Outlier Detection in GPS Networks with Fuzzy Logic and Conventional Methods

Ertan GÖKALP and Yüksel BOZ, Turkey

Key Words: Outlier, Conventional Method, Fuzzy Logic, Statistical Test, GPS

SUMMARY

It is assumed that the geodetic observations always have random errors and these errors are normally distributed. The measurements have random errors that deviate from the normal distribution are called as 'outliers'. Outlier investigation is one of the first interest areas of the statisticians and different approaches have been developed for this purpose so far. In the literature, there isn't a method as the best over all. In the conventional methods like 'Data Snooping (DS)', 'Tau', and 't' tests, outliers are determined iteratively by the statistical test theory and removed from the observation set. In the fuzzy logic approach, fuzzy set relations and the statistical tests are used together to detect the outliers. Fuzzy membership relations are used among the residuals and observation errors. In this study, commonly used conventional methods (statistical tests) and fuzzy logic method have been applied to several GPS networks with different characteristics. It has been aimed to find which algorithm is convenient among the methods used in this study for outlier detection in GPS observations. The baseline components ΔX, ΔY, and ΔZ of the GPS baselines have been taken as the measurements. The conventional methods have been applied to the networks at different significance levels. As the significance level gets greater, the sensitivity of the tests for outlier increase and more outliers appear. It is appropriate to determine the significance level regarding the number of the observations in the network. The capabilities of the fuzzy logic approach to be an alternative method have been examined. Also, it has been seen that the results of this method are completely dependent on the statistical tests. Unlike conventional methods, no measurement is removed from the observation set, thus the shape of the network isn't defected.

Outlier Detection in GPS Networks with Fuzzy Logic and Conventional Methods

Ertan GÖKALP and Yüksel BOZ, Turkey

1. INTRODUCTION

Outlier is an observation, which has a random error in a different distribution apart from the observation set. The random errors are small and inconspicuous. Since they do not have large magnitudes, they can be determined by statistical tests such as Data Snooping (DS), Tau, and t tests. These methods are conventional approaches to detect outliers and deal with the residuals. In Least Squares (LS) adjustment, an observation that fails the statistical test is removed from the observation set. In the adjustment that uses GPS-derived baseline components as observations, the baseline component identified as outlier and the other two baseline components are taken away from the observation set as well after each iteration. In order to consider more accurately the observation that has a test statistic close to the critical value of the statistical test, a fuzzy logic based method can be used with statistical tests. In the fuzzy logic approach, the final decision about outliers is given by means of the observation errors. Definition of the membership functions plays an important role in outlier detection. The outliers are obtained altogether at the end of the method; no observation is removed from the observation set.

2. CONVENTIONAL METHODS

2.1 Data Snooping

DS was suggested by Baarda and is applicable when the theoretical variance $\left(\sigma_0^2\right)$ of the observation of unit weight is known. In this study, since the theoretical variance is not known, the a priori variance $\left(s_0^2\right)$ obtained from loop closures using Ferrero equation is used instead of it. In the DS method, it is assumed that only one outlier is present in the observation set. In addition using this method iteratively, more than one outlier can be detected and their locations can be estimated (Berberan 1995).

Ferrero equation is as follows:

$$s_0 = \sqrt{\frac{w^T w}{9 n_t}} \tag{1}$$

where w is the closure of a triangle related to each baseline component, 9 is the number of the observations belongs to a triangle, n_t is the number of the triangles of GPS network (Gökalp and Boz 2004).

DS is realized using normalized residuals. Therefore, the test statistics are compared with the critical values from normal distribution table.

The test statistic is

$$T_i = \frac{(Pv)_i}{s_0\sqrt{(PQ_{vv}P)_{ii}}} \tag{2}$$

where v is the vector of residuals, P is the weight matrix of the observations, Q_{vv} is the cofactor matrix of the residuals.

The critical value is

$$q = N_{1-\alpha_0/2} = \sqrt{F_{1,\infty,1-\alpha_0}} = \sqrt{\chi^2_{1,\infty,1-\alpha_0}} \tag{3}$$

Here α_0 is the significance level, N represents the normal distribution, F represents the Fischer table, χ represents the Chi-Square table.

The significance level for a single observation α_0 is computed from

$$\alpha_0 = 1-(1-\alpha)^{1/n} \cong \alpha/n \tag{4}$$

where α is the total significance level and usually chosen as 5%, n is the number of observations (Biacs et al. 1990).

2.2 Tau Test

If the a priori variance is not known or a value cannot be assigned to it before adjustment, the a posteriori variance (m_0^2) calculated at the end of adjustment is used for outlier detection. The residuals normalized using the a posteriori variance are not normally distributed. They are in Tau (τ) distribution (Schwarz and Kok 1993).

$$m_0^2 = \frac{v^T Pv}{f} = \frac{v^T Pv}{n-u+d}$$

(5)

Here f is degree of freedom, n is the number of the observations, u is the number of unknown parameters, and d is the datum defect.

The test statistic is

$$T_i = \frac{(Pv)_i}{m_0\sqrt{(PQ_{vv}P)_{ii}}} \tag{6}$$

The critical value of τ table can be obtained as follows (Heck 1981):

$$q = \tau_{f,1-\alpha_0/2} = \sqrt{\frac{f \times t^2_{f-1,1-\alpha_0/2}}{f-1+t^2_{f-1,1-\alpha_0/2}}} \tag{7}$$

where t represents the student (t) table, and τ represents the Tau table.

2.3 The t Test

If an observation (l_i) includes an error (Δ_i), outlier detection using the a posteriori variance obtained from the invalid adjustment model is not appropriate. In this situation it is a more accurate approach to compute the m_0 value from the residuals that are free from the model errors.

$$\bar{m}_0^2 = \frac{1}{f-1}\left(f\, m_0^2 - \frac{v_i^2}{(Q_{vv})_{ii}} \right) \tag{8}$$

Here $\bar{m}_0^2$ is the a posteriori variance calculated from the residuals free from the model errors. The test statistics of the observations as follows:

$$T_i = \frac{(Pv)_i}{\bar{m}_0\sqrt{(PQ_{vv}P)_{ii}}} \tag{9}$$

The critical value of t table is

$$q = t_{f-1,1-\alpha_0/2} \tag{10}$$

2.4 Fuzzy Logic Method

The conventional methods are insufficient in case of existing more than one outlier in the observation set. As seen from relation between residuals and errors

$$v = -Qvv\,P\,\Delta \tag{11}$$

the magnitudes of residuals depend on observation errors. In equation (11), the multiplication $Qvv\,P$ equals to redundancy matrix R. The expanded form of R and equation (11) is as follows:

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = - \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_n \end{bmatrix} \tag{12}$$

A normal observation can be seen as an outlier due to other observation errors. In order to overcome this problem, the final decision about outlying observations is made using the observation errors. Fuzzy membership values are assigned to them and judgment about an observation to be an outlier or not are given according to the magnitudes of these membership values. The procedures to this stage are presented as follows (Sun 1994):

At initial stage, the residuals are separated into two main classes via a statistical test: abnormal residuals $C\{v_i\}$, which fail the test and normal residuals $D\{v_i\}$, which pass the test. Then, membership function values of the residuals are defined in one of these classes.

$$\mu_C(v_i) = \begin{cases} 0 & T_i \le q \\ \dfrac{1.0}{1.0 + \left(\dfrac{e}{T_i - q}\right)^2} & T_i > q \end{cases} \tag{13}$$

Here, e is the standardization component denoting the significant deviation of the test statistic from the critical value (Konak and Dilaver 1998). After determining the membership values of residuals in set C, corresponding values in set D are obtained utilizing from the complementation property of fuzzy sets.

$$\mu_D(v_i) = 1 - \mu_C(v_i) \tag{14}$$

In order to get the membership values of the observation errors, the redundancy matrix is used in addition to the membership values of residuals. Each row element in R is divided to the greatest value in the corresponding row.

$$\tilde{r}_{ij} = \frac{\left| r_{ij} \right|}{\max_{j} \left(\left| r_{ij} \right| \right)} \qquad (i, j = 1,2,\ldots,n) \tag{15}$$

This new generation matrix is called relative redundancy matrix.

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = - \begin{bmatrix} \tilde{r}_{11} \Delta_1 + \tilde{r}_{12} \Delta_2 + \ldots + \tilde{r}_{1n} \Delta_n \\ \tilde{r}_{21} \Delta_1 + \tilde{r}_{22} \Delta_2 + \ldots + \tilde{r}_{2n} \Delta_n \\ \vdots \\ \tilde{r}_{n1} \Delta_1 + \tilde{r}_{n2} \Delta_2 + \ldots + \tilde{r}_{nn} \Delta_n \end{bmatrix} \tag{16}$$

The rows of the relative redundancy matrix indicate the relative contributions of all observation errors to an individual residual and the columns indicate the relative contribution of an individual error to all residuals.

It is assumed that observations, the most likely affected by gross errors, are that have the greatest contribution to the most likely abnormal residuals and also have the least contribution to the most likely normal residuals.

Let G be the set of the observation errors have greatest effect on the most likely abnormal residuals and L be the set of the observation errors have least effect on the most likely normal residuals. The intersection set H includes the observation errors so called 'gross errors'. The related observation to the error in set H is treated as outlier. But some procedures are necessary to make the final decision.

In order to obtain the membership values of the observation errors in the set G, the maximum relative contribution of the i^{th} observation error to the residuals that have membership values equal or greater than 0.5 in the set C is looked for. Then, this relative value is multiplied by the membership value of the corresponding residual as follows:

$$\tilde{r}_{ji} = \max_{k = u, v, \ldots, w} \left(\tilde{r}_{ki} \right) \tag{17}$$

$$\mu_G(\Delta_i) = \tilde{r}_{ji} \times \mu_C(v_j) \tag{18}$$

In order to obtain the membership values of the observation errors in set L, the maximum relative contribution of the i^{th} observation error to the residuals that have membership values equal or greater than 0.5 in the set D is looked for. Then, the complementary value of this relative value is multiplied by the membership value of the corresponding residual as follows:

$$\tilde{r}_{mi} = \max_{k=x,y,...,z} \left(\tilde{r}_{ki}\right) \tag{19}$$

$$\mu_L(\Delta_i) = (1.0 - \tilde{r}_{mi}) \times \mu_D(v_m) \tag{20}$$

The membership values of the observation errors in the intersection set H are acquired by means of intersection property of fuzzy sets.

$$\mu_H(\Delta_i) = \min\left(\mu_G(\Delta_i), \mu_L(\Delta_i)\right) \qquad (i = 1, 2, ..., n) \tag{21}$$

Now there is a question to be answered that which observation errors will be considered as gross errors? A critical value must be defined that the membership values will be compared. The critical value ($C\mu_H$) can be calculated taking into account the relative effect values used during the calculation of the membership values of observation errors and corresponding membership values in the set H (Konak and Dilaver 1998).

$$p \Rightarrow \begin{cases} p_i = \tilde{r}_{ji} & \text{if} \quad \mu_H(\Delta_i) \in \mu_G(\Delta_i) \\ p_i = 1 - \tilde{r}_{mi} & \text{if} \quad \mu_H(\Delta_i) \in \mu_L(\Delta_i) \end{cases} \tag{22}$$

$$C\mu_H = \frac{\sum p_i \, \mu_H(\Delta_i)}{\sum p_i} \tag{23}$$

The observation errors, which have membership values greater than the critical value, are assumed as gross errors. But, this assumption must be verified. At this stage, the magnitudes of the errors exceeded the critical value are estimated and the significance of them are tested. The following procedure is proposed by Sun (1994):

$$K_{n\times m} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{24}$$

Here, n is the number of observations, m is the number of observation errors that exceeded the critical value, and K is the location matrix of gross errors. For each gross error, 1 is written at corresponding row. The weight matrix and the magnitudes of gross errors are calculated by means of this location matrix.

$$Pss = K^T PK - K^T PA\left(A^T PA\right)^{-1} A^T PK \tag{25}$$

$$\nabla s = -Pss^{-1} K^T P\, v \tag{26}$$

In equations (25) and (26), $Pss_{m\times m}$ is the weight matrix of gross errors, $P_{n\times n}$ is the weight matrix of observations, $A_{n\times u}$ is the design matrix, $v_{n\times 1}$ is the vector of residuals, and $\nabla s_{m\times 1}$ is the vector of gross errors.

As stated above, the magnitudes of gross errors (∇s) are tested using a statistical test and the decision about their significance is made.

3. APPLICATION AND PROGRAMMING

The conventional methods and fuzzy logic approach have been applied to two GPS networks named as Ordu-1 and Ordu-2 that were established in Ordu, Turkey. The properties of networks are presented below. In these networks, free adjustment has been performed. In order to adjust the measurements and detect outliers, two programs, one for conventional methods and one for the fuzzy logic method, have been written in MATLAB technical computing language. In application of conventional methods, the significance level of statistical tests has been obtained as 0.0003 from equation (4) for each network. Since, such a significance level may cause insensitivity for outliers, it has been considered as 0.001 in statistical tests. Meanwhile, the significance level has also been chosen as 0.01 and it has been seen that when the significance level gets greater, more outlier appears. So, the results at significance level 0.001 are presented in this study.

4. GPS NETWORKS

Ordu-1 GPS network consists of 15 points, 52 baselines, and 84 triangles (Figure 1). The a priori standard deviation has been calculated as 13.146 mm from equation (1). Ordu-2 GPS network consists of 14 points, 43 baselines, and 52 triangles (Figure 2). The a priori standard deviation of this network is 6.164 mm.

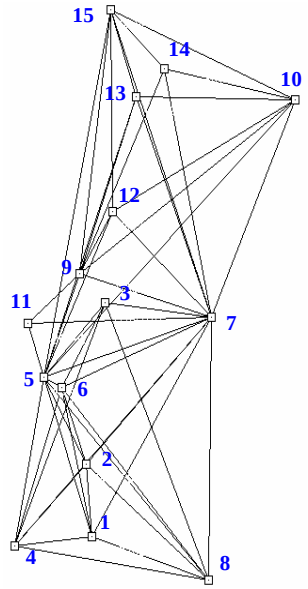

Figure 1 Ordu-1 GPS Network

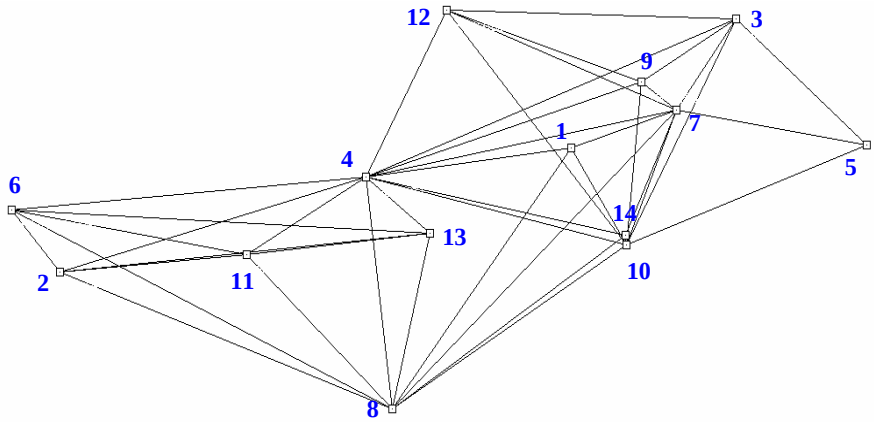

Figure 2: Ordu-2 GPS Network

5. RESULTS OF PROGRAMS

Table 1: Results of conventional methods for Ordu-1 GPS network

Statistical Test	İteration Number	Test Statistic (T_i)	Critical Value (q)	Outlier	Observation Number
Tau	1	5.6482	3.2342	ΔY_{8-5}	68
	2	3.2426	3.2327	ΔY_{8-4}	65
	3	2.8871	3.2311	passed the test	-
DS	1	5.3625	3.3003	ΔY_{8-5}	68
	2	2.6394	3.3003	passed the test	-
t	1	5.9633	3.3787	ΔY_{8-5}	68
	2	3.2753	3.3812	passed the test	-

As seen from Table 1, the results of three statistical tests are identical except for the observation ΔY_{8-4} in Tau test. But, it must be paid attention that the test statistic of this observation is very close to the critical value.

In application of fuzzy logic method, at the beginning (first test) and at the end (last test) of the method, the same statistical test has been used at significance level 0.001. Three different values have been assigned to the standardization component of membership function. If the membership value of an observation error in intersection set H it equals to the critical value, it is treated as gross error. The conditions mentioned here are applied to both GPS networks. The 68th observation seems as outlier after every application of fuzzy logic method in Ordu-1 GPS network (Table 2).

Table 2: Results of fuzzy logic method for Ordu-1 GPS network

First Test	Standardization Component	Outlier	Observation Number	$\mu_H(\Delta_i)$	Critical Value	Last Test	Test Statistic	Critical Value
Tau	0.01	ΔY_{8-5}	68	0.6767	0.5723	Tau	5.6482	3.2342
Tau	0.05	ΔY_{8-5}	68	0.6767	0.5723	Tau	5.6482	3.2342
Tau	0.1	ΔY_{8-5}	68	0.6767	0.5723	Tau	5.6482	3.2342
DS	0.01	ΔY_{8-5}	68	0.6767	0.6767	DS	5.3625	3.3003
DS	0.05	ΔY_{8-5}	68	0.6767	0.6767	DS	5.3625	3.3003
DS	0.1	ΔY_{8-5}	68	0.6767	0.6767	DS	5.3625	3.3003

The results related to Ordu-2 GPS network are as follows:

Table 3: Results of conventional methods for Ordu-2 GPS network

Statistical Test	İteration Number	Test Statistic (T_i)	Critical Value (q)	Outlier	Observation Number
Tau	1	4.3307	3.2192	ΔZ_{8-4}	39
	2	3.4606	3.2168	ΔX_{14-4}	118
	3	3.4126	3.2142	ΔY_{8-2}	35
	4	3.2401	3.2113	ΔZ_{14-7}	123
	5	2.9776	3.2083	passed the test	-
DS	1	3.8802	3.3003	ΔY_{8-4}	39
	2	2.6570	3.3003	passed the test	-
t	1	4.4290	3.4032	ΔZ_{8-4}	39
	2	3.4946	3.4073	ΔX_{14-4}	118
	3	3.4798	3.4116	ΔY_{8-2}	35
	4	3.6038	3.4163	ΔZ_{14-7}	123
	5	3.1201	3.4214	passed the test	-

In Table 3, the observations close to the critical value are identified as outliers. Effects of other observation errors may cause such a result. Some of them are most likely normal observations.

Table 4: Results of fuzzy logic method for Ordu-2 GPS network

First Test	Standardization Component	Outlier	Observation Number	$\mu_H(\Delta_i)$	Critical Value	Last Test	Test Statistic	Critical Value
Tau	0.01	ΔZ_{8-4}	39	0.6204	0.5316	Tau	4.3307	3.2192
Tau	0.05	ΔZ_{8-4}	39	0.6204	0.5316	Tau	4.3307	3.2192
Tau	0.1	ΔZ_{8-4}	39	0.6204	0.5316	Tau	4.3307	3.2192
DS	0.01	ΔZ_{8-4}	39	0.5951	0.5951	DS	3.8802	3.3003
DS	0.05	ΔZ_{8-4}	39	0.5951	0.5951	DS	3.8802	3.3003
DS	0.1	ΔZ_{8-4}	39	0.5951	0.5951	DS	3.8802	3.3003

In fuzzy logic method, the 39^{th} observation ΔZ_{8-4} seems as outlier. This observation also has been conspicuous at the first iteration in conventional methods.

6. CONCLUSIONS AND RECOMMENDATIONS

In GPS networks with redundant observations, choosing the significance level as 0.001 is sufficient to realize outlier detection procedure. Working with great significance levels produces unreliable results. In conventional methods, a normal observation may seem as outlier at the end of iterations and may be removed from observation set. Thus, the shape of the network is defected. On the other hand, more reliable results are obtained with fuzzy logic method. In contrast to the conventional methods, the observations that have test statistics close to the critical value are examined more accurately. It has not any iteration and removal about observations. There is no disadvantage of shape defection of network. The outliers are determined altogether at the end of the method.

ACKNOWLEDGEMENT

We would like to thank Karadeniz Technical University Research Fund for their support to our study.

REFERENCES

Berberan, A., 1995, Multiple Outlier Detection. A Real Case Study, Survey Review, 33, 255 41-49.

Biacs, Z. F., Krakiwsky, E. J., and Lapucha, D., 1990, Reliability Analysis of Phase Observations in GPS Baseline Estimation, Journal of Surveying Engineering, 116, 4, 204-224.

Gökalp, E. and Boz, Y., 2004, Uyuşumuz Ölçülerin Belirlenmesinde Bulanık Mantık ve Geleneksel Yaklaşımların İrdelenmesi, TUJK 2004 Çalıştayı Mühendislik Ölçmelerinde Jeodezik Ağlar Çalıştayı, 14-16 October, Zonguldak.

Heck, B., 1981, Der Einfluss einzelner Beobachtungen auf das Ergebnis einer Ausgleichung und die Suche nach Ausreissern in den beobachtungen, AVN, 88, 17-34.

Konak, H. and Dilaver, A., 1998, Jeodezik Ağlarda Uyuşumsuz Ölçülerin Yerelleştirilmesinde Kullanılan Yöntemlerin Davranışları-II Fuzzy Logic (Bulanık Mantık) Yaklaşımı, Harita ve Kadastro Mühendisliği, 85, 91-109.

Schwarz, C. R. and Kok, 1993, J. J., Blunder Detection and Data Snooping in LS and Robust Adjustments, Journal of Surveying Engineering, 119, 4, 127-136.
Sun, W., 1994, A New Method for Localisation of Gross Errors, Survey Review, 32, 252, 344-358.

BIOGRAPHICAL NOTES

Ertan Gökalp is an associate professor at Karadeniz Technical University (KTU), Turkey. He graduated from the Department of Geodesy and Photogrammetry Engineering at KTU in 1986. He got his M.Eng. degree from the Department of Surveying Engineering at University of New Brunswick (UNB), Fredericton, Canada in 1991. He got his Ph.D. degree from the Department of Geodesy and Photogrammetry Engineering at KTU in 1995. He is currently working at the Department of Geodesy and Photogrammetry Engineering at KTU. His interest areas are GPS (Global Positioning System), Engineering Surveying, and Satellite Geodesy. He is a member of Chamber of Surveying Engineers.

Yüksel Boz is a M.Sc. student at Karadeniz Technical University (KTU), Turkey. He graduated from the Department of Geodesy and Photogrammetry Engineering at Karadeniz Technical University (KTU) in 2002. His interest areas are Satellite Geodesy and GPS. He is a member of Chamber of Surveying Engineers.

CONTACTS

Ertan Gökalp
Karadeniz Technical University
Department of Geodesy and Photogrammetry Engineering
61080 Trabzon
TURKEY.
Tel. + 90 462 3772770
Fax + 90 462 3280918
Email: ertan@ktu.edu.tr

Yüksel Boz
Karadeniz Technical University
Department of Geodesy and Photogrammetry Engineering
61080 Trabzon
TURKEY.
Tel. + 90 462 3772760
Fax + 90 462 3280918
Email: yboz@ktu.edu.tr