

Detection of gross anomalies in datasets used for official statistics production

Francesca Torti and Domenico Perrotta

Abstract The contribution proposes an approach to identify in datasets collected by statistical offices gross anomalies such as collecting and recording errors or manifest false declarations. The approach is based on the Forward Search, a robust outlier detection technique, and its use for automatic data transformations. The work is motivated by concrete problems in international trade and agricultural census data.

Key words: Forward Search, multivariate outliers, Box Cox transformations, data quality, census data, trade data.

1 Introduction

The statistical institutions (e.g. the national statistical offices, Eurostat and international organisations such as United Nations Statistics Division, WTO, OECD, etc.) have to ensure the accuracy of the published official data. Many factors can affect data quality. In collecting numerical data a wrong unit of measurement may be used. Then, during the usual stages of coding, data entry or data editing, human operators may introduce recording errors such as the inversion of contiguous digits, the substitution of one digit with another, the inclusion or omission of one or more digits, the omission of separator of decimal values (commas or dots). These kinds of errors may obviously affect differently the final statistics, depending on the order of magnitude of the quantity involved and the position of the mistake in the number. Statistical analyses must not be exposed to these gross errors and their identification is therefore of undoubted importance.

Francesca Torti
Dipartimento di Economia, Università di Parma, Italy, e-mail: francesca.torti@nemo.unipr.it

Domenico Perrotta
European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizens, Ispra, Italy, e-mail: domenico.perrotta@ec.europa.eu

Recording and collecting errors can be treated as any anomalous observation, i.e. as outliers [5]. Of course, not all anomalous observations are errors. Therefore procedures for cleaning data from collecting and recording errors should be used with caution to not discharge relevant data. For example, in international trade certain price outliers are informative of specific market dynamics or may lead to cases of unfair competition or fraud. Similarly, in agricultural census certain anomalous declarations of breeding could refer to cases that we need to highlight in subsequent phases of the analysis. In this paper we will focus on multivariate survey data of the Italian census on agriculture and bivariate trade data collected by the national customs and published after appropriate aggregations by Eurostat. We describe our approach to detect gross outliers in multivariate data in Section 2. Then, in Sections 3 and 4 we demonstrate the approach on a multivariate dataset, the agricultural census data from the Italian Statistical Office (ISTAT), and a bivariate international trade dataset taken from the COMEXT database of Eurostat.

2 Methodologies

The identification of outliers is the main subject of robust statistics, which main goal is to build estimators independent from model assumption deviations. The robust statistic discipline has grown considerably in the last two decades and several methodologies have been proposed in contexts including regression, multivariate statistic and time series analysis. For the robust analysis of multivariate data one of the most popular methods is the Minimum Mahalanobis Distance by Rousseeuw and Van Driessen [8]. It consists in building a subset of the data with the smallest determinant, computing the desired estimators with reference to this subset and declaring outliers on the basis of these estimators. The Minimum Mahalanobis Distance estimators, after proper calibrations built to control the test size, have quite good power, as shown in [6]. However the Minimum Mahalanobis Distance, as most high breakdown estimators, assume that the data are normal. Unfortunately, in the contexts addressed in Sections 3 and 4, deviations from normality happen with a certain frequency and these cases would produce an increase in the number of false outliers. An approach to treat non-normal data consists in exploring proper transformations to normality and apply robust estimators on the transformed data. A family of parametric transformations that can be used for general data was proposed by [9]. This family generalises the well known Box and Cox exponential family [3] that is applicable to strictly positive data like those of this paper. In multivariate data the transformation can be applied to one variable at the time (univariate transformation) or to all variables at the same time. In the following we will consider for simplicity the univariate transformation case. The Box and Cox family is:

$$z(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda y^{\lambda-1}} & \lambda \neq 0 \\ y \log(y) & \lambda = 0 \end{cases} \quad (1)$$

where $\hat{y} = \exp(\sum(\log(y_i)/n))$ represents the geometric mean of the data. If $\lambda = 1$ no transformation is needed, if $\lambda = \frac{1}{2}$ or $\lambda = 0$ or $\lambda = -1$, the square root, the logarithmic and the reciprocal transformations are respectively required. λ is found as the solution of the multiple regression model $z(\lambda) = X\beta + \varepsilon$. The statistic used in the Box-Cox test for transformation is the so called *score test*, which indicates the goodness of a specific transformation value λ .

The big problem behind the use of the score test is that it is computed using all the data and therefore it is affected by the presence of outliers. To avoid this problem, one could remove backward one observation at the time, compute the score test for the $n - 1$ remaining observations and estimate the influence of the removed observation on the score test. Outliers would have big influence and a significant score test. Unfortunately this traditional backward approach is subject to the well known masking effect, that takes place when multiple outliers are present and they are placed so that one masks the presence of another and none appear significant in terms of score test (or in terms of other statistics in other contexts). A natural remedy to the issue is to remove k observations at the time instead of 1, and estimate the score test on the remaining $n - k$ observations. But also this solution is of no practical use for increasing values of k , because the problem is combinatorial: the combinations to test would be $\binom{n}{k}$, being n the dataset size.

We propose an approach that can at the same time estimate robustly the score test, select the best transformation parameter, transform the data and detect the outliers on the transformed data. The approach is based on the *Forward Search* of Atkinson, Riani and Cerioli [1, 2]. The main idea of the Forward Search is to start from a small homogeneous subset of the data, chosen for instance as intersection of units selected with bivariate boxplots, and iteratively include observations based on the Mahalanobis distance measure. The step from a m -size sub-sample to the following consists in taking the $m + 1$ observations with smaller Mahalanobis distance values. The good performances of the Forward Search in terms of both test size and power have been proved in the multivariate context by [6].

While the search progresses from the initial subset $S(m_0)$ to $S(n)$, one can monitor many statistics relative to the subset or to the observations not yet included. A robust tool to understand the percentage of observations in accordance with the different values of the transformation parameters λ is the forward plot of the score test statistic for transformation for different values λ , using a separate search for each λ . The trajectories of the score tests together with their confidence bands, can be combined in a single picture named the *fan plot* [1]. One can choose the most appropriate transformation value by looking at the overall progression of the the different trajectories and by selecting the trajectory which, during the central-last part of the search, is more likely to lie inside the envelopes; then one can identify a subset of homogeneous data in the transformed data and detect the outliers. Typically for large and complex datasets many transformation parameters should be studied. Here we limit the analysis to the parameter values $\lambda = -1, -0.5, -0.25, 0, 0.25, 0.5, 1, 1.5$ and, if no trajectory is inside the bands, we check if there are two λ with trajectories progressing just above and below the bands and reasonably close to them, and we choose their average.

3 First case study: agricultural census data

The data collected for the agricultural census can be grouped according to the kind of farming or breeding that each firm declares to own and they represent the quantity of farming or breeding owned. Any agricultural firm can declare one or more farming or breeding types. All these are response variables of a multivariate problem. The data can be hierarchically clustered according to geographical areas, typically regions, provinces and communes. We can assume that data from the same geographical area are homogeneous. Possible outliers are likely to originate from mendacious or erroneous declarations and recording errors. Severe errors should be removed to ensure the quality of the census analysis. In this Section we discuss three representative examples of what we found in agricultural census data and we show how to treat them with the Forward Search methodology proposed before. To simplify the exposition, we will analyse data with two variables, say X_1 and X_2 .

The scatterplot matrix of the first dataset is in the top-left panel of Figure 1. The shape of the points is quite far from the typical ellipsoidal shape of a Normal distribution. The Forward Search applied to this dataset detects three clear outliers represented with red circles. Other robust methods are expected to produce the same result. However, if we apply to each variable the transformation procedure described in Section 2, we obtain the two fan plots on the bottom panels of the same Figure. From the fan plot of variable X_1 (on the left) we can see that the score test for transformation parameter $\lambda = 0.25$ lies inside its 99% envelopes for almost all the steps of the Forward Search, while in the fan plot of variable X_2 (on the right) it is $\lambda = -0.25$ which lies inside the envelopes. These are the optimal transformations parameters suggested by the Forward Search. The top-right panel of the same figure gives the scatterplot matrix of the data transformed with these parameter values. The elliptical shape, characteristic of normal data, is now evident and the transformed data look homogeneous. In fact, now the Forward Search detects no outlier.

A second representative case is shown on the top-left panel of Figure 2. Also in this case the original data are not elliptical and three outliers are detected by the Forward Search (red circles). According to the fan plots on the bottom panels, the most appropriate Box-Cox transformation parameters in this case are 0.5 for variable X_1 and -0.25 for variable X_2 . Once transformed (top-right panel), the data have a more elliptical shape and the Forward Search now detects only one outlier. We have highlighted this outlier as a filled red circle also in the scatterplot of the data in the original scale (top-left panel). At first sight the result seems counter-intuitive: the filled red circle coincides with none of the three outliers detected in the original scale data (red circles) and, besides, it seems situated well in the cloud of the data. In reality, it corresponds to the observation with the lowest value (3) of variable X_2 .

The third representative dataset is reported on the left panel of Figure 3. These data, characterised by a very high dense area in correspondence of $X_1 = 10000$ and $X_2 = 0$, are again far from the elliptical normal shape. However, contrary to the previous cases, the fan plot for variable X_1 (right panel) suggests that none of the tested transformation values are suitable for such data. All score test statistic trajectories deviate from the bands since the first steps of the search. At this point the fan plot

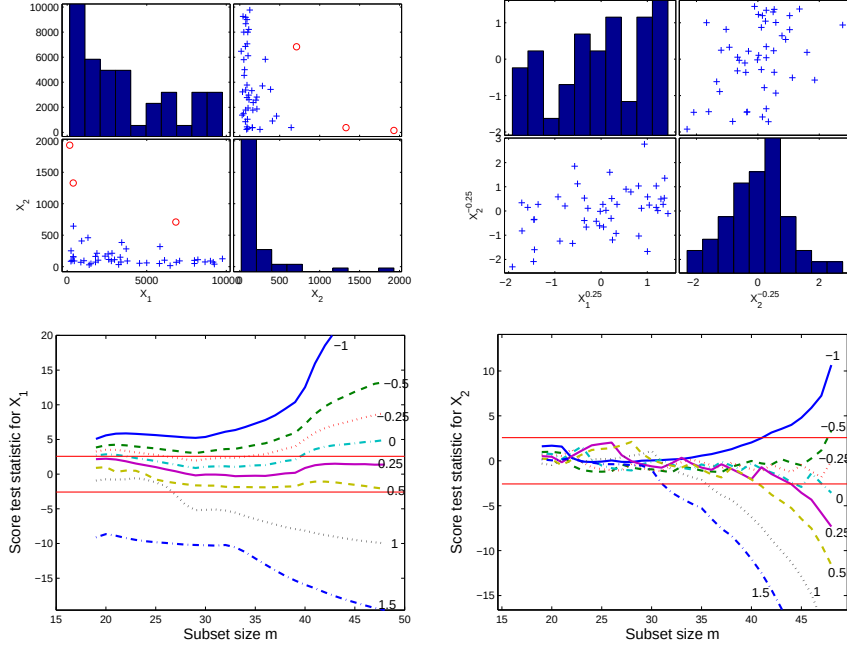


Fig. 1 Agricultural census dataset 1 (48 observations). On the top-left panel, the original dataset with detected outliers (red circles). On the top right panel, the transformed data: no outliers detected. On the bottom, fan plots for variables X_1 (left) and X_2 (right): in both cases at least one trajectory stays well within the envelopes.

for variable X_2 , that we do not report, is useless. The Box-Cox transformation is not appropriate to these data but other transformation models might be tried.

4 Second case study: international trade data

We use a set of trade flows extracted from the Eurostat COMEXT database to present a forth scenario, where the fan plot trajectories do not progress constantly within the Box Cox bands. While the agricultural census data are typically grouped in sets of the order of one hundred observations, international trade data can include thousands of observations. This specific dataset includes five years of monthly aggregates of imports of a fishery product in the European Union. Its main features are the potential presence of multiple populations and the variability increasing from bottom-left to top-right. The variable X_2 is the trade value in thousands of Euros and X_1 is the trade volume in Tons. Given the theoretical relation between trade value and volume, the dataset could be addressed more naturally as regression problem, eventually considering the transformation of only the response (trade value) or of both variables simultaneously (see [1], p.121). However, in what follows we show

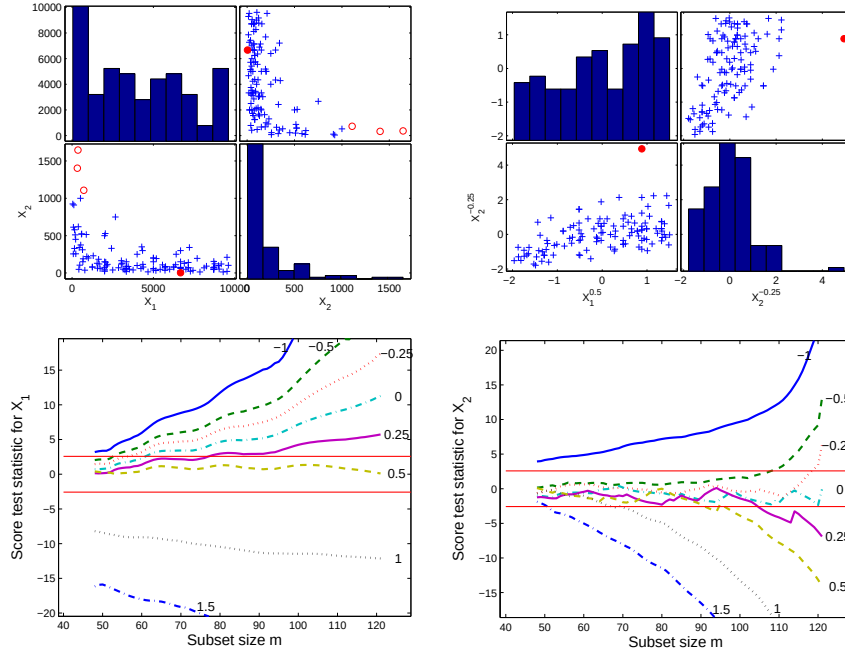


Fig. 2 Agricultural census dataset 2 (121 observations). On the top-left panel the original scale dataset, with 3 outliers (empty red circles). The filled red circle represents the outlier detected in the transformed data (top right panel). On the bottom, fan plots for variables X_1 (left) and X_2 (right). In both cases at least one trajectory stays well within the envelopes.

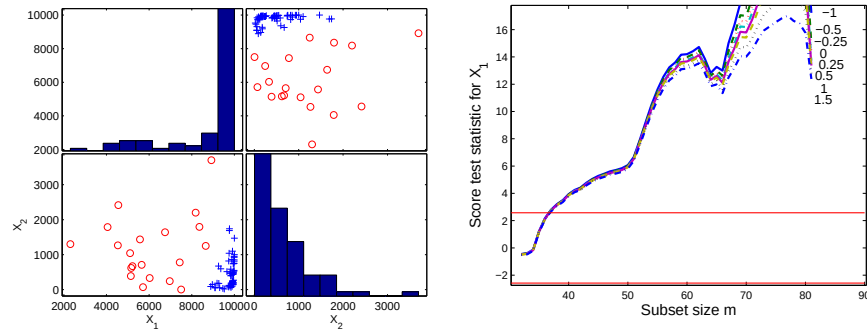


Fig. 3 Agricultural census dataset 3 (81 observations). On the left panel, the original scale dataset with many outliers detected (red circles). On the right panel, fan plot for variable X_1 : since all trajectories deviate from the envelopes, the Box-Cox transformation model is not appropriate.

that a large amount of information can be extracted even by approaching the data as multivariate problem, as in the previous Section, and by transforming the two variables independently, with the results shown in the bottom panels of Figure 4.

There is no trajectory completely inside the bands, but the methodology of Section 2 estimates that a reasonable transformation parameter is the average of the two λ associated with the trajectories lying just above ($\lambda = -0.25$) and just below ($\lambda = 0$) the bands. On the original data the Forward Search identifies a dense homogeneous sub-population and declares abnormal so many observations that the plot, not represented here, becomes uninformative. On the contrary, on the transformed data the Forward Search detects only 12 outliers, most of which almost overlapping near the origin of the axes (top right scatterplot of Figure 4). The top left scatterplot of Figure 4 highlights their position in a zoom of the data in original scale.

One of these 12 outliers is the trade flow with highest unit price in the dataset (110.9 Euro) and the other 11 are flows characterised by few tons of product of extremely low price: their mean price is of 3.6 Euro while the population mean price is of 14.91 Euro. Note that, being the transformation parameter negative, the low prices appear above the bulk of the transformed data while they are obviously in the lower part of the original scale data. The 11 low price outliers are among the 16 lowest price flows in the dataset. The other 5 lowest prices can be easily identified in the scatterplot of the transformed data for their close position to the outliers. The fact that they were not detected as outliers might be interpreted as a failure of the procedure due to the fact that λ was selected by interpolation between two sub-optimal cases.

5 Conclusion

We have discussed a Forward Search based procedure to find multivariate outliers in data that require transformation of the variables. We used datasets ranging from few tens to few thousands of observations, so that to test the statistical and computational properties of the procedure. The four cases were selected to show different transformation scenarios. When the original data deviate from normality assumptions, outliers may be miss-declared and results in transformed data should be more reliable. To test more extensively the procedure we need extending the approach to multivariate joint transformations. This paper has not touched upon the regression context. However we anticipate that if one applies the above procedure only to the response variable of the international trade dataset, no transformation is suggested but then the outliers detected on the basis of the minimum deletion residual (see [1], p. 24) are many and certainly not limited to potential collecting and recording errors. Again, we need extending the approach to the simultaneous transformation of both response and explanatory variables. The procedure can be finally extended to more complex datasets affected by multiple populations along the lines of [7].

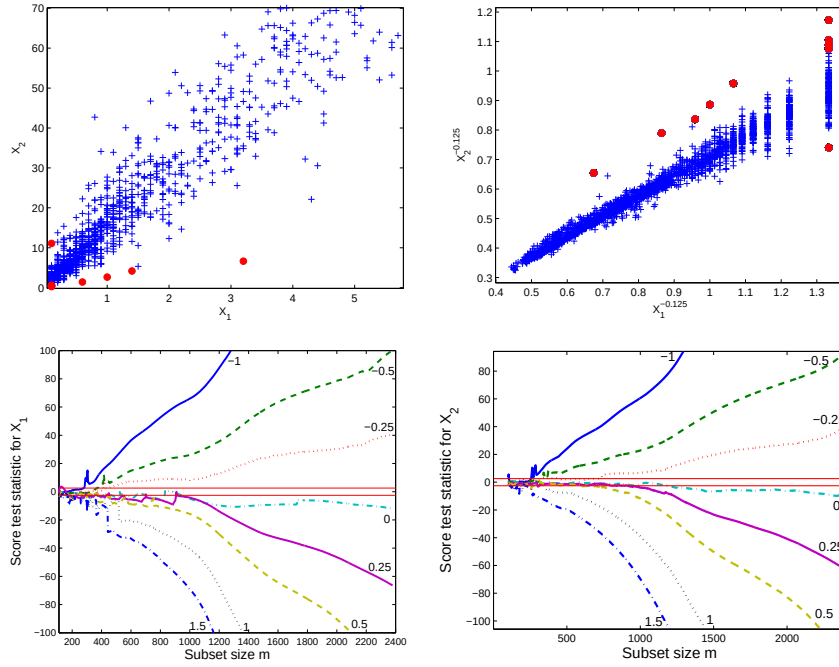


Fig. 4 Fishery data (2379 observations). On the top right panel the scatterplot of the transformed data with 12 outliers, also highlighted in a zoom of the original scale scatterplot (top left panel). On the bottom, fan plots for variable X_1 (left) and X_2 (right): in both cases the transformation value is set to be $\lambda = \text{average}(0; -0.25) = -0.125$, as no trajectory is inside the bands.

References

1. A. C. Atkinson and M. Riani. *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York, 2000.
2. A. C. Atkinson, M. Riani, and A. Cerioli. *Exploring Multivariate Data with the Forward Search*. Springer-Verlag, New York, 2004.
3. G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2):211–252, 1964.
4. A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, 54:761–771, 1992.
5. O. Luzi, T. De Waal, M. Di Zio, U. Guamera, J. Hoogland, B. Hulliger, D. Kilchmann, A. Manzari, J. Pannekoek, and C. Tempelman. Recommended practices for editing and imputation in cross-sectional business surveys. Technical report, Eurostat, 2007.
6. M. Riani, A.C. Atkinson, and A. Cerioli. Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society*, 71:1–21, 2009.
7. M. Riani, A. Cerioli, A. C. Atkinson, D. Perrotta, and F. Torti. *Mining Massive Datasets for Security Applications*, chapter Fitting Robust Mixtures of Regression Lines to European Trade Data. IOS Press, 2008.
8. P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.
9. I.K. Yeo and R.A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.