OUTLIER DETECTION IN SURVEYING NETWORKS*

Edward Preweda [1]

[1] AGH University of Science and Technology, **Poland**

ABSTRACT

The paper refers to the robust estimation methods, which allows to eliminate outliers in surveying networks. Network adjustment is performed by the method of least squares. A key problem is the correct selection of weights, resulting from the different standard deviations of observations. In the case of gross errors their impact on the results of the alignment can be minimized by reducing the weight of outstanding observations. The second solution is the elimination of such observations as they were detected and re-alignment this network. In addition to the presentation of the well-known features, damping solution, iterative solution was presented based author idea. The calculation is illustrated on the one-dimensional random variable. Also presented the final results of the flat network adjustment by the proposed algorithm to eliminate outliers.

Keywords: outliner, surveying network, robust estimation

INTRODUCTION

During execution geodetic measurements or during data processing heavy errors may occur. They cause distortion of estimators obtained by least square method. If at the same time there are many gross errors, their detection is difficult, especially in networks of linear angle. In practice, there are different ways to search for outliers. Sometimes it better to align the network in stages, dividing it into smaller modules. Thanks to verify the conformity of observation in "local coverage". Another method is to eliminate the most "suspicious" observations. Eliminating it is based more on intuition than on concrete. Such variant alignment are unfortunately very time-consuming. The study used several methods of strong estimation, the general rule is to iterate over-weighting observations based on the analysis of the amendments to the observation. Also proposed proprietary solution, consisting of an iterative analysis of the impact of the elimination of individual observations on the value of the residual variance and ranked on the basis of observations for which there is a presumption of gross errors occur. On the basis of the established order, these observations are eliminated from the solution of the problem, until a stable value of the length of the confidence interval for the residual variance. Presents the results of network adjustment angular-linear copyrighted program using selected methods of estimation and robust method for removing gross error by the proposals described below.

* This work is financed from funds for science realized at AGH University of Science and Technology, allocated for the year 2014

ROBUST ESTIMATION

In the least squares method, once established weights remain unchanged until the end of solving the system of equations. This means that the standard deviations on the basis of which the weights are determined, is assigned to the 100 percent probability. The result of the solution is very strongly associated with the selection of weights, thus setting the stage for their values cannot make mistakes. However, in practice, the standard deviations are not known with 100% probability, so it is reasonable requests to change their values, for example, in terms of the corresponding confidence interval for a specified level of significance for the standard deviation. Such a solution, however without restricting the scope to which the weights are changed, is used in rough estimation. Its basis is the function used to over-weight. In the literature [1], [2], [3], [4], [6], [8] you will encounter a number of features designed for this purpose. Appropriate function should meet the following criteria:

- take positive values (weight cannot be negative),
- be an even (symmetric with respect to the axis of the function),
- achieve one and only one maximum for the parameter equal to 0 (for deviation equal to 0 weight is greatest),
- the derivative of a function must be a step (create this "threshold" for gross errors),
- convex surrounded maximum (the second derivative is less than 0).

After reviewing these criteria can be concluded that the density function of the normal distribution satisfies these conditions. However, the impact of very large errors in the case of decreasing very slowly. Others considered curves are functions of distributions developed by authors such as Cauchy, Welsch, Tukey, Huber and Andrew. These functions have the following form:

Cauchy:

$$f(x)=\frac{c^2\log\left(1+\left(\frac{x}{c}\right)^2\right)}{2};\quad f'(x)=\frac{x}{\left(1+\left(\frac{x}{c}\right)^2\right)^2};\quad f''(x)=\frac{1}{1+\left(\frac{x}{c}\right)^2};\qquad c>0 \qquad (1)$$

Welsch:

$$f(x)=\frac{c^2\left[1-\exp\left(-\left(\frac{x}{c}\right)^2\right)\right]}{2};\ f'(x)=x\exp\left[-\left(\frac{x}{c}\right)^2\right];\ f''(x)=\exp\left[-\left(\frac{x}{c}\right)^2\right];\ c>0 \ (2)$$

Tukey:

$$if\,|x| \le c\;\; f(x) = \frac{c^2\left[1-\left(1-\left(\frac{x}{c}\right)^2\right)\right]}{6};\; f'(x) = x\left[1-\left(\frac{x}{c}\right)^2\right]^2;\; f''(x) = \left[1-\left(\frac{x}{c}\right)^2\right]^2; c>0$$
$$if\,|x| > c\;\; f(x) = 0; \quad f'(x) = 0 \tag{3}$$

Huber:

$$if\,|x| \le c \qquad f(x) = \frac{x^2}{2}; \qquad f'(x) = x; \quad f''(x) = 1; \quad c>0 \tag{4}$$

$$if\,|x| > c \qquad f(x) = k\left(|x| - \frac{k}{2}\right); \quad f'(x) = k\cdot \mathrm{sgn}(x);\; f''(x) = \frac{k}{|x|}; \quad c>0 \tag{5}$$

Andrew:

$$if\,|x| \le c\pi \quad f''(x) = \frac{\sin\left(\frac{x}{c}\right)}{\left(\frac{x}{c}\right)}; \quad c>0$$
$$if\,|x| \le c\pi \quad f''(x) = 0; \quad c>0 \tag{6}$$

A similar to rought estimation method is Danish method, which is based on the intuitive idea that a large correction for the observation may indicate a fault load thereof thick. Alignment proceeds iteratively. After the n-th iteration for each observation verifies the criterion $\frac{|\hat{v}_i^n|}{\sigma_0^n}\sqrt{p_i} < c$, where:

$\hat{v}_i^n$ - amendment of the i-th observation in the n-th iteration,

σ_0^n - standard deviation calculated in the n-th iteration

p_i - weight output (fixed a'priori in the first iteration) for the i-th observation

c - constant in the range 1÷3, in a sense symbolizes accepted level of probability.

In subsequent iterations, if the criterion is met, the weight of the observation remains unchanged otherwise:

$$p_i^{n+1} = p_i^n \exp\left(\frac{1}{c}\times\frac{|\hat{v}_i^n|}{\sigma_0^n}\sqrt{p_i}\right) \tag{7}$$

This is not a statistical method and the author's opinion should only be used for the detection of gross errors in order to eliminate them. Detected outliers should be removed and re-alignment should be performed using a priori weights.

METHODS FOR DETECTION OF OUTLIERS

Gross error detection methods are based on mathematical statistics and observations resulting from the practice, and many empirical experience. In either case, it is extremely difficult to develop a universal tool in the sense of the utility that could handle any situation of outliers. Using only statistical methods, it is necessary to verify the relevant statistical hypotheses on the adopted level of significance. Fundamental difficulty in this case is to determine the value of this level, because it determines the test result. Too mild initial assumption, which will result in wide confidence interval that will not be detected all outliers. In turn, the strict assumption may cause the outliers will be treated ones that actually represent the test object or phenomenon. There is a risk of making a mistake I or type II.

Summary of the principle of detection methods and gross errors in the observations, for estimating parameters of the least squares method is shown in [8]. The most commonly used methods are: Baardy, Pope, Chen-Kavouras-Chrzanowski, Cross-Price, Dinga-Coleman, Ethrog's.

THE ESSENCE OF THE AUTHOR'S METHODS OF DETECTING GROSS ERRORS

In the case of gross error which in a given system is little or errors have a very large value, the detection is relatively straightforward. Definitely more difficult task is to detect outliers present in large numbers and to limit the assumed level of confidence. Using only statistical methods can easily make a mistake I or type II . Considering the problem in terms of the development of software application using several methods with the detection of outliers and at the same time giving the user the freedom to determine the level of confidence , basically boils down to the task to a problem that has not been resolved . Too many options available, and thus the possible solutions, it is of course justified in a scientific sense. For practitioners, however, may be at least embarrassing. These considerations have led the author to seek a solution that combines statistical models of practice resulting from its own experience in the field of geodetic leveling numerical methods.

The proposed solution is based on automatic analysis of the impact of the elimination of individual observations on the length of the confidence interval for the residual variance, which has an asymmetric distribution χ^2 . It should be specify how the above-mentioned tests, the level of significance, but its value will be critical in detecting outliers. An important factor determining the end of the iterative process in this case is to verify the hypothesis concerning the quotient of the length of the confidence intervals for the variance of residuals $\frac{\sigma_0^{2)i}}{\sigma_0^{2)i+1}}$ in two next iterations. In order to streamline and generalization of the calculation of the specific cases, the task is carried out by numerical methods using decomposition SVD (Singular Value Decomposition) and matrix pseudo-inverse [7]. In the first pass, takes place a ranking of observation for possible occurrence of gross errors. At first places are allocated observations whose potential removal has the greatest impact on reducing the length of the confidence interval. On the basis of the established order, these observations are removed from the solution of the problem, until a stable value of the length of the confidence interval for

the residual variance $\sigma_0^2 \}^n$. After removing each observation rank of the matrix of the normal equations is verified numerically for the possible occurrence of the defect (allows a solution by SVD) and the variance ratio test performed. The proposed method requires many calculations, however, give an effective and straightforward solution to the problem of detection of outliers. In the last stage the classic solution of the least squares method without gross errors.

VERIFICATION OF SELECTED METHODS OF ADJUSTMENT OF OBSERVATIONS WITH THICK ERRORS

The object of analysis is the actual angular-linear network shown in Figure 1 for illustrative purposes confidence ellipse for designated points are also included. The network is established to two points adopted for error-free. As a result of this network alignment method of least squares estimators obtained far in excess of the limit values. Interesting juxtaposition and alignment observations posted in Figures 2-screenshot from application written by the author and Table 1-also calculate by the same application.

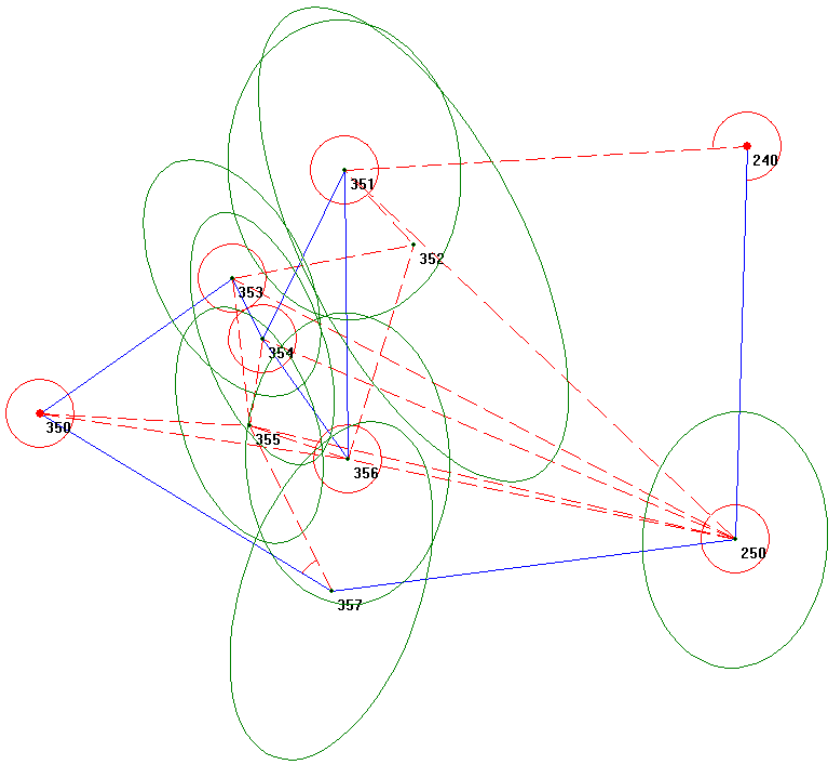
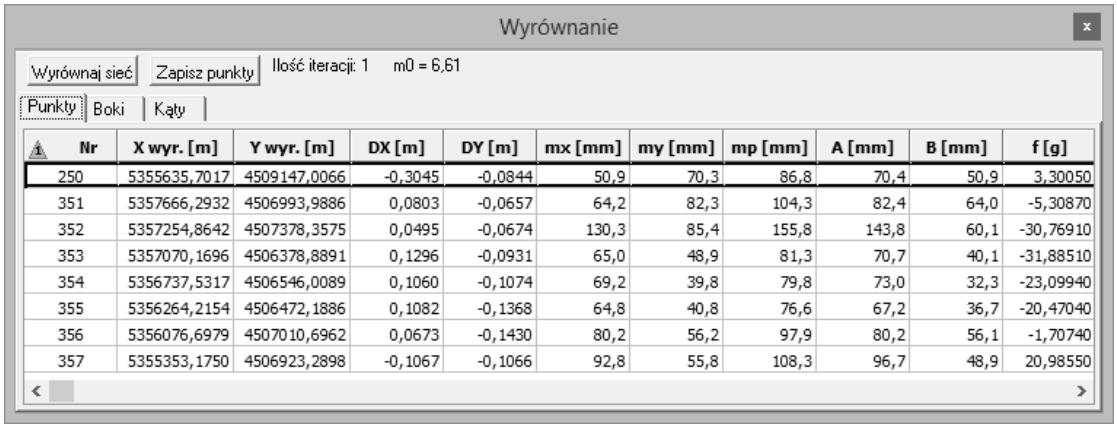


Figure 1. Sketch of angular-linear networks with confidence ellipses - leveling by LSM



Nr	X wyr. [m]	Y wyr. [m]	DX [m]	DY [m]	mx [mm]	my [mm]	mp [mm]	A [mm]	B [mm]	f [g]
250	5355635,7017	4509147,0066	-0,3045	-0,0844	50,9	70,3	86,8	70,4	50,9	3,30050
351	5357666,2932	4506993,9886	0,0803	-0,0657	64,2	82,3	104,3	82,4	64,0	-5,30870
352	5357254,8642	4507378,3575	0,0495	-0,0674	130,3	85,4	155,8	143,8	60,1	-30,76910
353	5357070,1696	4506378,8891	0,1296	-0,0931	65,0	48,9	81,3	70,7	40,1	-31,88510
354	5356737,5317	4506546,0089	0,1060	-0,1074	69,2	39,8	79,8	73,0	32,3	-23,09940
355	5356264,2154	4506472,1886	0,1082	-0,1368	64,8	40,8	76,6	67,2	36,7	-20,47040
356	5356076,6979	4507010,6962	0,0673	-0,1430	80,2	56,2	97,9	80,2	56,1	-1,70740
357	5355353,1750	4506923,2898	-0,1067	-0,1066	92,8	55,8	108,3	96,7	48,9	20,98550

Figure 2. Aligned coordinates - LSM

The position of the error in the network is not required to exceed 50 mm. The network has a weak structure, as evidenced by lines of constant probability density ellipses, there is an assumption about the possibility of the occurrence of outliers. First, strong estimation performed based on the selected schedule listed in the second section. Due to the extensive calculations and assumptions intermediate, final results are shown below (Table 1).

Table 1. Alignment by different algorithm

Alignment by Cauchy algorithm, m_0=1,67								
No	$\hat{X}$ [m]	$\hat{Y}$ [m]	σ_x [mm]	σ_Y [mm]	m_p [mm]	Average error ellipse		
						A [mm]	B[mm]	f [g]
250	5355635,8421	4509147,0762	60,8	22,4	64,8	61,0	21,9	5, 62940
351	5357666,2142	4506994,0389	21,7	22,3	31,1	23,51	20,4	43,29500
352	5357254,8058	4507378,4182	37,4	25,3	45,1	42,2	16,0	-33,44390
353	5357070,0426	4506378,9759	23,6	17,3	29,2	27,4	10,3	-36,98100
354	5356737,4269	4506546,1042	24,4	16,7	29,6	28,0	9,5	-35,06740
355	5356264,1091	4506472,3059	22,7	18,S	29,3	26,9	11,S	-40,38050
356	5356076,6262	4507010,8164	26,1	21,4	33,7	28,6	17,9	-35,22450
357	5355353,2282	4506923,3660	37,7	20,5	42,9	40,9	12,9	26,90440
Alignment by Welsch algorithm, m_0=1,05								
No	$\hat{X}$ [m]	$\hat{Y}$ [m]	σ_x [mm]	σ_Y [mm]	m_p [mm]	Average error ellipse		
						A [mm]	B[mm]	f [g]
250	5355636, 0224	4509147 0972	55,6	14 7	57,5	55,8	13 9	5, 80330
351	5357666,2113	4506994,0581	13,8	14,5	20,1	15,2	l3,1	38,93800
352	5357254,8157	4507378,4274	23,8	16,1	28,7	26,8	10,4	-32,99380
353	5357070,0384	4506378,9832	15,1	11,1	18,7	17,5	6,5	-37,16940
354	5356737,4239	4506546,1176	15,6	10,9	19,0	18,0	6,3	-35,70070
355	5356264,1059	4506472,3271	14,5	12,4	19,1	17,4	7,9	-42,36560
356	5356076,6287	4507010,8413	16,7	14,4	22,0	18,4	12,2	-37,59160
357	5355353,2929	4506923,3997	27,3	14,6	31,0	29,9	8,1	27,88040
Alignment by Tukey algorithm, m_0=1,01								
No	$\hat{X}$ [m]	$\hat{Y}$ [m]	σ_x [mm]	σ_Y [mm]	m_p [mm]	Average error ellipse		
						A [mm]	B[mm]	f [g]
250	5355636, 0224	4509147 0972	54,	14, 3	56 3	54 7	13 5	5,80720
351	5357666,2113	4506994,0581	13,4	14,1	19,5	14,8	12,7	38,67520
352	5357254,8157	4507378,4274	23,1	15,6	27,9	26,0	10,1	-32,96950
353	5357070,0384	4506378,9832	14,6	10,8	18, 1	17,0	6,3	-37,17420
354	5356737,4239	4506546,1176	15,1	10,6	18,5	17,4	6,1	-35,72080
355	5356264,1059	4506472,3271	14,1	12,l	18,5	16,9	7,7	-42,44460
356	5356076,6287	4507010,8413	16,2	14,0	21,4	17,8	11,8	-37,68500
357	5355353,2929	4506923,3997	26,6	14,3	30,2	29,1	7,9	27,91000
Alignment by Huber algorithm, m_0=1,01								
No	$\hat{X}$ [m]	$\hat{Y}$ [m]	σ_x [mm]	σ_Y [mm]	m_p [mm]	Average error ellipse		
						A [mm]	B[mm]	f [g]

250	5355636,0307	4509147,0980	54,5	14,3	56,3	54,7	13,5	5,80730
351	5357666,2112	4506994,0589	13,4	14,1	19,5	14,7	12,7	38,66670
352	5357254,8163	4507378,4278	23,1	15,6	27,9	26,0	10,1	-32,96870
353	5357070,0383	4506378,9835	14,6	10,7	18, 1	17,0	6,3	-37,17440
354	5356737,4239	4506546,1181	15,1	10,6	18,5	17,4	6,1	-35,72140
355	5356264,1059	4506472,3279	14,1	12,0	18,5	16,9	7,7	-42,44710
356	5356076,6289	4507010,8423	16,2	14,0	21,4	17,8	11,8	-37,68800
357	5355353,2957	4506923,1012	26,6	14,3	30,1	29,1	7,8	27,91090

Below are the three stages of the calculations according to the method author. Based on the ranking of observation due to the possibility of errors thick was placed on the first places of the two angles (357-350-353) and (356-350-357) and a length of between 240-250 points. In this order, these findings were eliminated, resulting in subsequent stages of the results shown in Figure 3.
Step 1 - remove the observation angle (357-350-353):
Step 2 - Removing the observation angle (357-350-353) and (356-350-357)
Step 3 - remove the above mentioned observation angle and length (240-250)

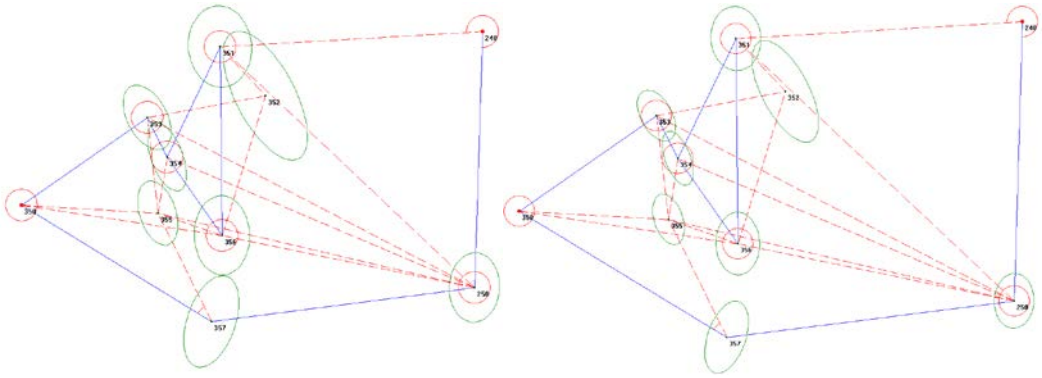
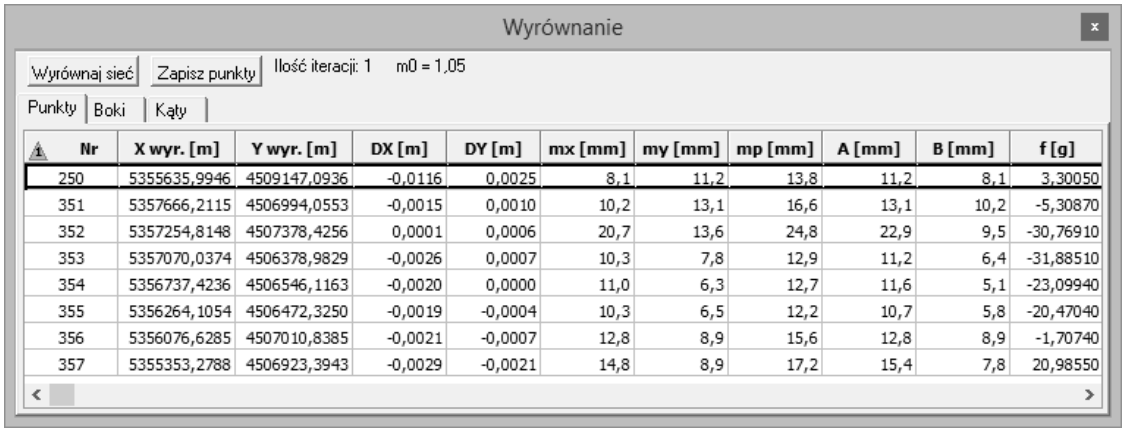


Figure 3. Sketch of angular-linear networks with confidence ellipses -the author idea - step 1 (left) and step 3 (right)

Subsequently angle was removed (356-350-357) in the third step the distance 240-250 was removed. The results are illustrated in Figure 3 (right side) and Figure 4.



Nr	X wyr. [m]	Y wyr. [m]	DX [m]	DY [m]	mx [mm]	my [mm]	mp [mm]	A [mm]	B [mm]	f [g]
250	5355635,9946	4509147,0936	-0,0116	0,0025	8,1	11,2	13,8	11,2	8,1	3,30050
351	5357666,2115	4506994,0553	-0,0015	0,0010	10,2	13,1	16,6	13,1	10,2	-5,30870
352	5357254,8148	4507378,4256	0,0001	0,0006	20,7	13,6	24,8	22,9	9,5	-30,76910
353	5357070,0374	4506378,9829	-0,0026	0,0007	10,3	7,8	12,9	11,2	6,4	-31,88510
354	5356737,4236	4506546,1163	-0,0020	0,0000	11,0	6,3	12,7	11,6	5,1	-23,09940
355	5356264,1054	4506472,3250	-0,0019	-0,0004	10,3	6,5	12,2	10,7	5,8	-20,47040
356	5356076,6285	4507010,8385	-0,0021	-0,0007	12,8	8,9	15,6	12,8	8,9	-1,70740
357	5355353,2788	4506923,3943	-0,0029	-0,0021	14,8	8,9	17,2	15,4	7,8	20,98550

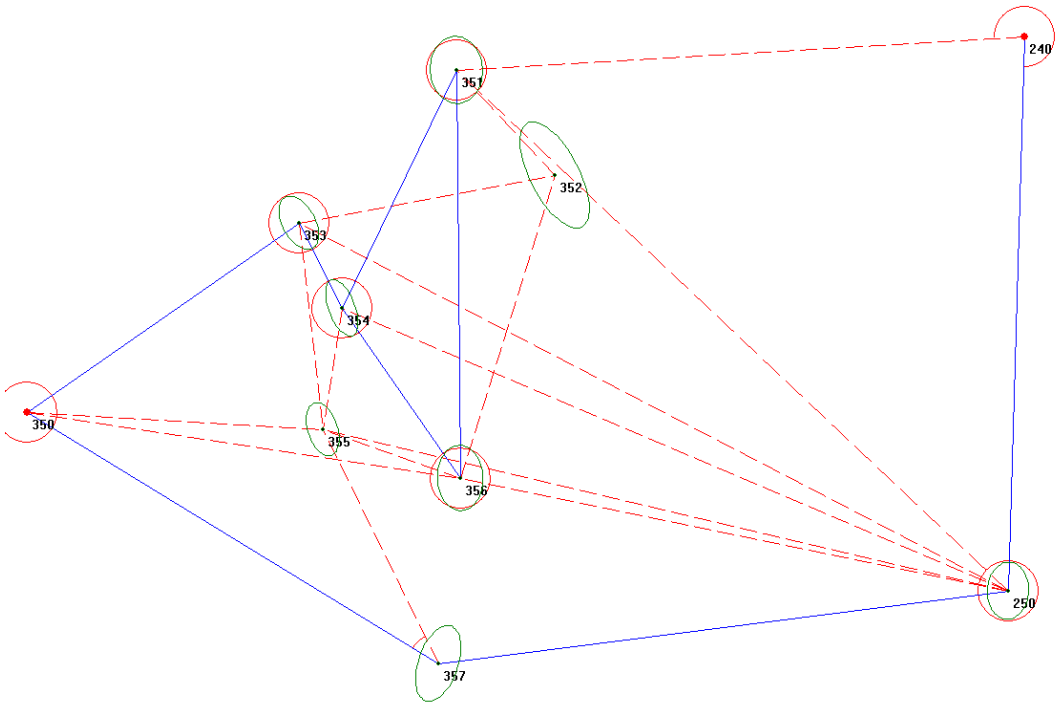


Figure 4. Aligned coordinates - algorithm by Edward Preweda- the screenshot of author's applications - Stage 3

CONCLUSIONS

One of the key factors affecting the results of parameter estimation method of least squares is the correct choice of weights of observation. In the case of gross errors, the results of the calculations are disturbed. It is very difficult to give a simple relationship, which would allow for the elimination of the influence of gross error. Estimation methods harder minimize the impact of outliers. As shown in the example, the use of different methods and with different levels of significance assumption leads to different results in terms of both the estimated expected value and the variance. The author proposes an iterative solution based on the analysis of the quotient of the residual variance in subsequent iterations, which admittedly requires many calculations, however, leads to a specific purpose, which is to eliminate the gross errors of a set of observations. The algorithm, although complex, is so unique that it is possible for the purpose of its software utility. Probably still requires specific testing under various conditions and minor modifications.

REFERENCES

[1] Baarda W.: 1968. A testing procedure for use in geodetic network. Netherlands Geodetic Commission, Publications on geodesy, Vol. 2, No.5, Delft, 97 pp.

[2] Caspary W.F.: 1988. Concept of Network and Deformation Analysis. Mon. 11, School of Surveying, University of New South Wales, Kensington, 183 pp.

[3] Ding X., Coleman R., 1996, Sensitivity Analysis in Gauss-Markov Models, *Journal of Geodesy*, Vol. 70 (8), pp.480-488.

[4] Kadaj R.: 1980: Rozwinięcie koncepcji niestandardowej metody estymacji. GiK, Vol. XXIX, nr 3/4.

[5] Koch K. R.: 1987. Parameter Estimation and Hypothesis Testing in Linear Models. Springer, New York, 378 pp.

[6] Pope A. J.: 1976. The Statistics of Residuals and the Detection of Outliers. Tech. Rep. NOS65 NGS1, Rockville, Md, 617 pp.

[7] Preweda E.: 2013, Rachunek wyrównawczy ⇒ modele statystyczne [Adjustment computations ⇒ statistical models]. Kraków, PROGRES, 2013, 387 pp

[8] Prószyński W., Kwaśniak M.: 2002. Niezawodność sieci geodezyjnych. Oficyna Wydawnicza PW, Warszawa 144 pp.