Identifying inliers

Michael Greenacre
Department of Economics and Business
Universitat Pompeu Fabra
& Barcelona Graduate School of Economics
Barcelona, Catalonia, Spain

H. Öztaş Ayhan
Department of Statistics
Middle East Technical University
Ankara, Turkey

Abstract

The problem of outliers is well-known in statistics: an outlier is a value that is far from the general distribution of the other observed values, and can often perturb the results of a statistical analysis. Various procedures exist for identifying outliers, in case they need to receive special treatment, which in some cases can be exclusion from consideration. An inlier, by contrast, is an observation lying within the general distribution of other observed values, generally does not perturb the results but is nevertheless non-conforming and unusual. For single variables, an inlier is practically impossible to identify, but in the multivariate case, thanks to interrelationships between variables, values can be identified that are observed to be more central in a distribution but would be expected, based on the other information in the data matrix, to be more outlying. We propose an approach to identify inliers in a data matrix, based on the singular value decomposition. An application is presented using a table of economic indicators for the 27 member countries of the European Union in 2011, where inlying values are identified for some countries such as Estonia and Luxembourg.

Keywords: imputation, inlier, outlier, singular value decomposition

1. Introduction

An outlier on a single variable is an observed value that is unusually far from the general distribution of the other values. A multivariate outlier is an observation vector on several variables that is far away from the other multivariate observations, in terms of the measure of distance in the multivariate space. Various approaches to the identification of univariate and multivariate outliers exist in the statistical literature: see, for example, Rousseeuw and van Someren (1990), Peña and Prieto (2001), Filzmoser (2005).

By contrast, an inlier is an observation that is unusually within the distribution of the other values, when it is expected to be more extreme. Clearly, for a single variable in isolation, the identification of an inlier is practically impossible. But for multivariate data, thanks to the relationships between variables, inliers can be identified. For example, while it is common that the value of a single variable be close to its mean, it is unusual that the values of all variables for a single observation vector be close to their respective means. Such inliers may indicate fabricated data, since the perpetrator of fraudulent data might create data equal or approximately equal to the means of subgroups in the data, since they would tend to go unnoticed, and not perturb the final analysis while increasing the sample size. Such cases have been detected in clinical trials and other biomedical experiments (Evans, 1996; Buys et al, 1999). The approach to identifying such inliers has been to compute Mahalanobis distances, either regular or robust versions, between the observation vectors and various subgroup means, to see if they are unusually close. A similar idea is found in assessing model fit: while attention is usually placed on deciding whether a model does not fit by looking at the right tail of the chi-square distribution, for example, it is equally unlikely that the fit statistic lies in the left tail, very close to the null hypothesis of perfect fit. Having a p-value for a chi-square goodness-of-fit test as high as 0.99 could indicate that the data have been fabricated, since the fit is "too good to be true" – see, for example, the controversial criticism by R.A. Fisher of the results of Mendel's breeding experiments: Fisher (1936), also Edwards (1986).

The term "multivariate inlier" is also used in prediction in a multivariate context. With more and more variables the multidimensional space of the data expands and there can be huge "holes" where no data exist, making prediction for an "inlier" in this part of the space difficult, for example by methods such as nearest neighbours – see the chemometric literature, where spectrometry, chromatography and image analysis are used to infer chemical concentrations, a methodology called multivariate calibration (Martens and Næs, 1989).

As a last sense of the term, "inliers" sometimes indicate all observations that are not outliers, when there is a large proportion of outliers in a data set – see, for example, Zhang and Košecká (2003).

We offer the following definition of an inlier, slightly adapted from the definition in a UN publication (UNECE, 2000):

> An ***inlier*** is a data observation that lies in the interior of a data set and is unusual or in error. Because inliers are difficult to distinguish from the other data values, they are sometimes difficult to find and – if they are in error – to correct.

(The principal difference in our definition is the phrase "is unusual or in error", whereas the UNECE definition referred to above simply says "in error" – we thus believe that inliers are not necessarily in error).

In this paper we consider the specific problem of inlying values (not whole observation vectors of values) in a table of data – these are values on particular variables that are unusually interior to the distributions of those respective variables. In order to identify such values we make a prediction of what they are expected to be given all the other values in the data table. If a data value is predicted to be outlying whereas its observed value is close to the mean, then this is an inlying value, and is unusual for that fact. In other words, we can say that by predicting outliers we are able to identify observed inliers. In most cases these are not errors in the data, although this approach could alert to possible erroneous values.

2. Imputation using the singular value decomposition

For a given data matrix $\mathbf{Y}$ of data, with the n rows representing sampling units and the m columns the variables, we want to capitalize on the relationships between the variables in order to predict what a particular value y_{ij} is expected to be, given all the other data in the matrix. One possibility would be to perform a regression analysis of the j-th variable on the other variables and use the value predicted by the regression. However, this approach treats the whole j-th column as a response vector, whereas it is just the (i,j)-th value that is to be predicted, with the remaining values in the j-th column (those apart from the i-th one) to be also considered as predictors. Fortunately, an imputation algorithm using the singular value decomposition (SVD) is just what we need here, since it uses all the data in the matrix, apart from the (i,j)-th value of interest, to predict that value.

The SVD[*] is a matrix decomposition theorem which expresses a rectangular matrix **B** as the product of three matrices of simple structure, an orthornormal matrix **U**, a diagonal matrix $\mathbf{D}_\alpha$ of positive numbers in descending order, and the transpose of an orthonormal matrix **V:**

$$\mathbf{B} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}, \text{ where } \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I} \quad \alpha_1 \geq \alpha_2 \geq \cdots > 0 \tag{1}$$

An alternative expression of the SVD is as a weighted combination of rank 1 matrices formed by the p columns of **U** and **V**, with the weights being the singular values $\alpha_1, \alpha_2, \cdots, \alpha_p$ (p is the rank of **S**):

$$\mathbf{B} = \alpha_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \alpha_2\mathbf{u}_2\mathbf{v}_2^\mathsf{T} + \cdots + \alpha_p\mathbf{u}_p\mathbf{v}_p^\mathsf{T} \tag{2}$$

The usefulness of the SVD is that if one retains the first k terms, say, of (2) then the resultant matrix is a rank k least-squares approximation of **S**, for example for $k = 2$:

$$\mathbf{B} \approx \hat{\mathbf{B}} = \alpha_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \alpha_2\mathbf{u}_2\mathbf{v}_2^\mathsf{T} \tag{3}$$

Deciding on the number of dimensions k to use in the approximation can be based on various criteria such as the scree plot and a permutation test adapted to the particular data set, to be described in the applications of Section 3.

In order to assess whether any data element of **S** is unusual, either outlying or inlying, each element can be regarded as missing in turn, and its value can be imputed using a missing value algorithm based on the approximation (3) reminiscent of an expectation-minimization (EM) procedure:

1. Consider the (i,j)-th element s_{ij} as missing.
2. Insert a reasonable first approximation of s_{ij} into **S**, call this matrix with this one element changed **S***.
3. Perform the SVD of **S***.
4. Estimate s_{ij} from the approximation as in (3), where k terms of the SVD are included.
5. Insert the new value of s_{ij} into **S*** and repeat from step 4 until convergence is achieved.

The above steps are repeated for each element of the matrix **S**, so that finally we have an estimation of the whole matrix, each element having been estimated from all the other elements. Then these estimates are compared with the originally observed values, and the elements for

[*] Greenacre and Stephens (2011) give a musical explanation of the SVD, one of the most useful results in matrix algebra, and one of the fundamental tools of machine learning – see, for example, Hastie, Tibshirani and Friedman (2009).

which these are the most different are classified as possible inlying or outlying values. An outlier will be a value for which the observed value is far from the mean, for example, whereas the estimated value lies much closer to the mean. An inlier will be a value for which the observed value is quite central in the distribution of that variable, while the estimated value by our procedure is far from the mean.

3. Application

Table 1 consists of six economic indicators for the 27 European Union countries for 2011, gleaned from the Eurostat web site (Eurostat, 2011).

Table 1

Six economic indicators for the 27 European Union countries in 2011[1]. CPI=consumer price index (index, =100 in 2005), UNE=unemployment rate in 15–64 age group (percentage), INP=industrial production (index, =100 in 2005), BOP=balance of payments (€ /capita[2]), PRC=private final consumption expenditure (€ /capita[2]), UN%=annual change in unemployment rate (percentage points).

		CPI	UNE	INP	BOP	PRC	UN%
Belgium	BE	116.03	4.77	125.59	908.6	6716.5	-1.6
Bulgaria	BG	141.20	7.31	102.39	27.8	1094.7	3.5
CzechRep.	CZ	116.20	4.88	119.01	-277.9	2616.4	-0.6
Denmark	DK	114.20	6.03	88.20	1156.4	7992.4	0.5
Germany	DE	111.60	4.63	111.30	499.4	6774.6	-1.3
Estonia	EE	135.08	9.71	111.50	153.4	2194.1	-7.7
Ireland	IE	106.80	10.20	111.20	-166.5	6525.1	2.0
Greece	EL	122.83	11.30	78.22	-764.1	5620.1	6.4
Spain	ES	116.97	15.79	83.44	-280.8	4955.8	0.7
France	FR	111.55	6.77	92.60	-337.1	6828.5	-0.9
Italy	IT	115.00	5.05	87.80	-366.2	5996.6	-0.5
Cyprus	CY	116.44	5.14	86.91	-1090.6	5310.3	-0.4
Latvia	LV	144.47	12.11	110.39	42.3	1968.3	-3.6
Lithuania	LT	135.08	11.47	114.50	-77.4	2130.6	-4.3
Luxembourg	LU	118.19	3.14	85.51	2016.5	10051.6	-3.0
Hungary	HU	134.66	6.77	115.10	156.2	1954.8	-0.1
Malta	MT	117.65	4.15	101.65	359.4	3378.3	-0.6
Netherlands	NL	111.17	3.23	103.80	1156.6	6046.0	-0.4
Austria	AT	114.10	2.99	116.80	87.8	7045.5	-1.5
Poland	PL	119.90	6.28	146.70	-74.8	2124.2	-1.0
Portugal	PT	113.06	9.68	89.30	-613.4	4073.6	0.8
Romania	RO	142.34	4.76	131.80	-128.7	1302.2	3.2
Slovenia	SI	118.33	5.56	105.40	39.4	3528.3	1.8
Slovakia	SK	117.17	9.19	156.30	16.0	2515.3	-2.1
Finland	FI	114.60	5.92	101.00	-503.7	7198.8	-1.3
Sweden	SE	112.71	6.10	100.50	1079.1	7476.7	-2.3
U.K.	UK	120.90	6.11	90.36	-24.3	6843.9	-0.8

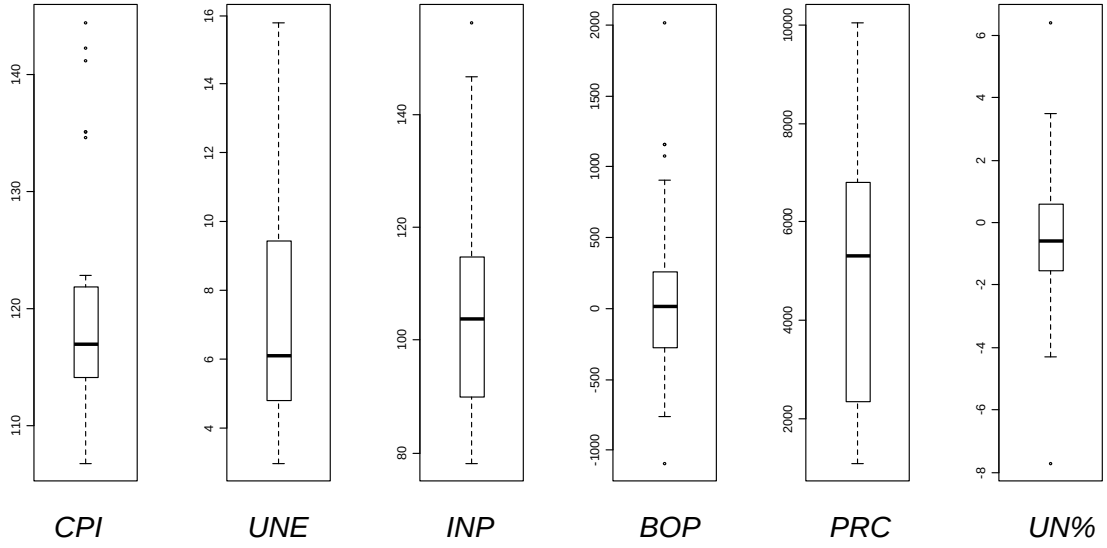
[1] CPI November 2011, UNE June 2011, INP September 2011, BOP second quarter 2011, PRC first quarter 2011, UN% second quarter 2011 compared to second quarter 2010

[2] Per capita computed with respect to population in 15–64 age group

Each country is thus defined by a six-dimensional vector of values. It is easy to see from simple boxplots of the six variables in Figure 1, compared to Table 1, that there are several outliers in inflation rates (CPI) for the recent eastern-European countries entering the EU, that Luxembourg is an outlier in balance of payments (BOP) and that Greece an outlier in increase in unemployment rate (UN%). However, there also several inlying values not obvious from the univariate analyses, as we shall soon see.

Figure 1

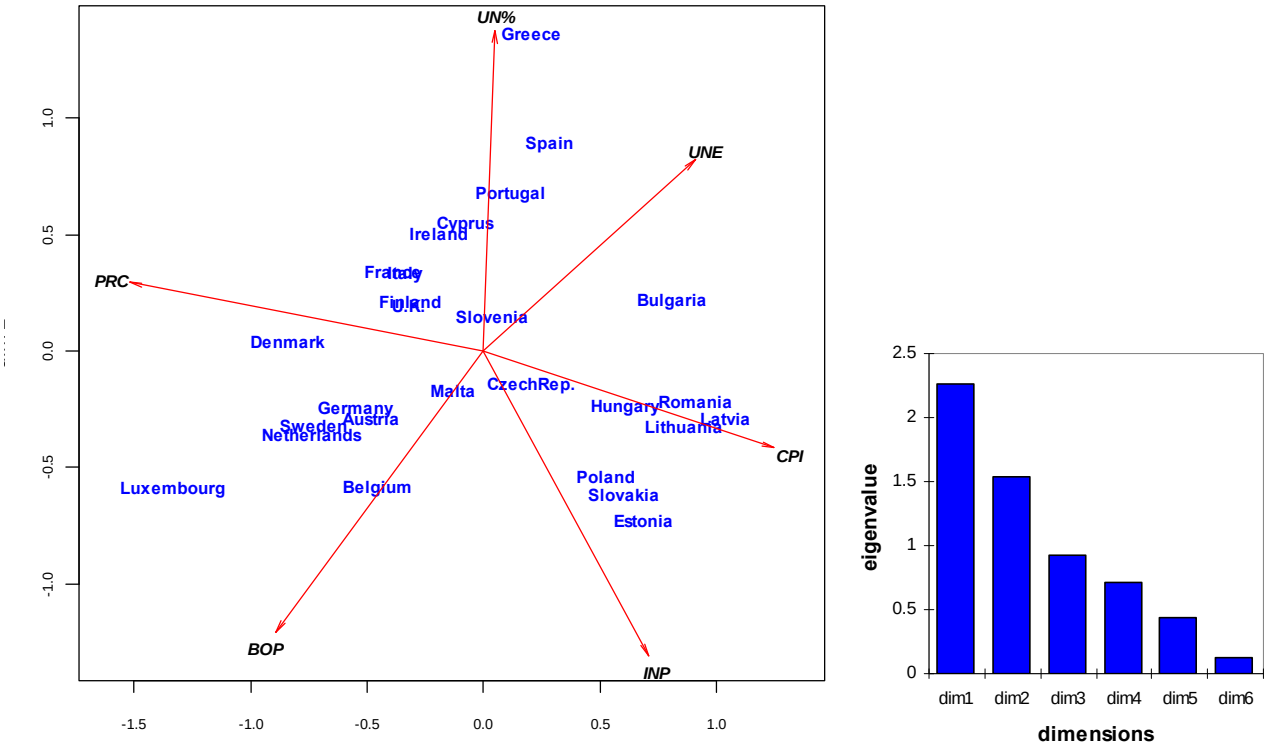
Boxplots of the six variables of Table 1.



Performing a principal component analysis (PCA) biplot (Gabriel, 1971; Greenacre 2010) on the standardized data shows the inherent multivariate structure of the data (Figure 2), for example that unemployment rate is negatively correlated with balance of payments (UNE and BOP point in opposite directions), also that inflation rate (CPI) is negatively correlated with private consumption (PRC). A percentage of 63.3% of the variance of the six variables is explained by this two-dimensional biplot. These two dimensions are the only ones worth interpreting according to several rules of thumb used to decide on the dimensionality of biplots. For example, the first two dimensions are the only two to have variances greater than 1, and the scree plot, shown in Figure 2, also shows these two dimensions well separated from the other four. Furthermore, a permutation test gives the following p-values for the first two dimensions: p = 0.006 and p = 0.037 respectively, while for all the other dimensions the p-value is above 0.9. The permutation test is achieved by randomly permuting the values in each column of the matrix, then performing a PCA and recording the eigenvalues. This is repeated 9999 times to

Figure 2

PCA biplot of standardized data in Table 1. The percentage of inertia explained is 63.3%. The scree plot of the eigenvalues is shown alongside.



generate a permutation distribution of the eigenvalues under the null hypothesis of no relationship between the indicators – see Greenacre (2010) for further details of this permutation test. Our conclusion is to consider the relationships in the data to be inherently two-dimensional, and the remaining variance to be random variation, hence our use of the two-dimensional SVD approximation to impute the data values.

After the imputation exercise described in Section 2 was performed on each of the 27 × 6 = 162 values in the table, the distribution of observed minus imputed values is shown in Figure 3. The distribution is close to normal, with standard deviation sd = 1.04, and there are 9 of the 162 values (i.e., 5.6%) in the table beyond the limits ±2sd, listed in Table 2. Estonia is prominent in this list, having both an outlier and an inlier. From its other indicators one would expect hardly any decrease in unemployment, but the decrease has been 7.70 percentage points, hence an outlier. On the other hand, its industrial production is imputed as much higher than the actual index value of 111.5, hence an inlier. The other inliers are Slovakia, with a near-average CPI but expected to be much higher, and Romania, with a less than average unemployment rate but expected to be much higher (similarly, Romania's change in unemployment rate is imputed as –4.8 percentage points, but it actually increased by 3.2 percentage points). Luxembourg has an

expected CPI of 91.8 based on its other indicators, a deflation that is clearly unrealistic, but this signals the excellent indicators enjoyed by Luxembourg. Finally, an outlier worth noting is Spain, for which one would expect an unemployment rate of only7.4%, but in reality it is 15.8%, the highest in this data set for 2011.

Figure 3

Histogram of observed minus imputed values, for the 27×6 = 162 cells in Table 1.

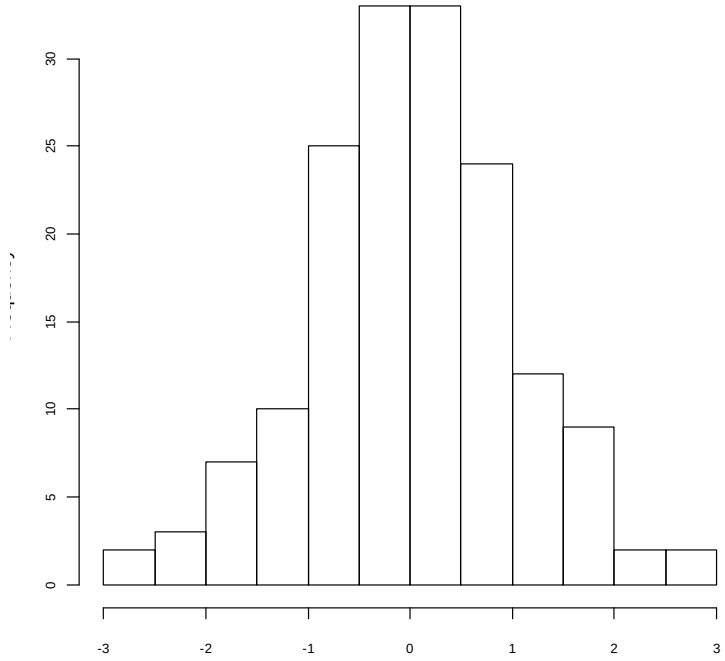


Table 2

Values beyond the limits ±2sd in Figure 3, followed by their original values (standardized form just before in parentheses) and imputed values (standardized form also just before). The observations in boldface red can be considered inliers, because their values are imputed to be far from the variable mean, yet their observed values are more central. Observations in blue are the opposite: their observed values are extreme whereas one would expect their values to be more central.

Value	(stand.) Orig. value		(stand.) Imput. value		Country, variable
–2.79	(–2.63)	–7.70	(–0.16)	–0.13	Estonia, UN%
–2.64	**(0.28)**	**111.50**	**(2.92)**	**161.38**	**Estonia, INP**
–2.33	(–1.09)	85.51	(1.24)	129.63	Luxembourg, INP
–2.26	**(–0.33)**	**117.2**	**(1.92)**	**141.0**	**Slovakia, CPI**
–2.19	**(–0.71)**	**4.76**	**(1.48)**	**11.67**	**Romania, UNE**
2.34	(2.65)	156.3	(0.31)	112.1	Slovakia, INP
2.49	**(–0.24)**	**118.2**	**(–2.73)**	**91.8**	**Luxembourg, CPI**
2.66	(2.79)	15.8	(0.13)	7.4	Spain, UNE
2.97	(1.38)	3.20	(–1.59)	–4.81	Romania, UN%

This approach can be used to detect errors in the data. For example, suppose that Spain's annual increase in unemployment, which is 0.7 percentage points, was inaccurately recorded as -0.7. This change is enough to identify this value as unusual after imputation, because the imputed value is 6.2. The previous observed minus imputed difference of 5.5 was not in the region outside ±2sd but the difference of 6.9 would be.

4. Discussion and conclusion

We have considered the identification of inliers in a multivariate data set. Rather than identifying inlying observation vectors, we have concentrated on identifying inlying values in a data table. This is in parallel to the frequent practice of identifying unusual outlying values in a data table, which might overly affect statistical estimates such as correlations or regression coefficients, and which then necessitate alternative estimation procedures such as nonparametric or robust methods. Inlying values, which are hidden in the data and generally do not affect estimates measurably, are unusual in the opposite sense: one would expect them to be more extreme. We have proposed a simple, yet effective, method for identifying inlying values, based on the matrix approximation property of the singular value decomposition. Inlying data that are correctly measured can not be modified or dispelled – it is their identification and interpretation that is interesting. But the identification of inliers can sometimes signal an incorrect measurement, and thus be useful for improving data quality.

References

Buys, M. et al [13 authors] (1999). The role of biostatistics in the prevention detection and treatment of fraud in clinical trials. *Statistics in Medicine*, 18, 3435–3451.

Edwards, A.W.F. (1986). Are Mendel's results really too close? *Biological Reviews*, 61, 295–312.

Eurostat (2011). Economic indicators.
URL: http://epp.eurostat.ec.europa.eu/portal/page/portal/euroindicators/data, last accessed 28 May 2013.

Evans, S.J.W. (1996). Statistical aspects of the detection of fraud. In S. Lock and F. Welss, eds. *Fraud and Misconduct in Medical Research, 2nd edition*. British Medical Journal Publishing Group, London, pp. 226–239.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. Australian Journal of Statistics, 34, 127–138.

Fisher, R.A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1, 115–137.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 453–467.

Greenacre, M.J. (2010). *Biplots in Practice*. BBVA Foundation, Madrid. Download from URL: http://www.multivariatestatsitics.org.

Greenacre, M.J. and Stephens, G. (2011). It had to be U – the SVD song.
URL: www.youtube.com/watch?v=JEYLfIVvR9I, last accessed 28 May 2014.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd edition*. Springer, New York.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. John Wiley, Chichester, UK.

Peña, D. and Prieto, F.J. (2001). Multivariate outlier detection and robust covariance matrix estimation (with discussion). Technometrics, 43, 286–310.

UNECE (2000). *Glossary of Terms on Statistical Data Editing*. United Nations Statistical Commission and Economic Commission for Europe.
URL: http://www.unece.org/fileadmin/DAM/stats/publications/editingglossary.pdf, last accessed 28 May 2014.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.

Winkler, W. E. (1997). Problems with inliers. Paper presented at the European Conference of Statisticians, Prague. URL: http://www.unece.org/stats/documents/1997/10/data_editing/22.e.pdf, last accessed 28 May 2014.

Zhang, W. and Košecká, J. (2006). A new inlier identification scheme for robust estimation problems. *Proceedings of Robotics: Science and Systems II*, Philadelpha, USA. URL: http://www.roboticsproceedings.org/rss02/p18.html, last accessed 28 May 2014.