Empirical tests of performance of some
M – estimators

Marek Banaś[1], Marcin Ligas[2]

[1]The Bronisław Markiewicz State Higher School of Technology and Economics in Jarosław
Institute of Technical Engineering
16 Czarnieckiego St., 37-500 Jarosław, Poland
e-mail: marek.banas@pwste.edu.pl

[2]AGH University of Science and Technology
Faculty of Mining Surveying and Environmental Engineering
Department of Geomatics,
30 Mickiewicza Al., 30-059 Krakow, Poland
e-mail: marcin.ligas@agh.edu.pl

Abstract: The paper presents an empirical comparison of performance of three well known M – estimators (i.e. Huber, Tukey and Hampel's M – estimators) and also some new ones. The new M – estimators were motivated by weighting functions applied in orthogonal polynomials theory, kernel density estimation as well as one derived from Wigner semicircle probability distribution. M – estimators were used to detect outlying observations in contaminated datasets. Calculations were performed using iteratively reweighted least-squares (IRLS). Since the residual variance (used in covariance matrices construction) is not a robust measure of scale the tests employed also robust measures i.e. interquartile range and normalized median absolute deviation. The methods were tested on a simple leveling network in a large number of variants showing bad and good sides of M – estimation. The new M – estimators have been equipped with theoretical tuning constants to obtain 95% efficiency with respect to the standard normal distribution. The need for data – dependent tuning constants rather than those established theoretically is also pointed out.

Keywords: M – estimation, iteratively reweighted least squares, tuning constant, outliers, network adjustment

1. Introduction and motivation

Problem of handling outlying observations was already considered by D. Bernoulli in his works in 18th century (Stigler, 2010). Also, P. S. Laplace may be considered as a pioneer of what is now known as robust methods with his method of finding the values of q unknown quantities from n observational equations. His method consisted in imposing the conditions that the algebraic sum of the residuals should be zero, and

that their sum, all taken with the positive signs, should be a minimum. By introducing these conditions, he was able to solve over – determined systems of equations. This method he applied to the deduction of the shape of the earth from measurements of arcs of meridians, and also from pendulum observations (Dunnington, 1955). A. M. Legendre one of the "inventors" of least squares method was aware of its sensitivity to blunders and wrote: "If among these errors are some which appear too large to be admissible, then those observations which produced these errors will be rejected, as coming from too faulty experiments and the unknowns will be determined by means of the other observations, which will then give much smaller errors" (Rousseeuw and Leroy, 1987). For the first time the term "robust" was used by G. Box in his paper (Box, 1953) but new great steps and new subdiscipline in statistics called robust estimation really emerged through works by J. W. Tukey (Tukey, 1960; 1962), P. J. Huber (Huber, 1964; 1967) and F. Hampel (Hampel, 1971; 1974).

Robust methods in their present form may be divided into two groups i.e. passive and active. The first group involves methods based on statistical tests such as iterative data snooping (Baarda, 1968) or τ – test (Pope, 1976; Prószyński and Kwaśniak, 2002). These methods are rigorous ones because of complete removal of an observation identified as an outlier. The second group is represented by methods based on robust estimation. The main idea behind the methods is to gradually decrease the influence of outlying observations by reducing their weights. A special place among the latter is occupied by M-estimation introduced by P. J. Huber (1964), as a generalization of maximum-likelihood estimation.

This work is devoted to the last mentioned group i.e. M – estimators. It compares performance of three well known M – estimators of Huber, Tukey and Hampel and some new M – estimators (as far as the authors knowledge goes) as well. The new estimators have been motivated by weighting function known from orthogonal polynomials theory (Jacobi orthogonal polynomials), kernel density estimation (Epanechnikov and tricube kernels) and also from probability distribution functions (Wigner's semicircle probability distribution). One may easily notice some similarities among the mentioned functions and weighting function used in M – estimation. For instance, the function $f(e) = c\,exp\left(\dfrac{-e^2}{2}\right)$ may be associated with standardized normal distribution, Gaussian kernel used in kernel density estimation (KDE), weighting function in Hermite orthogonal polynomials and also with Welsch M – estimator as well as with Danish weighting function without a neutral interval. Comparing Tukey's weighting function one notices its similarity to biweight (quartic) kernel used in KDE. Cauchy's M – estimator is derived from Cauchy's probability distribution. These analogies justify the authors' search for new M – estimators in these branches of mathematics.

Besides the comparison itself the paper shows some imperfections of M – estimation due to dependence of model residuals only. The procedure of comparison incorporates also robust measures of scale i.e. interquartile range and normalized median absolute

deviation; in the estimation process since the residual variance does not belong to the class of robust measures of scale. The new M – estimators have been endowed with theoretical tuning constants assuring 95% efficiency with respect to the standard normal distribution. It is pointed out, however that the data – driven tuning constants (here chosen by trial and error) may limit the number of iterations in the iteratively reweighted least squares procedure while maintaining satisfactory results.

2. M – estimation

The idea behind the M – estimation (maximum likelihood type estimation) introduced by Huber in the sixties of 20th century, discussed and further developed by others relies on replacing objective (loss) function of the least squares method with a less rapidly increasing function (i.e. less increasing than a square). This "new" principle (new objective function ρ) may be expressed as:

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho\left(l_i - \mathbf{a}_i^T \mathbf{x}\right) \to \min \tag{1}$$

where:

l_i – the ith observation
$\mathbf{a}_i^T$ – the ith row of the design matrix (Jacobi matrix – of first derivatives)
$\mathbf{x}$ – vector of model parameters to be estimated
e_i – error corresponding with l_i (disturbance of the model)
n – number of observations

The new loss function satisfies the following conditions:
– $\rho(e_i) \geq 0$, nonnegativity
– $\rho(0) = 0$, is zero when its argument is zero
– $\rho(e_i) = \rho(-e_i)$, is symmetric (even function),
– $\rho(e_i) \geq \rho(e_j)$ for $|e_i| > |e_j|$, monotonicity in $|e_i|$,

There are two other functions on the basis of which M – estimators may be characterized (and also derived in a heuristic way). These are *influence function* and *weighting function*. Basing on influence function M – estimators are divided into three categories (Chen and Yin, 2002) i.e: *monotone* – influence function ψ is a monotone function (e.g. Huber), *soft redescenders* – influence function ψ decreases asymptotically to 0 with increasing $|e_i|$ (e.g. Cauchy), *hard redescenders* – influence function ψ is 0 for large $|e_i|$ (e.g. Tukey, 1962; Hampel, 1974).
To make (1) minimum, it is required to solve the following equation:

$$\sum_{i=1}^{n} a_{i,j} \psi\left(l_i - \mathbf{a}_i^T \mathbf{x}\right) = 0 \qquad j = 1, \ldots, p \tag{2}$$

where $\psi(e)$ is the first derivative of the objective function with respect to residuals e, i.e. $\dfrac{\partial \rho(e)}{\partial e}$ and is called the influence function (p – number of parameters to be estimated). The last function, particularly important and characteristic for M – estimators is the weighting function and is defined as:

$$w(e) = \frac{\psi(e)}{e} \qquad (3)$$

That is, i-th observation with larger residual $|e_i|$ has smaller $w(e_i)$. Additionally, $w(e_i)$ should approach zero if $|e_i|$ is infinitely large. Weighting function should satisfy the following conditions:
– $w(e)$ is continuous, symmetric (even function)
– $w(e)$ decreases when $|e|$ increases
–   equal to one when its argument is zero $w(0) = 1$
– $\lim_{e \to \infty} w(e) = 0$

Weighting functions with other properties are considered in Wiśniewski (2014).
Rewriting expression (2) with the use of weighting function results in:

$$\sum_{i=1}^{n} a_{i,j}\, w_i \left(l_i - \mathbf{a}_i^T \mathbf{x}\right) = 0 \qquad j = 1, \ldots, p \qquad (4)$$

where $w_i = w\left(l_i - \mathbf{a}_i^T \mathbf{x}\right)$

$$\sum_{i=1}^{n} a_{i,j}\, w_i \left(l_i - \mathbf{a}_i^T \mathbf{x}\right) = 0 \Rightarrow \sum_{i=1}^{n} a_{i,j}\, w_i\, l_i = \sum_{i=1}^{n} a_{i,j}\, w_i\, \mathbf{a}_i^T \mathbf{x} \qquad j = 1, \ldots, p \qquad (5)$$

In matrix notation (5) may be rewritten as:

$$\mathbf{A}^T \mathbf{P} \mathbf{A}\, \mathbf{x} = \mathbf{A}^T \mathbf{P} \mathbf{L} \qquad (6)$$

where the weighting matrix $\mathbf{P}$ consists of w_i entries defined beforehand.
Equation (6) leads to the well known solution of weighted least squares:

$$\hat{\mathbf{x}} = \left(\mathbf{A}^T \mathbf{P} \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{P} \mathbf{L} \qquad (7)$$

The optimization problem defined by (1) is usually carried out by means of *iteratively reweighted least squares* (IRLS). This method is usually chosen because of its mathematical simplicity and common understanding i.e. it is easy to implement in the standard least squares framework. The IRLS method relies on the use of adaptive weights suppressing the influence of observations with large values of residuals in subsequent iterations (for details consult e.g. Draper and Smith, 1998; Wiśniewski 2009).

Since the residual variance is strongly affected by outlying observations, in this study, besides the latter mentioned the following scaling factors (variance factors) have been used (Wilcox, 2005; Duchnowski, 2011).

Inter-quartile range

$$IQR(\hat{\mathbf{e}}) = \hat{\mathbf{e}}_{0.75} - \hat{\mathbf{e}}_{0.25} \tag{8}$$

where: subscripts 0.75 and 0.25 denote the third and first quartiles; respectively

Normalized Median Absolute Deviation

$$MADN(\hat{\mathbf{e}}) = \frac{MAD(\hat{\mathbf{e}})}{0.6745} \tag{9}$$

3. M – estimators used in this study

Besides the well known M – estimators i.e. Huber, Hampel and Tukey's we introduce new redescending M – estimators (hard redescenders). Table 1 shows Huber, Tukey and Hampel's weighting functions (both in descriptive and graphical form) and influence function (only graphical representation). The weighting functions for the new M – estimators were motivated by weighting functions applied in orthogonal polynomials theory, kernel density estimation as well as one derived from Wigner semicircle probability distribution. Table 2 presents the underlying functions that were the base for derivation of the new weighting functions. The new weighting functions adopted names from the underlying functions.

Table 1. Weighting and influence function for Huber, Tukey and Hampel M – estimators
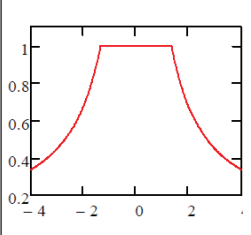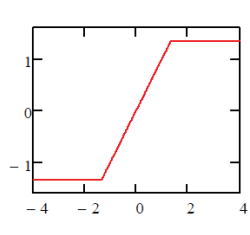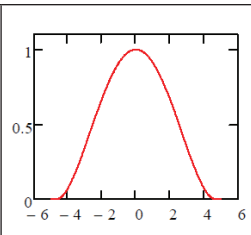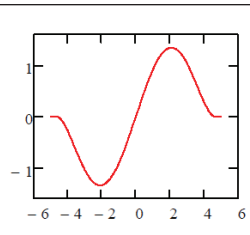
M – estimator (*weighting function*)	Weighting Function	Influence Function
Huber		
$w(\hat{e}_{scaled_i}) = \begin{cases} 1 & for \quad \left| \hat{e}_{scaled_i} \right| \le t^c \\ \dfrac{t^c}{\left| \hat{e}_{scaled_i} \right|} & for \quad \left| \hat{e}_{scaled_i} \right| > t^c \end{cases}$		
Tukey		
$w(\hat{e}_{scaled_i}) = \begin{cases} \left[1 - \left(\dfrac{\hat{e}_{scaled_i}}{t^c} \right)^2 \right]^2 & for \quad \left| \hat{e}_{scaled_i} \right| \le t^c \\ 0 & for \quad \left| \hat{e}_{scaled_i} \right| > t^c \end{cases}$		

Table 1.
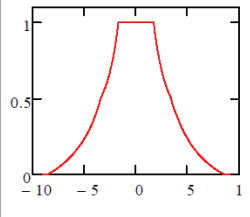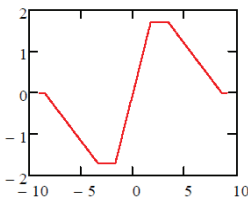
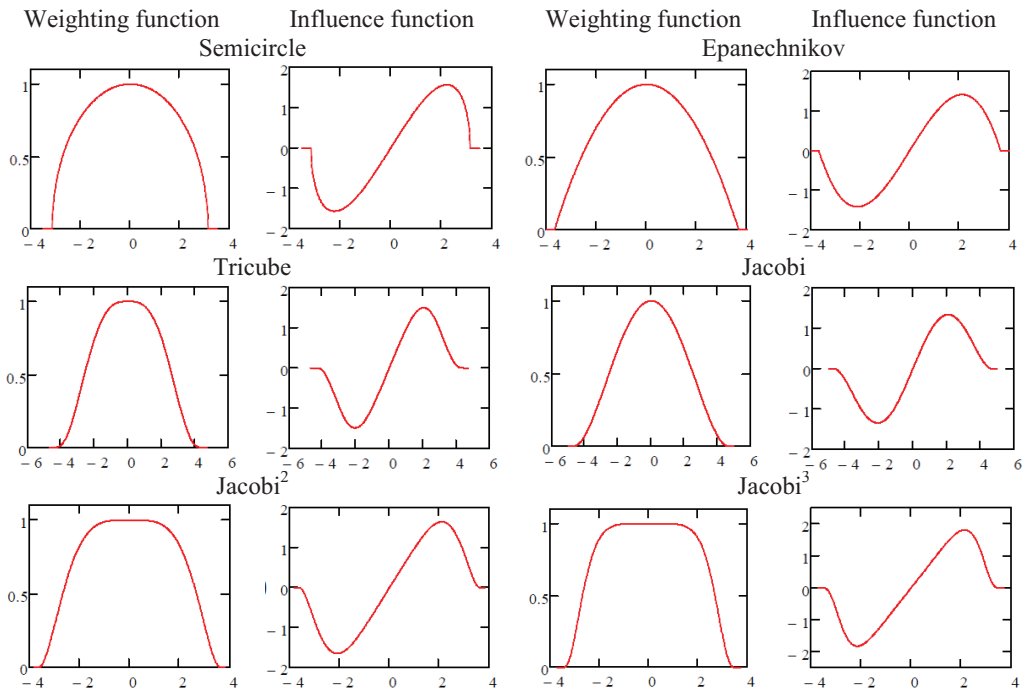| M – estimator (*weighting function*)	Weighting Function	Influence Function
Hampel		
$w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases} 1 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert \le t_1^c \\[2mm] \dfrac{t_1^c}{\left\lvert\hat{e}_{scaled_i}\right\rvert} & for \ t_1^c < \left\lvert\hat{e}_{scaled_i}\right\rvert \le t_2^c \\[2mm] t_1^c\,\dfrac{t_3^c-\left\lvert\hat{e}_{scaled_i}\right\rvert}{\dfrac{t_3^c-t_2^c}{\left\lvert\hat{e}_{scaled_i}\right\rvert}} & for \ t_2^c < \left\lvert\hat{e}_{scaled_i}\right\rvert \le t_3^c \\[2mm] 0 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert > t_3^c \end{cases}$		

Table 2. The underlying functions and new derived weighting functions

Underlying function	Derived weighting function
Wigner semicircle distribution $f(x)=\dfrac{2}{\pi R^2}\sqrt{R^2-x^2}$ $-R \le x \le R,\ f(x)=0 \text{ if } R < \lvert x\rvert$ (Wigner, 1955)	Semicircle weighting function $w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}\sqrt{1-\left(\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2} & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert \le t^c \\[3mm] 0 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert > t^c \end{cases}$
Epanechnikov kernel $K(x)=\dfrac{3}{4}\left(1-x^2\right)\mathbf{1}_{\{\lvert x\rvert \le 1\}}$ $\mathbf{1}_{\{\dots\}}$ is the indicator function (Epanechnikov, 1969)	Epanechnikov weighting function $w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}1-\left(\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert \le t^c \\[3mm] 0 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert > t^c \end{cases}$
Tricube kernel $K(x)=\dfrac{70}{81}\left(1-\lvert x\rvert^3\right)^3\mathbf{1}_{\{\lvert x\rvert \le 1\}}$ $\mathbf{1}_{\{\dots\}}$ is the indicator function (Hastie et al., 2009)	Tricube weighting function $w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}\left(1-\left(\dfrac{\lvert\hat{e}_{scaled_i}\rvert}{t^c}\right)^3\right)^3 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert \le t^c \\[3mm] 0 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert > t^c \end{cases}$
Jacobi (orthogonal) polynomials with respect to weighting function $w(x)=(1-x)^\alpha(1+x)^\beta$ $\alpha=\beta$ – Ultraspherical polynomials $\alpha,\ \beta=0$ – Legendre polynomials this coincides with least squares (Davis, 1963)	Jacobi weighting function $w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}\left(1-\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2\left(1+\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert \le t^c \\[3mm] 0 & for \ \left\lvert\hat{e}_{scaled_i}\right\rvert > t^c \end{cases}$

Table 2.

Underlying function	Derived weighting function
As above	Jacobi² weighting function $$w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}\left(1-\left(\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2\right)^2\left(1+\left(\dfrac{\hat{e}_{scaled_i}}{t^c}\right)^2\right)^2 & for\ \left|\hat{e}_{scaled_i}\right|\le t^c\\[2mm]0 & for\ \left|\hat{e}_{scaled_i}\right|> t^c\end{cases}$$
As above	Jacobi³ weighting function $$w\!\left(\hat{e}_{scaled_i}\right)=\begin{cases}\left(1-\left(\dfrac{\left|\hat{e}_{scaled_i}\right|}{t^c}\right)^3\right)^3\left(1+\left(\dfrac{\left|\hat{e}_{scaled_i}\right|}{t^c}\right)^3\right)^3 & for\ \left|\hat{e}_{scaled_i}\right|\le t^c\\[2mm]0 & for\ \left|\hat{e}_{scaled_i}\right|> t^c\end{cases}$$

Table 3. Graphical representation of weighting and influence functions for the new M – estimators

To save some space we do not present objective and influence functions for the new M – estimators. Influence functions may easily be derived from weighting functions by using formula (3). On the other hand, objective functions may be found by taking the antiderivatives of influence functions and shifting them by a constant to satisfy $\rho(0) = 0$ when necessary. Shapes of weighting and influence functions for the new M – estimators as being more informative than formulas are presented in Table 3.

All the estimators are endowed with theoretical tuning constants which give 95% efficiency with respect to the standard normal distribution. The tuning constants were computed numerically from the following formula (Huber, 1981; Shevlyakov et al., 2008):

$$eff = \frac{\left[\int\limits_{-t^c}^{t^c}\psi'(x)f(x)dx\right]^2}{\int\limits_{-t^c}^{t^c}[\psi(x)]^2 f(x)dx} \cong 0.95 \tag{10}$$

where:

eff – stands for efficiency
$x \sim N(0, 1)$
$f(x)$ – probability density of the standard normal distribution
ψ – influence function for any M – estimator

Table 4. Tuning constants for the new M – estimators assuring 95% efficiency with respect to the standard normal distribution

Semicircle	Epanechnikov	Tricube	Jacobi	Jacobi2	Jacobi3
$t^c = 3.137$	$t^c = 3.674$	$t^c = 4.417$	$t^c = 4.687$	$t^c = 3.618$	$t^c = 3.492$

For the M – estimators of Huber and Tukey tuning constants were adopted as 1.345, 4.685 respectively and for the Hampel's estimator three tuning constants were 1.7, 3.4, 8.5 (Hogg, 1979).

4. Numerical example

This simple numerical example is derived from Ghilani (2010) (original units: feet and miles are maintained). The original leveling network (Fig. 1 and Table 5) was sequentially contaminated with blunders. In the first test every single observation (measured height difference) was burdened with a 1 foot gross error i.e. 7 different contaminated models were tested overall (7 contaminated models × 9 weighting functions × 3 scaling factors = 189 cases). In the second test two observations at a time were contaminated with a 1 foot gross error (in a sequence 1–3, 1–5, 1–7, 2–4,

2–6, 3–5, 3–7, 4–6, 5–7) resulting in 9×9×3=243 cases in total. The overall number of tests (432) provides an informative insight how the methods work under different scenarios.

In this place, it is worth recalling two quantities that appear to be quite informative in blunder detection. The first one is the redundancy number (global and local) whilst the other is a correlation matrix for residuals. The global redundancy number is equal to $R = n - p$ i.e. the number of extra observations in the model. The local redundancy numbers are the diagonal elements of the matrix (difference between the unit matrix **I** and the "hat" matrix **H**):

$$\mathbf{R} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{A}\left(\mathbf{A}^T\mathbf{P}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{P} \tag{11}$$

thus, individual local redundancies read:

$$r_i = 1 - h_{ii} \tag{12}$$

In an uncorrelated case these values sum up to the global redundancy number and are all in the interval $0 \leq r_i \leq 1$. Individual values of local redundancy inform about detectability of blunders on individual observations and are firmly related to the reliability of geodetic networks. Large local redundancy number r_i (close to unity) means that a blunder greatly affects the residual and is easy to detect, on the other hand small value (approaching zero) of this local measure means that an outlier has small impact on the residual and will be hard to detect (Ghilani, 2010). The second quantity, the residual correlation matrix gives an answer as to which two residuals are significantly correlated and what is going further if a blunder occurs on one observation from the pair the other residual will be influenced by this blunder as well. Large values of correlation coefficients for the residuals may indicate which residuals are subject to masking (a bad observation becomes a good one) or swamping (a good observation becomes a bad one) effect. The correlation matrix may be easily derived from covariance matrix for the residuals $\mathbf{C}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}$ according to the following formula:

$$Corr(\hat{\mathbf{e}}) = \mathbf{D}\mathbf{C}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}\mathbf{D} \tag{13}$$

where **D** is a diagonal matrix containing the inverses of square roots of diagonal elements of $\mathbf{C}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}$ matrix (standard deviations for individual residuals). Hence, correlation coefficient between $i - th$ and $j - th$ residuals expressed by means of diagonal and off – diagonal entries of the covariance matrix reads:

$$\hat{\rho}_{ij} = \frac{\left(c_{\hat{\mathbf{e}}\hat{\mathbf{e}}}\right)_{ij}}{\sqrt{\left(c_{\hat{\mathbf{e}}\hat{\mathbf{e}}}\right)_{ii}\left(c_{\hat{\mathbf{e}}\hat{\mathbf{e}}}\right)_{jj}}} \tag{14}$$

The latter coefficient will particularly be helpful in explaining why all M – estimators considered in this study fail to detect outliers in some instances of this example. For more information on undetectability of gross errors we refer the reader to (Kwaśniak 2012; Prószyński 1997, 2010).

Tests were mainly performed with previously derived "theoretical" tuning constants but some cases were also checked with arbitrary ones (trial and error method) but main emphasis is put on the first mentioned. Table 6 presents simplified results (without error analysis) of the adjustment of the test leveling network without blunders, thus it constitutes the reference for further analysis. Table 7 reveals mutual correlations between residuals as well as redundancy numbers, the quantities which will be useful in explaining results of the adjustment. Table 8 introduces list of residuals obtained from least squares adjustment of the leveling net for "one – blunder" contaminated sets i.e. 7 sets of residuals. Table 9 presents maximum and average differences between adjusted heights (model parameters) from "clean" (without blunders) least squares adjustment and "contaminated" least squares adjustment for "one – blunder" contaminated sets. These two measures i.e. maximum and average difference will be repeatedly reported (in figures) in the part of this numerical example devoted to results of M – estimation. Tables 10 and 11 present the same quantities as revealed in Tables 8 and 9 but for the case of "two – blunders" contamination. From tables 10 and 7 it is immediately visible how the correlation between residuals might influence other residuals (innocent ones), this is marked with bold – face font. This is clearly visible for pairs 3-5 and 3-7 where M – estimation failed to detect blunders. For the contaminated pairs 3-5 and 3-7 innocent observation no. 4 obtained the highest values of residuals and as the M-estimators are based on individual residuals only and do not take into account interrelations between them. It is the main reason of their usual failure in these cases, but as shown in Figures 8, 9 (case of Tricube weighting function), Figures 10, 11 (Jacobi3 weighting function, to some extent) and Figures 12, 13 (cases of Semicircle, Epanechnikov, Tricube and Jacobi2 weighting functions) even then by some coincidence it may give proper results.
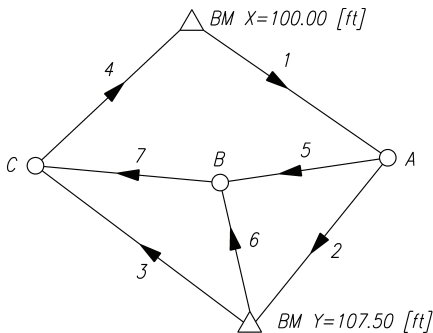
Fig. 1. Tested leveling net

Table 5. Observed elevation differences and weights used in adjustment

Line	Elevation difference [ft]	Length [miles]	Relative weights
1	5.10	4	3
2	2.34	3	4
3	−1.25	2	6
4	−6.13	3	4
5	−0.68	2	6
6	−3.00	2	6
7	1.70	2	6

Table 6. Simplified results of the adjustment of the leveling network
(only adjusted heights and residuals)

Adjusted Elevations						
A		B			C	
105.1504		104.4892			106.1972	
Residuals						
1	2	3	4	5	6	7
-0.0504	-0.0096	0.0528	0.0672	-0.0188	0.0108	-0.0080

Table 7. Correlation matrix for the residuals and the local redundancy numbers
(correlation coefficients around or greater than the value of 0.4 are marked with a bold – face font)

Obs.	1	2	3	4	5	6	7	Redundancy number
1	1.000	**0.481**	-0.089	0.065	**0.454**	-0.242	0.167	0.720
2	**0.481**	1.000	0.110	-0.080	**-0.562**	0.300	-0.206	0.627
3	-0.089	0.110	1.000	**0.571**	-0.196	-0.316	**-0.535**	0.560
4	0.065	-0.080	**0.571**	1.000	0.143	0.230	**0.389**	0.707
5	**0.454**	**-0.562**	-0.196	0.143	1.000	**-0.534**	0.367	0.404
6	-0.242	0.300	-0.316	0.230	**-0.534**	1.000	**0.591**	0.538
7	0.167	-0.206	**-0.535**	0.389	0.367	**0.591**	1.000	0.444

Table 8. Residuals obtained from least squares adjustment for "one – blunder" contaminated sets

Obs.	1	2	3	4	5	6	7
1	**0.670**	**0.323**	-0.130	0.003	**0.296**	-0.264	0.083
2	**0.270**	**0.617**	0.070	-0.063	**-0.356**	0.204	-0.143
3	0.013	0.106	**0.613**	0.346	-0.040	-0.120	-0.214
4	0.107	0.014	**0.507**	**0.774**	0.160	0.240	**0.334**
5	**0.154**	**-0.250**	-0.112	0.043	**0.386**	**-0.268**	0.137
6	-0.096	0.153	-0.162	0.126	**-0.238**	**0.548**	**0.300**
7	0.059	-0.097	**-0.275**	**0.170**	0.148	**0.281**	**0.436**

Values presented in Tables 7 and 8 are worth comparing, it is immediately visible how the correlation between residuals influences residuals themselves. For example taking the residual on the first observation one notices its strong correlation with other two i.e. numbered 2 and 5 (bold face font in Table 7). Confronting this with the values from Table 8 one observes a kind of residual flow between correlated ones i.e. in general; stronger correlated ones obtain larger values of residuals. One may also notice that the effect of a blunder on the corresponding residual is proportional to

the redundancy number i.e. compare redundancy numbers from Table 7 and diagonal entries in Table 8. An explanation of this proportionality is described in e.g. Ghilani (2010).

Table 9. Maximum and average differences between adjusted heights (model parameters) from "clean" least squares adjustment and "contaminated" least squares adjustment for "one – blunder" contamination

	1	2	3	4	5	6	7
Max. diff.	0.2800	0.3733	0.4400	0.2933	0.3467	0.4622	0.2889
Ave. diff.	0.1422	0.1896	0.2311	0.1541	0.2296	0.2830	0.2296

Table 10 Residuals obtained from least squares adjustment for "two – blunders" contaminated sets

Obs.	1-3	1-5	1-7	2-4	2-6	3-5	3-7	4-6	5-7
1	0.5896	1.0163	0.8029	0.3763	0.1096	0.2163	0.0029	-0.2104	0.4296
2	0.3504	-0.0763	0.1371	0.5637	0.8304	-0.2763	-0.0629	0.1504	-0.4896
3	0.5728	-0.0805	-0.2539	0.3995	-0.0672	**0.5195**	**0.3461**	0.1728	-0.3072
4	0.5472	0.2005	0.3739	0.7205	0.1872	**0.6005**	**0.7739**	0.9472	0.4272
5	0.0612	0.5590	0.3101	-0.1877	-0.4988	**0.2923**	0.0434	-0.2055	0.5412
6	-0.2692	-0.3448	0.1930	0.2686	0.6908	-0.4114	0.1264	0.6641	0.0508
7	-0.2080	0.2142	0.5031	0.0809	0.1920	-0.1191	**0.1698**	0.4587	0.5920

Table 11. Maximum and average differences between adjusted heights (model parameters) from "clean" least squares adjustment and "contaminated" least squares adjustment for "two – blunders" contamination

	1-3	1-5	1-7	2-4	2-6	3-5	3-7	4-6	5-7
Max. diff.	0.4800	0.3556	0.3067	0.4267	0.3200	0.5333	0.7067	0.3467	0.4800
Ave. diff.	0.3733	0.1852	0.2119	0.3437	0.2000	0.4074	0.2919	0.2089	0.2933

The results of M – estimation will be listed for two cases (one outlier and two outliers) in the same scheme. Within each variant of a scaling factor i.e. residual variance, interquartile range and normalized MAD; the maximum difference and the average difference between the "clear" solution and the one "contaminated" obtained with M – estimation will be reported. The comment will be provided when necessary.

Case I – one outlier
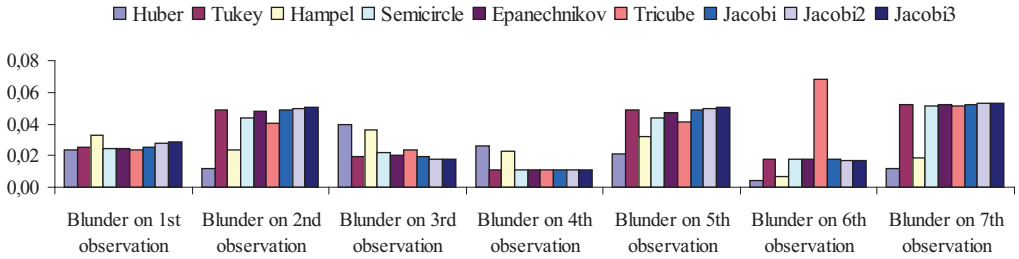
Scaling factor: residual variance

Fig. 2. Maximum absolute differences between the „clean" least squares solutions and M – estimations for contaminated datasets
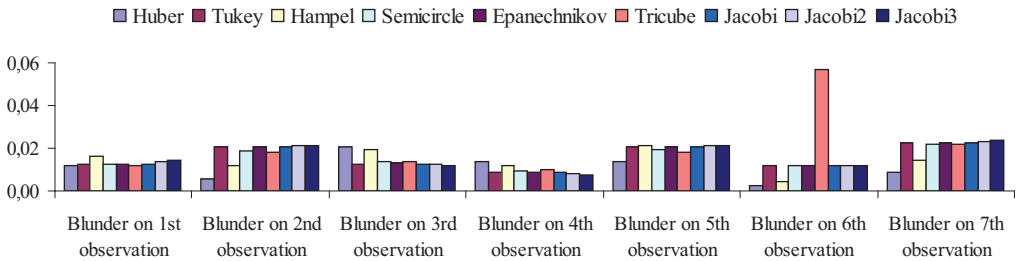

Fig. 3. Average absolute differences between the „clean" least squares solutions and M – estimations for contaminated datasets
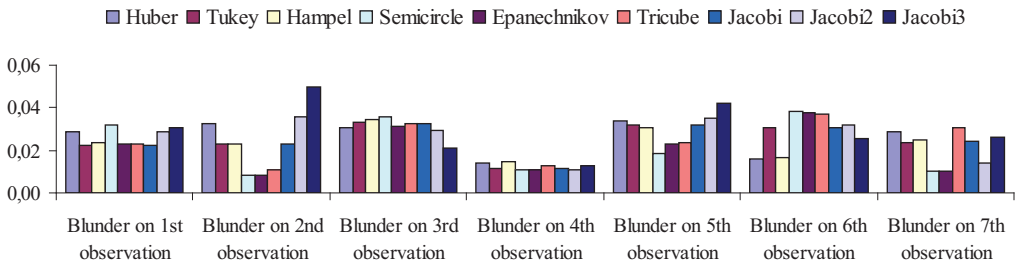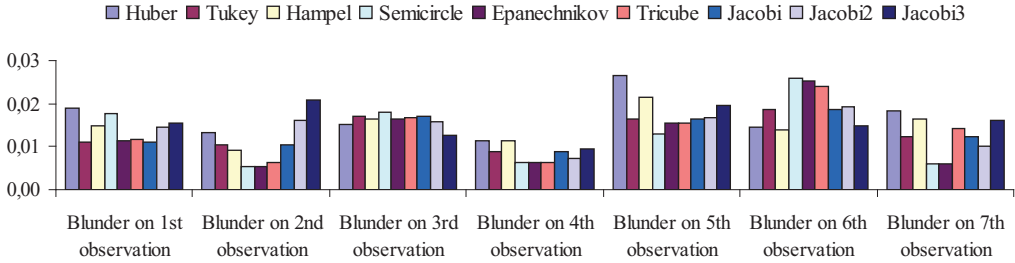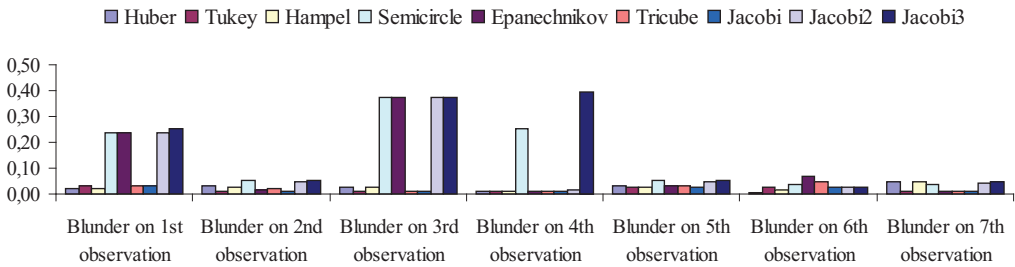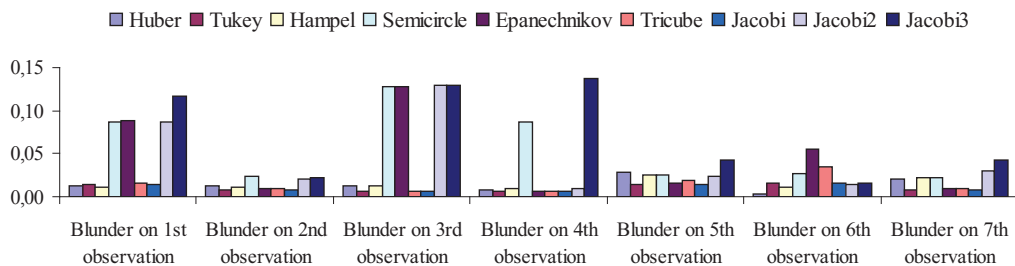
Scaling factor: interquartile range

Fig. 4. Maximum absolute differences between the „clean" least squares solutions and M – estimations for contaminated datasets

Fig. 5. Average absolute differences between the „clean" least squares solutions and M
– estimations for contaminated datasets

Scaling factor: normalized MAD

Fig. 6. Maximum absolute differences between the „clean" least squares solutions and M
– estimations for contaminated datasets

Fig. 7. Average absolute differences between the „clean" least squares solutions and M
– estimations for contaminated datasets

Despite the fact that the residual variance is considered as an improper scale factor
(it lacks robustness) it gave satisfactory results in this case. On the other hand in case
of the normalized MAD Semicircle, Epanechnikov, Jacobi2 and Jacobi3 weighting
functions gave unsatisfactory results as far as maximum and average difference

in model parameters (estimated heights) are concerned (residuals carry a trace of blunders detectability). The explanation of the above may be that the tuning constants were set up with respect to the standard normal distribution. And in case of different than residual variance scaling factor the scaled residuals may take values much higher than triple standard deviation and bounds determined by the tuning constants may be slightly overshot. This indicates that the choice of a tuning constant must be dependent on the adopted scaling factor in some reasonable way (this will be the subject of the authors' further researches). Additional tests (not presented in the content of the paper) carried out on arbitrary tuning constants selected by "trial and error method" proved that the number of iterations of each M – estimator may be reduced (in theory, for the cost of loss in efficiency with respect of a normal distribution) whilst maintaining satisfactory results. When comparing the maximum and average differences from Figures 2, 3 and 4, 5 and 6, 7 respectively one may notice their slight decrease (on average) in favour of M – estimators with the interquartile range as a scaling factor.

Case II – two outliers
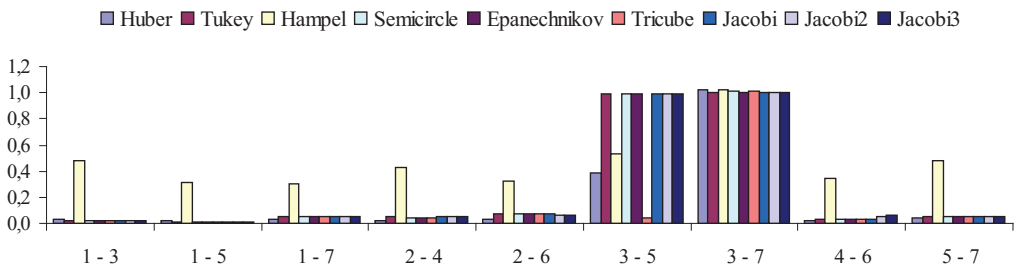
Scaling factor: residual variance

Fig. 8. Maximum absolute differences between the „clean" least squares solutions and M – estimations for contaminated datasets
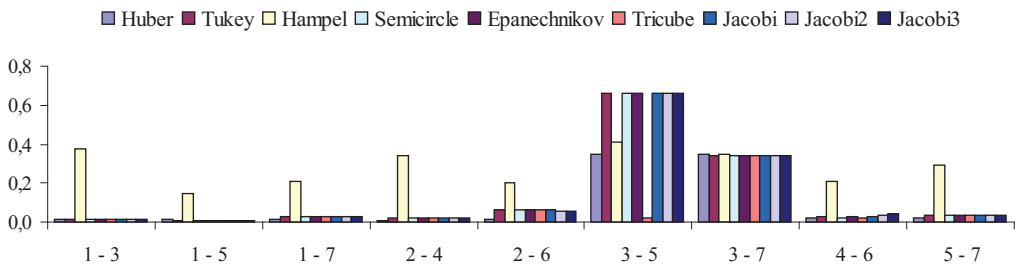

Fig. 9. Average absolute differences between the „clean" least squares solutions and M – estimations for contaminated datasets
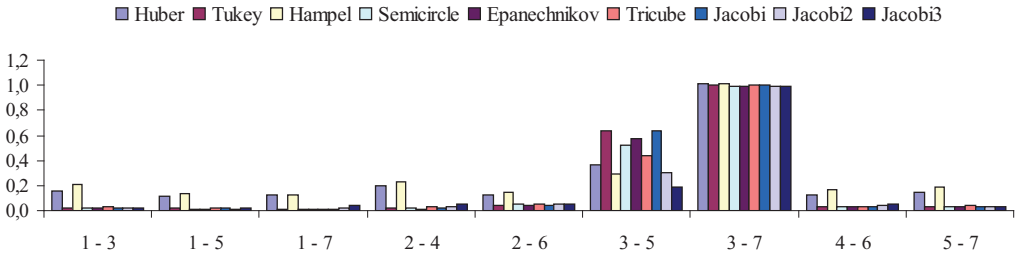
Scaling factor: interquartile range

Fig. 10. Maximum absolute differences between the „clean" least squares solutions and M
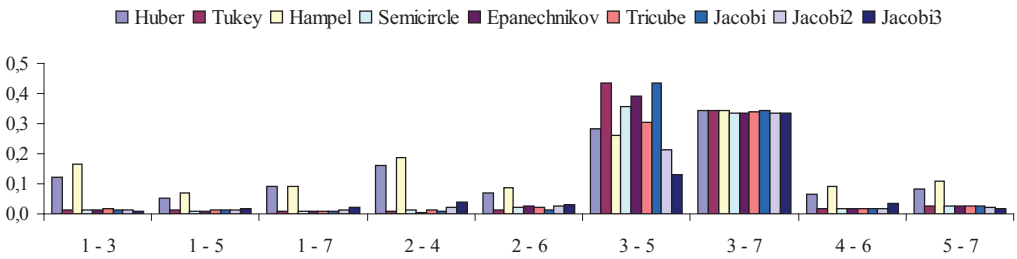– estimations for contaminated datasets

Fig. 11. Average absolute differences between the „clean" least squares solutions and M
– estimations for contaminated datasets
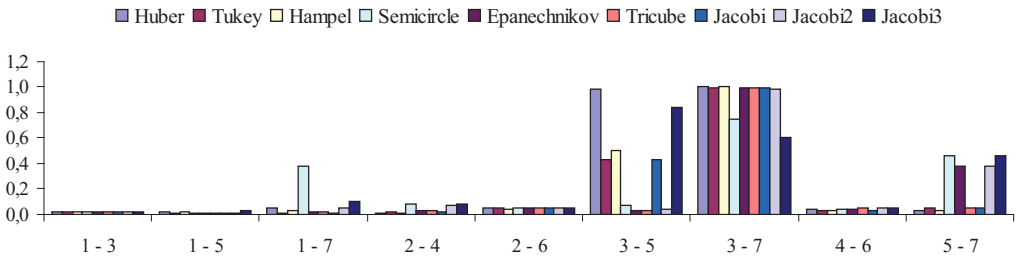
Scaling factor: normalized MAD

Fig. 12. Maximum absolute differences between the „clean" least squares solutions and M
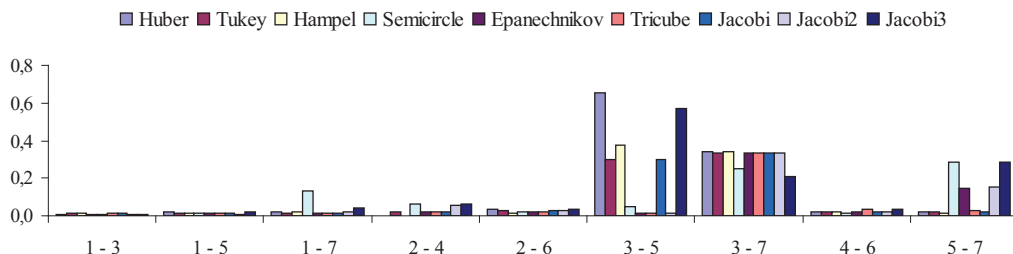– estimations for contaminated datasets

Fig. 13. Average absolute differences between the „clean" least squares solutions and M
– estimations for contaminated datasets

In case of residual variance scaling factor almost all estimators worked equally well. The exception is the Hampel's one that with adopted tuning constant immediately satisfied stopping criteria and fail to detect blunders. In all cases contaminated pair 3 – 7 was undetectable from the reasons described at the beginning of the numerical example section. For the pair 3 – 5 theoretically undetectable blunder was detected by one of the applied M – estimators (i.e. Tricube). For the case of interquartile range scaling factor two M – estimators had problems with proper detections i.e. the ones of Huber and Hampel. They gave unsatisfactory results as far as maximum and average differences in model parameters (estimated heights) are concerned but when one analyzes residuals blunders are quite visibly detected. The normalized MAD scaling factor may be considered as the best performing this time. It overcame almost all the problems of the above mentioned scaling factors. Additionally, a problematic pair 3 – 5 got four right solutions. The pair 3 – 7 remains unsolved in this case as well. As a summary, Table 12 presents the overall percentage of right blunder detections from all variants.

Table 12. The overall percentage of right blunder detections in all variants by the M – estimators used
in this study

Huber	Tukey	Hampel	Semicircle	Epanechnikov	Tricube	Jacobi	Jacobi2	Jacobi3
87.5%	87.5%	72.9%	81.3%	85.4%	91.7%	87.5%	85.4%	83.3%

Conclusions

Six new M – estimators (as far as the authors' knowledge is concerned) have been introduced in the paper and compared to the three well known with grounded position i.e. of Huber, Tukey and Hampel. If one expects the authors will select the one out of the nine tested as the best performing the authors will disappoint these expectations. Although the percentage of detectability of gross errors by each method is possible to obtain on the basis of presented example (simulated one with the knowledge what the result should be, and in reality no such knowledge is usually available); it is

only illusory because the performance of M – estimators is highly dependent on the dataset. As stated in the introduction this comparison, besides introducing new M – estimators aimed at showing imperfections of M – estimation due to dependence on model residuals only. Of course, in a large number of cases M – estimators are highly superior to the standard least squares technique. On the other hand a large number of M – estimators is dependent on the so-called tuning constant which inappropriately selected may also contribute to the failure of the entire procedure. Tuning constants may be selected either on theoretical grounds (e.g. to obtain a certain level of efficiency with respect to the standard normal distribution), data – driven (this approach is constantly being developed e.g. (Wang et al., 2007)) or just empirically by "trial and error" method. This form of arbitrariness is a drawback of M – estimation. In fact arbitrary choice of the tuning constant is able to limit the number of iterations in the reweighted least squares method. Also, since M – estimation is based on individual residuals only it is very vulnerable to masking and swamping effect, i.e. misidentification of observations – "innocent" as blunders and vice versa. The authors' future research will be focused on finding suitable data – driven tuning constants that involve the shape of weighting function, structure of data (scaling factor, sample size etc.) as well as on the use of additional information hidden in e.g. correlation matrix for residuals and redundancy numbers and probably other measures (geometry of the problem will be taken into account) which are very informative as to the issue of which observations are potentially the most endangered with misidentification effects.

Acknowledgments

References

Baarda, W. (1968). *A testing procedure for use in geodetic networks*. Publications on Geodesy – New Series, vol. 2, no. 5.

Box, G. (1953). Non-normality and tests on variances. *Biometrika*, 40, 318-335.

Chen, C. & Yin G. (2002). Computing the Efficiency and Tuning Constants for M-Estimation. Proceedings of the 2002 Joint Statistical Meetings, 478-482.

Davis, P. J. (1963). *Interpolation and Approximation*. New York: Blaisdell Publishing Company.

Draper, N. & Smith H. (1998). *Applied regression analysis*. 3th Edition, New York: John Wiley & Sons.

Duchnowski, R. (2011). Sensitivity of robust estimators applied in strategy for testing stability of reference points. EIF approach. *Geodesy and Cartography*, 60(2), 123-134

Dunnington, G. W. (1955). *Carl Friedrich Gauss: Titan of Science*. Washington: Mathematical Association of America.

Epanechnikov, V. A. (1969): Non-Parametric Estimation of a Multivariate Probability Density. Theory of Probability and Its Applications, 14(1), 153-158.

Ghilani, C. D. (2010). *Adjustment Computations – spatial data analysis*. 5th Edition, New Jersey: Wiley.

Hampel, F. R. (1971): A general definition of qualitative robustness. *The Annals of Mathematical Statistics*, 42 (6), 1887-1896.

Hampel, F. R. (1974): The influence curve and its role in robust estimation. *The Annals of Statistics*, 69 (346), 383-393.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edn, New York: Springer.

Hogg, R.V. (1979): Statistical robustness: one view of its use in applications today. *The American Statistician*, 33(3), 108-115.

Huber, P. J. (1964): Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), 73-101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics and Probability, 1, 221-233. Berkley: University of California Press.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons.

Kwaśniak, M. (2012). Badanie wpływu niezawodności wewnętrznej sieci geodezyjnej na efektywność wybranych podejść do wykrywania błędów grubych. *Prace Naukowe Politechniki Warszawskiej – Geodezja*, z. 49.

Pope, A. J. (1976). The Statistics of Residuals and the Detection of Outliers – NOAA Technical Report NOS 65 NGS 1. Rockville: United States Department of Commerce.

Prószyński, W. (1997). Measuring the robustness potential of the least-squares estimation: geodetic illustration. *Journal of Geodesy*, 71, 652-659.

Prószyński, W. & Kwaśniak M. (2002). *Niezawodność sieci geodezyjnych*. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.

Prószyński, W. (2010). Another approach to reliability measures for systems with correlated observations. *Journal of Geodesy*, 84, 547-556.

Rousseeuw , P. J. & Leroy A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons.

Shevlyakov, G., Morgenthaler S. & Shurygin A. (2008): Redescending M – estimators. *Journal of Statistical Planning and Inference*, 138, 2906-2918.

Stigler, S. M. (2010): The changing history of robustness. *The American Statistician*, 64(4), 277-281.

Tukey, J. W. (1960): A survey of sampling from contaminated distributions. In: *Contributions to Probability and Statistics*, I. Olkin et al. (ed.). Stanford: Stanford University Press.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33, 1-67.

Wang, Y.G., Lin X. & Zhu M., Bai Z., (2007): Robust estimation using the Huber function with a data – dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16(2), 1-14.

Wigner, E. (1955): Characteristic Vectors of Bordered Matrices with Infinite Dimensions. *The Annals of Mathematics*, 62(3), 548-564.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. San Diego: Elsevier Academic Press.

Wiśniewski, Z. (2009): *Rachunek wyrównawczy w geodezji*. Olsztyn: Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego.

Wiśniewski, Z. (2014). M-estimation with probabilistic models of geodetic observations. *Journal of Geodesy*, 88, 941-957.

Praktyczne porównanie kilku M – estymatorów

Marek Banaś[1], Marcin Ligas[2]

[1]Państwowa Wyższa Szkoła Techniczno-Ekonomiczna im. ks. B. Markiewicza w Jarosławiu
Instytut Inżynierii Technicznej,
ul. Czarnieckiego 16, 37-500 Jarosław
e-mail: marek.banas@pwste.edu.pl

[2]AGH Akademia Górniczo-Hutnicza
Wydział Geodezji Górniczej i Inżynierii Środowiska
Katedra Geomatyki,
al. A. Mickiewicza 30, 30-059 Kraków
e-mail: ligas@agh.edu.pl

Streszczenie

W artykule przedstawiono empiryczne porównanie trzech dobrze znanych M – estymatorów (Huber'a, Tukey'a oraz Hampel'a) jak również kilku nowych. Nowe estymatory motywowane były funkcjami wagowymi wykorzystywanymi w teorii wielomianów ortogonalnych, estymacji jądrowej oraz jeden motywowany przez funkcję gęstości „półokręgu" Wigner'a. Każdy z estymatorów został użyty do wykrywania obserwacji odstających w skażonych zbiorach danych. Obliczenia wykonano za pomocą „reważonej" metody najmniejszych kwadratów. Ze względu na fakt, iż wariancja resztowa (używana w konstrukcji macierzy kowariancyjnych) nie jest odpornym estymatorem skali, w testach wykorzystano również odporne miary takie jak: rozstęp ćwiartkowy oraz znormalizowane odchylenie medianowe. Testy wykonano na prostej sieci niwelacyjnej w dużej ilości wariantów ukazujących dobre i złe strony M – estymacji. Nowe estymatory zostały wyposażone w teoretyczne stałe odcinania zapewniające 95% efektywność względem standaryzowanego rozkładu normalnego. Kwestia rozwijania metod bazujących na stałych odcinania pochodzących z danych została również pokrótce poruszona.