On the Performance of Some Location Estimators Regarding the Skewness of Data

Vasil Cvetkov[1*], Gergana Antova[2], Christina Mickrenska[3], Danail Nedyalkov[4]

[1*] University of Architecture, Civil Engineering and Geodesy, Sofia 1046, Bulgaria, tzvetkov_fgs@uacg.bg; ORCID 0000-0001-9628-6768
[2] University of Architecture, Civil Engineering and Geodesy, Sofia 1046, Bulgaria, antova_fgs@uacg.bg; ORCID 0000-0002-0926-8033
[3] University of Architecture, Civil Engineering and Geodesy, Sofia 1046, Bulgaria, kmikrenska_fgs@uacg.bg; ORCID 0000-0002-7054-1799
[4] University of Architecture, Civil Engineering and Geodesy, Sofia 1046, Bulgaria, dned_far@uacg.bg

https://doi.org/10.29227/IM-2026-02-02-040

Submission date: *2026-22-04* | Review date: *2026-30-04*

Abstract

This study investigates the performance of various location estimators across increasing skewness levels (0.00–1.70), evaluating their relative efficiency relative to the sample mean. The analysis is based on simulations of standard normal data contaminated with 10% extreme values to mimic real-world deviations. The findings reveal substantial variation in robustness and bias among estimators as skewness increases. Using the sample mean as a benchmark, the median exhibits a sharp increase in relative efficiency, rising from 66 to 781 as skewness intensifies. Robust estimators such as Hodges–Lehmann and Hampel show moderate increases in efficiency, indicating partial sensitivity to skewness. In contrast, the Danish, Andrews, Tukey, S-25, and MM estimators demonstrate strong robustness to contamination, with rapidly increasing efficiency values, making them well-suited for highly skewed distributions. The Huber family of estimators exhibits performance that depends on the tuning parameter: smaller constants (e.g., Huber (1.345)) provide greater robustness to contamination, whereas larger constants (e.g., Huber (4)) behave more like the mean, indicating reduced resistance to outliers. The L1-norm estimator displays only limited robustness across skewness levels. Overall, estimator performance is strongly influenced by skewness. The MM-estimator and the S-estimator (with a 25% breakdown point) emerge as the most robust and efficient choices, even under moderate skewness (greater than 0.10). The Danish estimator also demonstrates notable robustness. Conversely, the L1-norm and high-parameter Huber estimators appear less suitable for heavily skewed data.

Keywords: *Probability Theory, Central Tendency, Maximum Likelihood, Robust Estimation, Reliability of Results.*

1. Introduction

Accurate estimation of central tendency is essential in geodetic and engineering practice, where observational data form the basis for positioning, height determination, deformation analysis, and infrastructure design. Measurements obtained in the field are often affected by random errors, systematic effects, and occasional gross errors due to instrument limitations, environmental influences, or human factors. In such contexts, a reliable estimate of the central value is crucial for ensuring the precision, consistency, and stability of computed coordinates, elevations, and other derived quantities. Even small biases in central tendency estimation may propagate through calculations and lead to significant deviations in final results, which can affect both the safety and the economic efficiency of engineering projects.

Beyond descriptive purposes, central tendency plays a key role in statistical inference and hypothesis testing. Many classical statistical procedures, including confidence interval estimation and significance testing, rely on assumptions about the underlying distribution and the accuracy of location estimates. If the chosen estimator is sensitive to outliers or deviations from normality, statistical tests may yield misleading conclusions, increasing the risk of incorrect decisions. This is particularly critical in geodetic quality control, where decisions about data acceptance, network adjustment, and deformation detection depend on reliable statistical assessments.

Conventional estimators, such as the sample mean, are optimal under ideal conditions but can perform poorly when data are skewed or contaminated. In such cases, robust estimators offer a more reliable alternative, as they are less influenced by extreme values and better reflect the true central tendency of the data. The selection of an appropriate estimator is therefore not only a matter of statistical efficiency but also of practical reliability.

The objective of this study is to compare the performance and reliability of various location estimators in the presence of skewed data, with particular emphasis on their suitability for geodetic and engineering applications.

2. Methodology, empirical and real examples

2.1 Methodology

As all estimators considered in this study are well known and extensively discussed in the literature [1–7], only a brief overview is provided.

2.1.1 Arithmetic mean. Let us have the sample $x_1, x_2, \ldots, x_n$. Then, the sample mean $\bar{x}$ can be given by equation (1).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

2.1.2 Median. To compute the median, the values $x_1, x_2, \ldots, x_n$ must first be arranged in ascending order. If the sample size n is odd, the median is the middle value in the ordered sequence. If n is even, the median is defined as the arithmetic mean of the two central values. By definition, the median divides the sample such that 50% of the observations lie below it and 50% lie above it.

2.1.3 Hodges–Lehmann nonparametric estimator. The Hodges-Lehmann estimator applied in our research is given by equation (2) [1, 3, 7],

$$E^{HL} = median\left[\frac{1}{2}(\Delta^* + \Delta^{**})\right] \tag{2}$$

Where:

$$\Delta^* = \{x_1, x_2, \ldots, x_n\} \tag{3}$$

$$\Delta^{**} = \{x_1, x_2, \ldots, x_n\} \tag{4}$$

2.1.4 Danish Method. The weighting function of the Danish method is given by equation (5) [1].

$$w(\overline{v_i}) = \begin{cases} 1 & for\ |\overline{v_i}| \le 2\sigma \\ \exp(k\overline{v_i}^2) & for\ |\overline{v_i}| > 2\sigma \end{cases} \tag{5}$$

In our study, we used $k = 2$. In equation (5), v_i is the residual of the i^{th} observation, $\hat{\sigma}$ is the standard deviation, and $\overline{v_i}$ is the standardized residual, calculated by equations (6) and (7) in each iteration.

$$\overline{v_i} = v_i/\hat{\sigma} \tag{6}$$

$$\hat{\sigma} = 1.4826.med|v_i - med(v_i)| \tag{7}$$

2.1.5 L1-Norm. The weighting function for the L1-norm method, also known as the least absolute deviation method, is given by equation (8) [1, 8].

$$w(v_i) = \left|\frac{1}{v_i}\right| \tag{8}$$

If $v_i = 0$, then we set $v_i = 0.0000001$.

2.1.6 Huber's M-Estimator. The weighting function of Huber's method is given by equation (9) [1, 2].

$$w(\overline{v_i}) = \begin{cases} 1 & for\ |\overline{v_i}| \le k \\ \frac{k}{|\overline{v_i}|} & for\ |\overline{v_i}| > k \end{cases} \tag{9}$$

In our study, we use four values of $k = \{1.345, 2, 3, \text{and } 4\}$. If $v_i = 0$, then we set $v_i = 0.0000001$.

2.1.7 Hampel's M-Estimator. The weighting function of Hampel's method is given by equation (10).

$$w(\overline{v_i}) = \begin{cases} 1 & for\ |\overline{v_i}| \le k_1 \\ \frac{k_1}{|\overline{v_i}|} & for\ k_1 < |\overline{v_i}| \le k_2 \\ \frac{k_1}{k_3 - k_2}\left(\frac{k_1}{|\overline{v_i}|} - 1\right) & for\ k_2 < |\overline{v_i}| \le k_3 \\ 0 & for\ |\overline{v_i}| > k_3 \end{cases} \tag{10}$$

In our study, we used $k_1 = 2$, $k_2 = 4$, $k_3 = 8$ [1]. If $v_i = 0$, then we set $v_i = 0.0000001$.

2.1.8 Andrew's M-Estimator. Andrew's weighting function is given by equation (11) [1, 4].

$$w(\overline{v_i}) = \begin{cases} \frac{sin\left(\frac{|\overline{v_i}|}{k}\right)}{\frac{|\overline{v_i}|}{k}} & for\ |\overline{v_i}| \le \pi k \\ 0 & for\ |\overline{v_i}| > \pi k \end{cases} \tag{11}$$

In our study, we used $k = 1.339$ {1]. If $v_i = 0$, then we set $v_i = 0.0000001$.

2.1.9 Tukey's M-Estimator. The weighting function of Tukey's method is given by equation (12).

$$w(\overline{v_i}) = \begin{cases} \left[1 - \left(\frac{\overline{v_i}}{k}\right)^2\right]^2 & for\ |\overline{v_i}| \le k \\ 0 & for\ |\overline{v_i}| > k \end{cases} \tag{12}$$

In our study, we used $k = 4.685$ and $k = 6$ [1].

2.1.10 S-Estimation. S-estimators, introduced by Peter J. Rousseeuw and Victor J. Yohai, were developed to handle datasets with a high proportion of contamination [5]. Therefore, the weight function of S-estimators can be given by equation (13).

$$w(\overline{v}_i) = \begin{cases} \left[1 - \left(\frac{\overline{v}_i}{c}\right)^2\right]^2 & for\ |\overline{v}_i| \le c \\ 0 & for\ |\overline{v}_i| > c \end{cases} \quad (13)$$

In our study, we applied c = 2.937, corresponding to a breakdown point (BDP) of 25% [5]. The S-estimation algorithm is given in [5, 9].

2.1.11 MM-Estimation. MM-estimators are robust estimators that combine high efficiency under normal errors with a 50% breakdown point under heavy contamination [5-6, 9]. Introduced in [6], they are obtained through a two-stage procedure: first, an S-estimator with c = 1.547 is used to ensure a high breakdown point; second, an M-estimator with a more efficient influence function is applied to improve asymptotic efficiency. In this study, the second stage employs Tukey's M-estimator with k = 4.685. Implementation details are provided in [5–6, 9].

2.2 Empirical tests by Monte Carlo simulations

To evaluate the performance of the proposed location estimators, samples with sizes ranging from 10 to 50 observations were generated from a standard normal distribution $N(0,1)$. To introduce contamination and varying levels of skewness (0.10, 0.20, 0.30, 0.50, 1.00, 1.30, 1.50, and 1.70), 10% of the observations were shifted by 1.00, 1.25, 1.50, 2.00, 3.00, 4.00, 5.00, and 6.00 units, respectively. For each contamination level, the procedure was repeated 10,000 times. The generated samples were then used to compute the location estimators described in Section 2.1. Their performance was assessed using the root-mean-squared error (RMSE), and since the true mean is E=0, equation (14) applies.

$$\theta_E = \sqrt{\frac{1}{10000}\sum_{i=1}^{10000}(E_i)^2} \quad (14)$$

On this basis, the empirical relative efficiency of each estimator, referenced to the RMSE of the sample mean $\theta mean$, is calculated according to equation (15), and the results are presented in Table 2.

$$e(Mean, Etimator) = (\theta_{mean}/\theta_E)^2 \quad (15)$$

2.3 Real data example

As a real-data example in this study, we used measurements of six elevation differences from a levelling test loop near Sofia, collected during tests of the precise level Ni 002 [10]. The processed data are presented in Table 1. The aim is to assess the reliability of the conclusions drawn from the Monte Carlo simulations by estimating the central tendency of these six section elevations and, based on these estimates, computing the loop closing errors for each of the considered statistics. The approach that yields the smallest loop closing error is regarded as the most reliable.

Tab. 1. Initial data of the real observation example [10].

Number of the loop	Elevations in the Levelling Sections (m)					
	1-2	2-3	3-4	4-5	5-6	6-1
1	7.07872	11.35026	0.82324	-4.11120	-9.19346	-5.94726
2	7.07900	11.35022	0.82314	-4.11197	-9.19313	-5.94658
3	7.07852	11.34996	0.82311	-4.11159	-9.19345	-5.94720
4	7.07912	11.35054	0.82292	-4.11162	-9.19331	-5.94655
5	7.07907	11.35000	0.82322	-4.11107	-9.19334	-5.94709
6	7.07898	11.34986	0.82338	-4.11124	-9.19356	-5.94670
7	7.07876	11.35008	0.82362	-4.11174	-9.19333	-5.94690
8	7.07862	11.34936	0.82314	-4.11166	-9.19306	-5.94689
9	7.07847	11.34986	0.82324	-4.11126	-9.19334	-5.94715
10	7.07839	11.34995	0.82328	-4.11142	-9.19318	-5.94660
11	7.07863	11.34894	0.82349	-4.11137	-9.19302	-5.94680
12	7.07854	11.34904	0.82368	-4.11140	-9.19287	-5.94688
13	7.07856	11.34915	0.82306	-4.11148	-9.19303	-5.94700
14	7.07837	11.34942	0.82331	-4.11114	-9.19276	-5.94632
15	7.07839	11.34908	0.82335	-4.11148	-9.19302	-5.94668

3. Results

Table 2 presents the empirical efficiency of the analysed estimators of central tendency across different levels of data skewness. Table 3 shows the estimated central values of the six section elevations listed in Table 1, along with the corresponding loop closing errors derived from these estimates. Fig. 1 illustrates the similarity among the estimated central values obtained using the different estimators. The differences between methods are computed relative to the Danish estimator, as it yields the smallest closing error.

Tab. 2. Empirical relative efficiency of the analysed statistics relative to the sample mean under varying distribution skewness (sample sizes = [10, 50]).

Statistics	Skewness								
	0.00	0.10	0.20	0.30	0.50	1.00	1.30	1.50	1.70
Mean	100	100	100	100	100	100	100	100	100
Median	65	88	103	121	162	268	408	576	781
Hodges–Lehmann	95	101	108	117	140	215	324	457	619
Danish	88	98	104	116	150	332	636	1001	1453
L1-Norm	84	95	102	109	120	139	154	169	186
Huber (1.345)	89	99	109	123	169	333	559	828	1150
Huber (2)	97	101	104	109	124	196	339	528	745
Huber (3)	99	100	102	103	110	138	201	302	425
Huber (4)	100	100	100	100	100	102	106	113	123
Hampel	96	100	104	108	124	200	353	546	697
Andrews	82	98	112	131	180	332	558	863	1263
Tukey (4.685)	82	98	112	131	180	331	555	851	1238
Tukey (6)	87	99	112	128	171	308	509	765	1089
S-Estimation (BDP 25%)	89	101	111	123	155	317	722	1099	1488
MM-Estimation	93	102	110	120	149	277	578	1073	1646

Tab. 3. Calculated location of the observed section elevations by various statistics and the closing errors of the test levelling loop based on them.

Statistic	Calculated Elevations in the Levelling Sections (m)						Closing Errors (mm)
	1-2	2-3	3-4	4-5	5-6	6-1	
Mean	7.07868	11.34971	0.82328	-4.11144	-9.19319	-5.94684	0.20
Median	7.07862	11.34986	0.82324	-4.11142	-9.19318	-5.94688	0.24
Hodges–Lehmann	7.07867	11.34969	0.82326	-4.11143	-9.19318	-5.94684	0.17
Danish	7.07863	11.34971	0.82326	-4.11142	-9.19319	-5.94684	0.15
L1-Norm	7.07867	11.34986	0.82328	-4.11142	-9.19318	-5.94688	0.33
Huber (1.345)	7.07868	11.34971	0.82326	-4.11142	-9.19320	-5.94685	0.18
Huber (2)	7.07868	11.34971	0.82328	-4.11144	-9.19319	-5.94684	0.20
Huber (3)	7.07868	11.34971	0.82328	-4.11144	-9.19319	-5.94684	0.20
Huber (4)	7.07868	11.34971	0.82328	-4.11144	-9.19319	-5.94684	0.20
Hampel	7.07867	11.34970	0.82323	-4.11140	-9.19320	-5.94680	0.20
Andrews	7.07862	11.34985	0.82326	-4.11143	-9.19320	-5.94684	0.26
Tukey (4.685)	7.07862	11.34984	0.82326	-4.11143	-9.19320	-5.94684	0.25
Tukey (6)	7.07862	11.34984	0.82326	-4.11143	-9.19320	-5.94684	0.25
S-25	7.07866	11.34973	0.82327	-4.11143	-9.19320	-5.94684	0.19
MM	7.07866	11.34972	0.82326	-4.11144	-9.19319	-5.94684	0.17

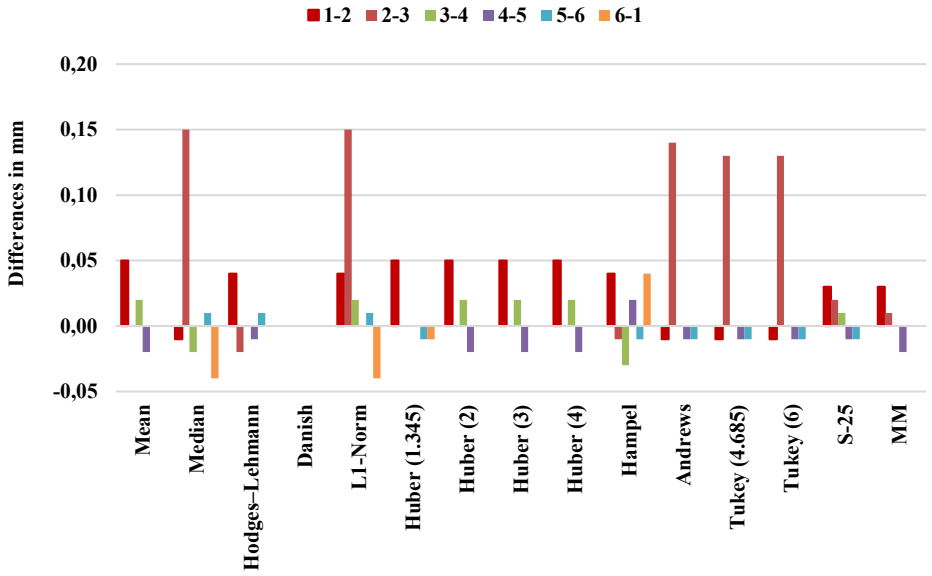

Fig.1. Comparison between the calculated central locations of the section elevations in the test loop near Sofia, given in Tab. 1.

According to Figure 1, the Danish method, MM-estimation, and S-estimation (with a 25% breakdown point) yield the most similar results. Most M-estimators, such as the L1-norm, Huber, and Hampel estimators, produce results close to those of the sample mean. Notably, some methods show distinct deviations for section 2–3, which is of particular interest.

4. Discussion

We begin by examining the results presented in Table 2. A clear relationship is observed between the reliability of the analyzed estimators and the degree of data skewness. The main findings can be summarized as follows.

The arithmetic mean proves to be the most reliable measure of central tendency only within a narrow skewness range of −0.10 to 0.10. Even at the boundaries of this interval, alternative estimators, such as the Hodges–Lehmann estimator, the MM-estimator, and the S-estimator (BDP = 0.25), often provide superior performance. This suggests that the mean is dependable only under near-perfect symmetry, a condition rarely met in practice.

For mildly skewed data (−0.50 to −0.10 and 0.10 to 0.50), several estimators outperform the mean. In addition to the Hodges–Lehmann, MM-, and S-estimators, M-estimators such as those of Hampel, Tukey, and Andrews, as well as the Danish estimator and the median, yield more reliable results. A similar pattern holds for moderate skewness (−1.00 to −0.50 and 0.50 to 1.00), where multiple robust estimators remain preferable to the mean.

In cases of strong skewness (below −1.00 or above 1.00), the advantage of robust methods becomes pronounced. The most reliable estimators in such situations include Huber (k = 1.345), Andrews, Tukey (k = 4.685), the Danish method, the MM-estimator, and the S-estimator (BDP = 0.25). The arithmetic mean performs poorly under these conditions and may even be the least suitable choice, while the median also shows limited adequacy.

Overall, the results support recommending the MM-estimator and/or the S-estimator (BDP = 0.25) whenever skewness lies outside the narrow interval of −0.10 to 0.10.

To validate these conclusions, a real-data test was conducted using repeated measurements from a levelling test loop near Sofia, Bulgaria. Since the loop closing error represents a true error, it serves as a reliable criterion for evaluating estimator performance. As shown in Table 3, the Danish method yields the smallest closing error, followed by the Hodges–Lehmann estimator, the MM-estimator, Huber (k = 1.345), and the S-estimator (BDP = 0.25). As expected, the arithmetic mean and the median produce less accurate results. Additional evidence of the robustness of the Danish, MM-, and S-estimators is their close agreement in the estimated elevation values.

Although the Andrews and Tukey estimators produce results very similar to these leading methods for five of the six sections, notable discrepancies arise for section 2–3. Similar deviations are observed for the L1-norm estimator and the median. This raises an important question regarding the underlying cause of these differences.

To answer the above question, we calculated the skewness S, kurtosis K, and excess kurtosis E of the samples by equations (16)-(18).

$$S = \frac{m_3}{m_2^{3/2}} \tag{16}$$

$$K = \frac{m_4}{m_2^2} \tag{17}$$

$$E = K - 3 \tag{18}$$

In equations (16) and (17), m_4 is the fourth sample moment about the mean, m_3 is the third sample moment about the mean, and m_2 is the second sample moment about the mean. They can be calculated by equation (19), where $i = \{2, 3, 4\}$, x_j and $\bar{x}$ are the j^{th} observation and the mean of each section sample.

$$m_i = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^i \tag{19}$$

The approximate (but precisely enough) standard errors of skewness and excess kurtosis are given in equations (20) and (21), respectively.

$$\sigma_S \approx \sqrt{\frac{6}{n}} \tag{20}$$

$$\sigma_E \approx \sqrt{\frac{24}{n}} \tag{21}$$

Tab. 4. Skewness and Excess Kurtosis of the real observation data.

Statistics	Elevations in the Levelling Sections (m)					
	1-2	2-3	3-4	4-5	5-6	6-1
Skewness S	0.58±0.63	-0.20±0.63	0.46±0.63	-0.47±0.63	0.21±0.63	0.14±0.63
Excess Kurtosis E	-1.03±1.27	-1.26±1.27	0.09±1.27	-0.03±1.27	-0.78±1.27	-0.67±1.27

According to Table 4, five of the six samples exhibit mild skewness, while the observations for section 1–2 show moderate skewness. None of the samples can be strictly classified as symmetric. However, in all cases, the standard deviation of skewness exceeds the estimated skewness values, which allows the assumption that the samples are effectively symmetric.

Regarding excess kurtosis, five of the six samples display platykurtic behaviour. The highest excess kurtosis is observed in the sample for section 2–3, although its value remains slightly below its corresponding standard deviation. Based on the three-sigma rule, this indicates no statistically significant deviation from normal kurtosis. Overall, the samples can be considered approximately normally distributed.

Despite this, the loop closing error computed using the arithmetic mean is not the smallest. Moreover, the estimated central values for section 2–3, associated with the highest excess kurtosis, show the greatest variability and inconsistency across estimators. This highlights a potential limitation of relying automatically on the arithmetic mean as the preferred measure of central tendency.

5. Conclusion

This study demonstrates that the performance of location estimators is highly sensitive to distributional skewness and contamination. The simulation results show that, while the sample mean serves as a useful benchmark, it becomes increasingly inefficient in the presence of skewed and contaminated data. Among the estimators considered, the MM- and S-estimators (with a 25% breakdown point) consistently exhibit the highest robustness and efficiency, even under moderate skewness, making them the most reliable choices in such conditions. The Danish estimator also proves to be a strong alternative, while the median, despite its increasing relative efficiency, reflects substantial sensitivity to skewness. In contrast, the L1-norm and Huber estimators with larger tuning constants show limited robustness and are less suitable for heavily skewed distributions.

The analysis of real measurement data fully supports the conclusions drawn from the Monte Carlo simulations, confirming the practical relevance of the findings. Importantly, the results suggest that deviations in kurtosis (excess kurtosis) may pose a greater risk to the reliability of central tendency estimation than skewness alone. This highlights the need for further research into the influence of kurtosis, particularly in the context of mid-sized samples, which remains a key direction for future work.

Acknowledgments

The authors express gratitude to the Centre for Scientific Research and Design at the UACEG for funding and support for the completed research project under contract No BN340/2026 on the "Analysis of the suitability of existing and new robust estimators and methods in the mathematical processing of geodetic measurements".

References

1. Duchnowski, R. and Wyszkowska, P., Robust procedures in processing measurements in geodesy and surveying: a review, Measurement Science and Technology, 35(5), (2024), 1-16. https://doi.org/10.1088/1361-6501/ad28ec
2. Huber P. J., Robust Estimation of a Location Parameter, Ann. Math. Statist., 35(1), (1964), 73–101. https://doi.org/10.1214/aoms/1177703732
3. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. Robust statistics: The approach based on influence functions. New York, NY: Wiley. (1986).
4. Andrews D. F. A Robust Method for Multiple Linear Regression, Technometrics, 16, (4), (1974), 523–531. https://doi.org/10.1080/00401706.1974.10489233
5. Rousseeuw P. and Yohai V., Robust Regression by Means of S-Estimators. In: Franke, J., Härdle, W., Martin, D. (eds) Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics, 26, (1984), 256-272, Springer, New York, NY. https://doi.org/10.1007/978-1-4615-7821-5_15
6. Yohai V., High Breakdown-Point and High Efficiency Robust Estimates for Regression, The Annals of Statistics, 15(2), (1987), 642-656. https://doi.org/10.1214/aos/1176350366
7. Duchnowski, R. Hodges–Lehmann estimates in deformation analyses. J. Geod. 87, (2013), 873–884. https://doi.org/10.1007/s00190-013-0651-2
8. Cvetkov, V. Evaluating levelling adjustment methods in deformation analysis, Budownictwo i Architektura / Civil and Architectural Engineering, 25(1), (2026). https://doi.org/10.35784/bud-arch.8789
9. Susanti L., Pratiwi H., Sulistijowati S., and Liana T., M Estimation, S Estimation, and MM Estimation in Robust Regression, International Journal of Pure and Applied Mathematics, 91(3), (2014), 349-360. https://doi.org/10.12732/ijpam.v91i3.7
10. Burilkov, T. First results from measurements with the new compensator level Ni 002. Izvestya GUGK, 3, (1974), 10–19. (in Bulgarian)