000300033
16. Giloni, A. & Padberg, M. (2001). Least Trimmed Squares Regression, Least Median Squares Regression and Mathematical Programming. *Matmetical and Computer Modelling*, 35:9-10, 1043-1060. http://doi.org/10.1016/S0895-7177(02)00069-9
17. Gui, Q. & Zhang, J. (1998). Robust biased estimation and its applications in geodetic adjustments. Journal of Geodesy, 72:7-8, 430-435. http://doi.org/10.1007/s00190005 0182
18. Hekimoglu, S. & Erenoglu, C. (2007). Jeodezik Ağlarda Uyuşumsuz Ölçülerin Klasik Yaklaşım ve Robust

Yöntemlerle Belirlenmesi. Jeodezi ve Jeoinformasyon Dergisi, 97, 3-14.
19. Hubert, M., Rousseeuw, P. J. & Van Aelst, S. (2008). High-breakdown robust multivariate methods. Statistical science, 23:1, 92-119. http://doi.org/10.1214/088342307000000087
20. İnal, C. & Yetkin, M. (2006). Robust yöntemlerle uyuşumsuz ölçülerin belirlenmesi. Selçuk Üniversitesi Mühendisler-Mimarlar Fakültesi Dergisi, 21, 3-4.
21. Knight, N. L. & Wang, J. (2009). A comparison of outlier detection procedures and robust estimation methods in GPS positioning. The Journal of Navigation, 62:4, 699-709 http://doi.org/10.1017/S0373463309990142
22. Sisman, Y. (2010). Outlier measurements analysis with the robust estimation. Scientific Research and Essays, 5:6, 668-678.
23. Sisman, Y. (2011). Parameter estimation and outlier detection with different estimation methods. Scientific Research and Essays, 6:7, 1620-1626 http://doi.org/10.5897/SRE10.1181.
24. Susanti, Y., Pratiwi, H., Sulistijowati, S., & Liana, T. (2014). M estimation, S estimation, and MM estimation in robust regression. International Journal of Pure and Applied Mathematics, 91:3, 349-360. http://doi.org/10.12732/jipam. v91i3.7
25. Valero, J. L. B., & Moreno, S. B. (2005). Robust estimation in geodetic networks. Física de la Tierra, 17, 7.
26. Yang, Y. (1999). Robust estimation of geodetic datum transformation. Journal of Geodesy, 73:5, 268-274. http://doi.org/10.1007/s001900050243
27. Yetkin, M. & Inal, C. (2011). L1 norm minimization in GPS networks. Survey Review, 43:323, 523-532. http://doi.org/10.1179003962611x131177 48892038
28. Ogundare, J. O. (2018). Understanding Least Squares Estimation and Geomatics Data Analysis. John Wiley ve Sons, Inc, 111 River Street, Hoboken, NJ 07030, USA
29. Sisman, Y. & Bektas, S. (2012). Linear regression methods according to objective functions. Acta Montanistica Slovaca, 17:3, 209-217.
30. Grafarend, E. W. & Sansò, F. (Editors) (2012). Optimization and design of geodetic networks. Springer Science & Business Media, Heidelberg, Berlin.
31. Schaffrin, B. (2019). Notes on Adjustment Computations Part I.
32. Wells, D., & Krakiwsky, E. (1971). The Method of least squares. University of New Brunswick: Canada
33. Muhlbauer, A., Spichtinger, P., & Lohmann, U. (2009). Application and comparison of robust linear regression methods for trend estimation. *Journal of Applied Meteorology and Climatology,* 48:9, 1961-1970. http://doi.org/10.1175/2009JAMC1851.1
34. Hekimoğlu, Ş., Erdogan, B., Soycan, M., & Durdag, U. M. (2014). Univariate Approach for Detecting Outliers in Geodetic Networks. Journals of Surveying Engineering, 140:2, 04014006, 1-8. http://doi.org/10.1061/(ASCE)SU.19435428.0000123
35. Kavouras, M. (1982). On the detection of outliers and the determination of reliability in geodetic networks. Department of Surveying Engineering Technical Report No. 87, University of New Brunswick, Fredericton, N.B., November.
36. Sisman, Y. (2005). Uyuşumsuz Ölçü Gruplarının Belirlenmesi Yöntemleri. Harita Dergisi, 133.
37. Rousseeuw, J. R. (1990). Robust estimation and identifying outliers. Handbook of statistical methods for engineers and scientists, 16-1.
38. Huber, P. J. (1981). Robust Statistics. John Wiley and Sons, Inc.
39. Hofmann, M., Gatu, C., & Kontoghiorghes, E. J. (2010). An Exact Least Trimmed Squares Algorithm for a Range of Coverage Values*, Journal of Computational and Graphical Statistics,* 19:1, 191-204, http://doi.org/10.1198/jcgs.2009.07091
40. Rousseeuw, P. J. (1984). Least Median of Squares Regression. Joırnal of the American Statistical Association, 79:388, 871-880. http://doi.org/10.1080/016 21459.1984.10477105
41. Toka, O., & Cetin, M. (2011). The comparing of S-estimator and M-estimators in linear regression. *Gazi University Journal of Science,* 24, 4, 747-752
42. Staudte, R. G., & Sheather, S. J. (2011). *Robust estimation and testing*. John Wiley ve Sons, Inc, 918.
43. Cizek P (2005). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning,* 136, 3967-3988, http://doi.org/10.1016/j.jspi.2005.05.004
44. Cizek, P., & Visek, J. A. (2000): Least trimmed squares, SFB 373 Discussion Paper, No. 2000, 53, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin.
45. Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS Regression for Large Data Sets. Data Min Knowl Disc 12, 29-45. https:/doi.org/10.1007/s10618-005-0024-4