Statistical Journal of the IAOS 32 (2016) 257–263
DOI 10.3233/SJI-150949
IOS Press

Iterative method for reducing the impact of outlying *data points:* Ensuring data quality

Svetlana Jesiļevska
Central Statistical Bureau of Latvia, Lāčplēša street 1, Rīga 1301, Latvia
Tel.: +371 27603109; E-mail: mozir@inbox.lv

Abstract. Data editing is essential to check the survey data for possible data problems. Outlying data values are frequently encountered in sample surveys. Consequently, in working with data, the correctness of the reported values must be verified, and if a reported value constitutes an outlier, its appropriate treatment needs to be considered. In this paper, the *Iterative method for the reducing the impact of outlying data points* is proposed. The novelty of the *Iterative method for the reducing the impact of outliers* is the following: an iterative approach for determining the outlying data points is proposed; outliers are determined considering the impact of conjoined factors; estimation of weight coefficients of the outliers and estimation of the total measurement error of the non-linear regression model is carried out.

Keywords: Outlier, data quality, iterative method, conjoined factor, regression model

1. Introduction

Nowadays there is an increasing demand of high quality, reliable and timely statistical data. Confidence in the quality of the statistical data is a matter of survival for a statistical office. Data quality results from the interaction between the attributes of the analytical data (such as its bias, precision, detection and quantitation limits, and other characteristics that together contribute to data uncertainty) and the intended use of the data [6]. Data quality effects in survey data arise from modes of collection [7], interviewers [14] and survey design [35]. The main data quality problems with survey data are non-response problems [17], coverage biases [9], measurement error [2] and outliers. In sample survey data the outliers do not come alone, but are almost always "accompanied" by missing values, large data sets, sampling weights, mixture of continuous and categorical variables etc. [34].

Outliers-data values that differ greatly from the majority of a set of data – are not a new concern [27]. Outlying data values frequently occur in sample surveys, especially in surveys measuring economic and financial phenomena. Outliers exist for the following reasons: incorrect data entry can cause data to contain extreme cases; failure to indicate codes for missing values in a dataset; the case did not come from the intended sample; the distribution of the sample for specific variables may have a more extreme distribution than normal (Statistical Solutions).

Chambers [3] classifies outliers into two groups: "*representative outlier values* (correctly measured sample values that are outlying relative to the rest of the sample data and for which there is no reason to believe that similar values do not exist in the non-sampled part of the survey population; *non-representative outlier values* (gross errors in the sample data, caused by deficiencies in survey processing (e.g. miscoding)".

Eurostat Statistics Explained Glossary makes the difference between outliers and inliers: *"an outlier is a data value that lies in the tail of the statistical distribution of a set of data values. In the distribution of raw data, outliers are often regarded as more likely to be incorrect. In contrast, an* ***inlier*** *is an erroneous data value which actually lies in the interior of a statistical distribution, making it difficult to distinguish it from good data values."* (Statistics Explained Glossary)

Univariate and multivariate outliers are distinguished. *"A univariate outlier is a data point that consists of an extreme value on one variable. A multivariate out-*

lier is a combination of unusual scores on at least two variables." (Statistical Solutions) Both types of outliers can influence the outcome of statistical analyses.

When the survey is completed, the statistician must perform the survey data pre-editing in order to clean data set: in order to detect and appropriately modify outlying values, such that the data set is suitable for further general use with standard statistical software.

Scientists have developed several methods to identify outliers. For example, during the Coference of European Statisticians, Work Session on Statistical Data Editing (held in Paris, France, on the 28–30 April, 2014) the following outlier detection methods were presented: Quantiles, MAD distance, Share in total [18]. The outliers can be detected by conducting hypothesis tests, e.g. Grubb's test [15], Dixon test [8]. Knorr et al. introduced a distance-based method to identify the outliers [20]. Spiros et al. introduce the method to detect outliers by using the multi-granularity deviation factor (MDEF) [30]. Recently, Kriegel et al. proposed the angle-based method that computes outlier scores based on the angles of the points with respect to other points [21]. However, the method can not detect outliers surrounded by other points.

Various approaches to the identification of univariate and multivariate outliers exist in the statistical literature: see, for example, Rousseeuw and van Someren [28], Peña and Prieto [25], Filzmoser [12]. Methods proposed for univariate outlier detection are based on (robust) estimation of location and scatter, or on quantiles of the data. A major disadvantage is that these rules are independent from the sample size. Moreover, by definition of most rules outliers are identified even for "clean" data (Filzmoser (a)). Multivariate outliers can be identified with the use of Mahalanobis distance and can also be recognized using leverage, discrepancy, and influence (Statistical Solutions). *Each* of these methods *has* strengths and weaknesses.

Penny and Jolliffe conducted a comparison study with six multivariate outlier detection methods. In particular, the methods depend on whether or not the data set is multivariate normal; on the dimension of the data set; on the type of the outliers; on the proportion of outliers in the dataset; and on the outliers' degree of contamination (outlyingness) [26].

The multivariate aspect of the data collected in surveys makes the task of outlier identification particularly challenging. The outliers can be completely hidden in one or two dimensional views of the data [34].

In this paper, author presents the *Iterative method for the reducing the impact of outlying data points*. The Iterative method aims to deal with multivariate outliers in the sample survey data. The Iterative method has not weaknesses described above. The novelty of the *Iterative method for the reducing the impact of outliers* is the following: an iterative approach for determining the outlying data points is proposed; outliers are determined considering the impact of conjoined factors; estimation of weight coefficients of the outliers and estimation of the total measurement error of the non-linear regression model is carried out.

2. The iterative method for the reducing the impact of outlying data points

Detection and treatment of values deviating extremely from other values of the data set is an old problem of statistics.

The aim of the outlier treatment is improving the estimation.

The iterative method for the reducing the impact of outlying data points described below aims to ensure statistical data quality during data pre-processing step.

The *Iterative method* involves a series of steps to identify outlying data points and reduce its impacts. During the first step, an indicator and a factor is selected for further analysis. The conjoined factor may be used during the analysis. During the second step, the best-fit regression model for further analysis is determined. The third step is the total estimate error of the chosen regression model measurement. The fourth step is the most extensive as it consists of four substeps. During the fourth step, firstly, potential outlier points are identified. After potential outlier points are evaluated for the reason of their existence and factual outlier points are identified. Then, the author suggests to minimize the impact of factual outliers on the results: weight ratio for the outlier data point is determined by the normal distribution law. As a result, regression's model recalculation with the corrected data is performed. The last step is validation. During the validation step, the received results should be analyzed and evaluated. If potential outliers are detected, come back to the 2^{nd} step and run a new *Iterative method* circle.

We shall take a closer look at these steps.

1st step: Select data for analysis

Often statistical survey questions can be linked to each other for in-depth analysis.

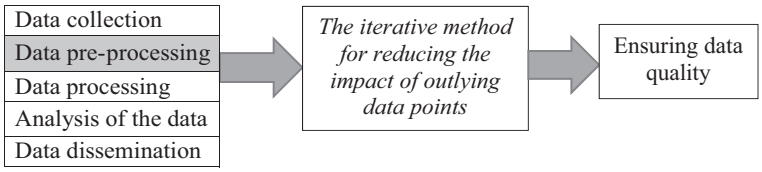

Fig. 1. Statistical data quality assurance Source: Author's scheme.

An example of such a survey is the Community Innovation Survey. In this paper the *Iterative method* will be approbated using the Community Innovation Survey 2012 data. The Community Innovation Survey 2012 was conducted in Latvia during 2013 and covered enterprises with 10 employees or more. The survey collected information about innovation activity in Latvian enterprises during the three-year period 2010–2012.

$$y_i \; - \; \text{survey data (an indicator)}; i = \overline{1,n}$$

In this example y_i is *the percentage of the total turnover of enterprises in 2012 from new or significantly improved products introduced during the three years 2010 to 2012 that were new to the market.*

$$x_j \; - \; \text{survey data (a factor)}; j = \overline{1,k}$$

In this example as a factor the conjoined factor is used and is calculated by the following formula:

$$x_j^{CF} = \frac{\check{x}_j^{TURN} + \check{x}_j^{EMPL}}{2},$$

where

$$\check{x}_j^{TURN} = x_j^{TURN} / \overline{x}_{\max}^{TURN};$$

$$\check{x}_j^{EMPL} = x_j^{EMPL} / \overline{x}_{\max}^{EMPL}$$

- j is a number of respondents, $j = \overline{1 \ldots k}$
- x_j^{TURN} is a turnover of the j^{th} enterprise;
- x_j^{EMPL} is a number of employees of the j^{th} enterprise;
- $\overline{x}_{\max}^{TURN}$ is an average of k_θ maximum turnover values, where k_θ is calculated as *2,33%* from k;
- $\overline{x}_{\max}^{EMPL}$ is an average of k_θ maximum values of number of employees, where k_θ is calculated as *2,33%* from k.

For large enterprises X_j^{CE} value is higher than for small enterprises.

A *z-score* of $2,33$ means that we are talking about the value that is $2,33$ standard deviations from the mean. In our case, this is equivalent to 4 enterprises.

One can notice that the conjoined factor was calculated using two determinants of the size of the enterprise: the turnover and the number of employees. The question of how enterprise size relates to the ability and propensity to innovate is one of the oldest in political economy [16]. A number of studies based on innovation counts have found that small enterprises have introduced more innovations per thousand employees than larger enterprises. Small enterprises are more innovative than large enterprises, or that small enterprises are more efficient innovators than large enterprises [4]. This interpretation, however, depends on the important assumption that, on average, the value of the innovations introduced did not increase systematically with the size of the innovating enterprises [33].

In the context of innovation surveys, the enterprise size (determined by turnover or number of employees) still is one of the main factors that has an impact on innovative performance of the enterprises.

2^{nd} step: Determine the best-fit model for further analysis

$\hat{y} = f(A, x_j^{CE})$ – regression model *(e.g. logistic, exponent, semi-logarithmic etc.)*

The best shape of the function is determined using quality evaluation criteria.

In this practical example, a logarithmic function is used.

3^{rd} step: Measure the total estimate error of the chosen regression model

Šķiltere and Danusēvičs [32] developed the Iterative method for evaluating parameters of truly non-linear trend models. This method is based on least square method and gradual calculation of the model parameters [32]. This approach is used in the *Iterative method for the reducing the impact of outlying data points.*

For an exponent function:

$$\hat{y} = a \cdot x^b \rightarrow lny = lna + blnx;$$

$$\Delta = \pm t_\alpha \sqrt{s_{\hat{y}}^2 + S_{lna}^2 + S_b^2 \exp(lnx - \overline{lnx})^2},$$

where

$$S_{lna} = \sqrt{\frac{S^2_{\ln \tilde{y}}}{n}}; S_b = \sqrt{\frac{S^2_{\ln \tilde{y}}}{S^2_{lnx} \cdot n}};$$

For a Pearl-Reed model:

$$\hat{y} = \frac{c}{1 + ae^{-bx}};$$

$$\Delta = \pm t_\alpha \sqrt{S^s_{\tilde{y}} + \exp\left(x^2_{\ln\left(\frac{c}{\tilde{y}-1}\right)}\right) + \exp\left(\frac{S^2_{\ln\left(\frac{c}{\tilde{y}-1}\right)}}{n} + \frac{S^2_{\ln\left(\frac{c}{\tilde{y}-1}\right)}}{nS^2_x}(x-\overline{x})^2\right)};$$

etc.

For a logarihtmic function:

$$\hat{y} = a + b \cdot lnx$$

$$\Delta = \pm t_\alpha \sqrt{S^2_{\tilde{y}} + S^2_a + S^2_b \exp(lnx - \overline{lnx})^2},$$

where

$S_{\hat{y}}$ – standard deviation of the model,

$$S_{\hat{y}} = \sqrt{\frac{\Sigma^n_{i=1}(y_i - \hat{y}_i)^2}{n - p - 1}};$$

where p is the number of independent variables or predictors.

S_a – standard deviation of the parameter a,

$$S_a = \sqrt{\frac{S^2_{\tilde{y}}}{n}};$$

S_b – standard deviation of the parameter b,

$$S_b = \sqrt{\frac{S^2_{\tilde{y}}}{S^2_{lnx} \cdot n}};$$

In this practical example, the total estimate error for a logarithmic function is calculated.

4th step: Outlier treatment

The aim of the outlier detection during the editing phase is to decide whether this value is a real value or an error. Errors have to be corrected. Real extreme values cannot be the objects of imputation.

1. Potential outlier point identification

$$y^{POP}_i = \begin{cases} y_i > \hat{y}_i + \Delta \\ y_i > \hat{y}_i - \Delta \end{cases},$$

where

- y_i – survey data; $i = \overline{1, n}$
- y^{POP}_i – potential outlier points (POP); $i = \overline{1, n}$
- $\hat{y}_i$ – regression model predicted values; $i = \overline{1, n}$
- Δ – the total estimate error of the chosen regression model

Potential outlier points are selected with statistical methods. Once potential outlier points have been identified they should be further evaluated to determine the reason for their existence. Potential outlier points should generally be kept as part of the data set unless there is reasonable evidence that they are the result of an error, standardization failure etc.

2. Factual outlier point identification and treatment

Qualitative analysis of potential outlier points: $y^{OP}_i \in [y^{POP}_i]; i = \overline{1, n}$.

In this sub-step, it's important to investigate the nature of the potential outlier before deciding what to do with it. There is always possibility to put in some subjective consideration when we analyze the nature of potential outlier points. Potential outlier points are evaluated for the reason of their existence.

Outliers can arise from several different mechanisms or causes. Anscombe [1] sorts outliers into two major categories: those arising from errors in the data, and those arising from the inherent variability of the data. Some more outlier causes, proposed by researchers:

Outliers from data errors. Outliers are often caused by human error: errors in data collection, recording, or entry [24].

Outliers from standardization failure. Outliers can be caused by research methodology, particularly if something anomalous happened during a particular subject's experience [24].

Outliers from faulty distributional assumptions. Incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers [19].

3. Reduce outlier influence by determining weight ratio to outlier data point

There are three main methods of dealing with outliers in a finite population [5]: reducing the weights of outliers (trimming weight); changing the values of outliers (winsorization, trimming); using robust estimation techniques such as M-estimation.

In author's view, an approach of excluding an outlier data point from the dataset reduces the amount of data therefore the author suggests to minimize the impact of outliers on the results: weight ratio for the outlier data point is determined by the normal distribution law: $\beta^{OP}_i = e^{-\frac{(y_i - \hat{y}_i)^2}{2S^2_{\hat{y}}}}$, where $S_{\hat{y}}$ – standard deviation of the regression model.

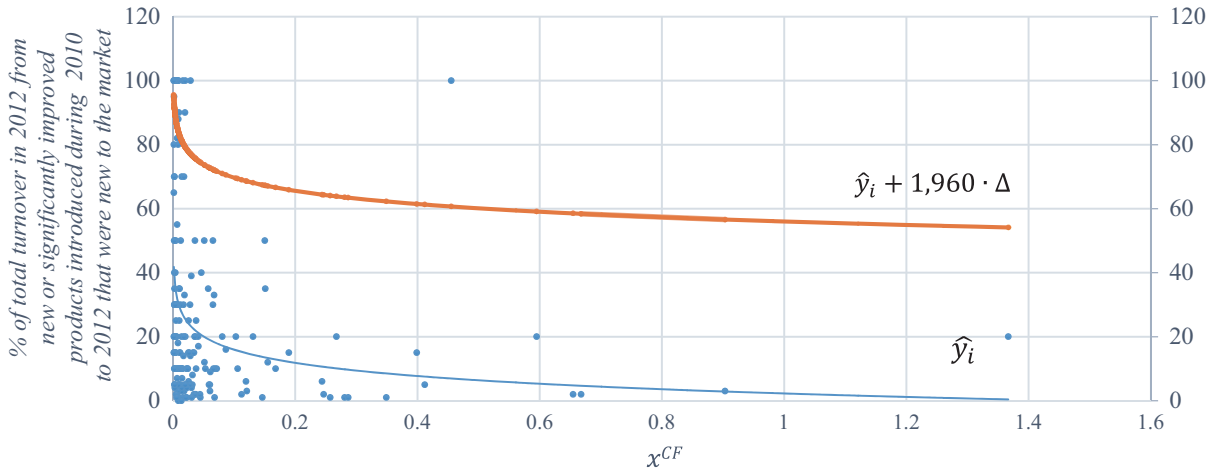

Fig. 2. Iterative method for the reducing the impact of outlying data points calculations. Source: Author's calculations based on the CIS2012 data.

Regression's model $\hat{y}_i^*$ recalculation with the corrected data

$$y_i^* = y_i^{OP} \beta_i^{OP}.$$

For the other data: $y_i = y_i \beta_i, \beta_i = 1$,

5^{th} step: Validation

During the validation step, the received results should be analyzed and evaluated. If potential outliers are detected, come back to the 2nd step and run a new *Iterative method* circle.

In the next sub-section, author provides a practical example.

3. Iterative method for the reducing the impact of outlying *data points* calculations

The percentage of the total turnover in 2012 from new or significantly improved products introduced during the three years 2010 to 2012 that were new to the market by conjoined coefficient presented in Fig. 2 (grey dots). The proportions of sampled enterprises by size according to the number of employees in the Community Innovation Survey carried out in Latvia in 2012 are the following: 63% of the surveyed enterprises are small (10–49 employed persons); 29% are medium (50–249 employed persons) and 8% are large enterprises (250 or more persons employed).

In theory, one can argue that the larger is an enterprise the higher is the percentage of the turnover from innovative products. In practice, in Latvia there is a high number of small enterprises. SMEs are of great importance to Latvia's business economy as they provide 78% of employment and 72% of value added, significantly higher than the EU averages (67% and 58% respectively) [10]. The share of manufacturing in Latvia's economy is one of the lowest among the new EU member states. Also, productivity level in Latvia's manufacturing is considerably below the EU average. Specialization in low technology sectors in Latvia is the key factor that reduces the productivity level in manufacturing [29]. All these peculiarities of the Latvian economic structure explain the data layout on the graph and the shape of the model curve on Fig. 2.

Thin grey line marked as $\hat{y}_t$ in Fig. 2 shows a regression model function that is a logarithmic function that describes the predicted values of the model.

According to the normal curve probability density function, 95% of the data will fall within 1.960 standard deviations of the mean. Depending on the stringency of the researcher's criteria, which should be defined and justified by the researcher, the following values can be discussed: of interval (cautious researcher), interval (moderate researcher) or even interval (liberal researcher). The author have chosen the threshold for an example discussed in this paper.

The data points that are not covered by interval in this practical example are potential outliers. The values were checked very carefully because these data values differ greatly from the majority of a set of data. Human intervention is an important part of this process. The errors with data were identified, and in order to minimize the impact of outliers on the results: weight

ratios for the outlier data points were determined by the normal distribution law: $\beta_i^{OP} = e^{-\frac{(y_i-\hat{y}_i)^2}{2S_{\hat{y}}^2}}$. Recalculated regression impact model $\hat{y}_i^*$ after minimization of the impact of the outliers is the following: $\hat{y}_i^* = -4,717 \in (x^{CF}) + 2,1895$. In this practical example, during the validation step, no more potential outliers were detected.

4. Conclusions

The author have presented the *Iterative method for reducing the impact of outlier data points*. This method provides a new approach for the reducing the impact of outliers and ensuring data quality. The novelty of the *Iterative method* is the following: an iterative approach for determining the outlying data points; outliers are determined considering the impact of conjoined factors; estimation of weight coefficients of the outliers and estimation of the total measurement error of the non-linear regression model is carried out.

References

[1] Anscombe, *Francis John.* Rejection of outliers, *Technometrics* **2** (1960), 123–147.

[2] P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, Measurement Errors in Surveys. Wiley, 1991.

[3] R.L. Chambers, Outlier robust finite population estimation, *Journal of the American Statistical Association* **81** (1986), 1063–1069.

[4] W.M. Cohen and S. Klepper, A Reprise on Firm Size and R&D, *Economic Journal* **106**(437) (1996), 925–951.

[5] B.G. Cox et al., Business Survey Methods. John Wiley & Sons, 1995.

[6] D.N. Crumbling, In search of representativeness: evolving the environmental data quality model, *Quality Assurance* **9** (2001), 179–190.

[7] E.D. de Leeuw and J. van der Zouwen, Data quality in telephone and face to face surveys: A comparative meta-analysis, *Telephone Survey Methodology* (1988), 283–300.

[8] R.B. Dean and W.J. Dixon, Simplified Statistics for Small Numbers of Observations, *Anal. Chem* **23**(4) (1951), 636–638.

[9] K.B. Duncan and E.A. Stasny, Using propensity scores to control coverage bias in telephone surveys, *Survey Methodology* **27**(2) (2001), 121–130.

[10] European Commission, 2014, SBA Fact Sheet–Latvia.

[11] Eurostat. Statistics Explained Glossary. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Outlier [online, accessed 8 November 2015]

[12] P. Filzmoser, Identification of multivariate outliers: A performance study, *Australian Journal of Statistics* **34** (2005), 127–138.

[13] P. Filzmoser (a), Univariate and Multivariate Outlier Detection with Application to Geochemical Data http://www.statistik.tuwien.ac.at/rmed03/abstracts/filzmoser.pdf [online, accessed 7 November 2015].

[14] R.M. Groves, L.J. Magilavy and N.A. Mathiowetz, The process of interviewer variability: Evidence from telephone surveys. In ASA Proceedings of the Section on Survey Research Methods, Alexandria, VA. American Statistical Association, 1981, pp. 438–443.

[15] F.E. Grubbs, Procedures for Detecting Outlying Observations in Samples, *Technometrics* **11** (1969), 1–21.

[16] B. Harrison, Lean and Mean. Basic Books, New York, 1994.

[17] M.A. Hidiroglou, J.D. Drew and G.B. Gray, A framework for measuring and reducing nonresponse in surveys, *Survey Methodology* **19** (1993), 81–94.

[18] Horváth, Gergely, Presentation and development of outlier treatment in HCSO. United Nations Economic Commission for Europe Conference of European Statisticians. Work Session on Statistical Data Editing, 2014, pp. 1–10.

[19] B. Iglewicz and D.C. Hoaglin, How to detect and handle outliers. Milwaukee, WI.: ASQC Quality Press, 1993.

[20] E.M. Knorr and T.N. Raymond, Algorithms for mining distance-based outliers in large datasets, in VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases, (San Francisco, CA, USA), Morgan Kaufmann Publishers Inc, 1998, pp. 392–403.

[21] H.-P. Kriegel, M.S. Hubert and A. Zimek, Angle-based outlier detection in high-dimensional data, in KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, (New York, NY, USA), 2008, pp. 444–452, ACM.

[22] H. Miller, The multiple dimensions of information quality, *Information Systems Management* **13**(2) (1996), 79–83.

[23] OECD, OECD glossary of statistical terms. Available on-line at http://stats.oecd.org/glossary/detail.asp?ID=5054.

[24] J.W. Osborne and A. Overbay, The power of outliers (and why researchers should always check for them), *Practical Assessment, Research & Evaluation* **9**(6) (2004), 1–12.

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

[26] K.I. Penny and I.T. Jolliffe, A comparison of multivariate outlier detection methods for clinical laboratory safety data, *The Statistician* **50**(3) (2001), 295–308.

[27] P.J. Rousseeuw and C. Croux, Alternatives to the median absolute deviation, *Journal of the American Statistical Association* **88**(424) (1993), 1273–1283.

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

[29] I. Skribane and S. Jekabsone, Structural weaknesses and challenges of the economic growth of Latvia, *Social Research* **1**(34) (2014), 74–85.

[30] Spiros Papadimitriou, Hiroyuki Kitagawa, Philip B. Gibbons, and Christos Faloutsos, LOCI: Fast outlier detection using the local correlation integral," in Proceedings of the 19th International Conference on Data Engineering: 2003, pp. 315–326, IEEE Computer Society Press.

[31] Statistical Solutions. Univariate and Multivariate Outliers. http://www.statisticssolutions.com/univariate-and-multivariate-outliers/ [online, accessed 8 November 2015].

[32] D. Šķiltere and M. Danusēvičs, Interval Forecasting Methods In Longterm Statistical Forecasting, *A Journal of the Inter-*

national Institute for General Systems Studies **11**(1) (2010), 11–20.

[33] B. Tether, Small and large firms: Sources of unequal innovations? CRIC Discussion Paper No 11, 1998, pp. 1–40.

[34] V. Todorov et al., Detection of Multivariate Outliers in Business Survey Data with Incomplete Information http://www.statistik.tuwien.ac.at/public/filz/papers/ADAC10.pdf [online, accessed 9 November 2015].

[35] C. Tucker, The estimation of instrument effects on data quality in the Consumer Expenditure Diary Survey, *Journal of Official Statistics* **8** (1992), pp. 41–61.