XXXI–émes Journées de Statistique
Grenoble 17–21 Mai 1999

Two Methods for Graphical Assessment of Multivariate Outliers

Anna Bartkowiak

Institute of Computer Science, University of Wrocław
Przesmyckiego 20, Wrocław 51–151 Poland, e-mail: aba@ii.uni.wroc.pl

1. Methods and Summary. There is a very large literature on detecting multivariate outliers. Most of the proposed methods are analytical – because humans can not see truly points in more then 3 dimensions. On the other hand, the graphical methods are most valuable: they permit to assess the importance of the data points (data vectors) suspected to be outliers. In the paper I show two such graphical methods, which seem to me very valuable: the *grand tour* [2,3] and the *parallel coordinates* plot [5,9]. I compare the results of the two methods when applied to four benchmark data: *The Bradu-Hawkins-Kass* data, the *modified wood gravity* data, the *bushfire* data and the *milk* data. All these data sets have their stories.

Figures 1 and 2 show graphs produced by the two methods. It appears that the two methods are complementing each other. The grand tour permits in an easy way to find suspected outliers; however these need to be verified. The parallel coordinates plot permits to do it. The results coincide with results obtained by the analytical hybrid method elaborated by Woodruff and Rocke (1994).

2. Comments on the results shown in Figure 1 and Figure 2. These figures show graphs obtained typically by the grand tour method and the parallel coordinates plots. The grand tour produced two plots: A) a scatterplot exhibiting momentary point projections with a concentration ellipse superimposed, and B) a count plot exhibiting how many times the given point was notified outside the concentration ellipse. After making several hundred of projections, the entire data set may be subdivided into two subgroups: the data points suspected to be outliers and the remaining ones constituting a 'clean' ('cleaned') data set. To verify and explain this subdivision we draw two parallel coordinates plots exhibiting C) the 'suspected' and D) the 'cleaned' data points.

For all the 4 analyzed data sets the parallel coordinates plots exhibit a different pattern for the cleaned and the suspected subgroups of the data vectors.

The concentration ellipses were drawn using the confidence coefficient $\beta = 0.95$ and a robustified algorithm for evalutions of the variances and the covariance of the projections.

Figure 1, top, shows the results obtained for the **Bradu-Hawkins-Kass data**, which are artificial data constructed with the aim to show influential points in a regression. These data were analyzed by many authors, a.o. by [6,7,8,2]; the source code of the data is included, a.o., in [8]. The projections obtained by the grand tour show clearly that we have to do with two groups of outliers: the points numbered 0–9 and 10–13 (please, note, that our numeration starts from zero!). They are visible in most of the projections. The parallel plot shows, what is the structure of these points (remind, representing individual data vectors) and in which respect they differ from the remaining ones.

Figure 1, bottom, shows the results of the grand tour for the **modified wood gravity data**. The source code of these data was taken from [7]. The tour plots for these data indicate clearly 4 outliers (no. 3, 5, 7, 18) and eventually no. 9, which seems to be simply a point located at the border of the data cloud. The parallel plot confirms the different structure of the data vectors no. 3, 5, 7 and 18. The suspected point no. 9 differs essentially from the 4 indicated outliers; its structure is similar to those of the remaining data set, albeit no. 9 keeps here an extreme position.

Figure 2, top, shows the results for the **bushfire** data. The set contains 38 data vectors (meaning pixels) of satellite measurements on 5 frequency bands (variables) used to locate bushfire scars. I took the source code from [1], where a reference is given to Campbell (1989). The data set was previously analyzed a.o. in [6, 7]. There are two subgroups of outliers: the points $S5=\{6\ 7\ 8\ 9\ 10\}$ which are relatively easy to detect by various robust methods, and the points $S7=\{31\ 32\ 33\ 34\ 35\ 36\ 37\}$ which are masked by the set $S5$ and thus are difficult to detect. The two sets are clearly visible in our plots.

Figure 2, bottom, shows the results for the milk **data**, originally used by Daudin *et al.* [4], elaborated a.o. also in [1] and [3]. Here the grand tour method yields about 17 suspected outliers. The parallel plot drawn for the cleaned set shows clearly that this contains fairly homogeneous data vectors; while the parallel plot drawn for the suspected outliers shows clearly their heterogeneity.

The Bibliography

[1] Atkinson A. (1994), Fast very robust methods for the detection for the detection of multiple outliers. *JASA* 89, 1329–1339.

[2] Bartkowiak A., and Szustalewicz A. (1997). Detecting outliers by a grand tour, *Machine Graphics & Vision* 6, 486–505.

[3] Bartkowiak A., and Szustalewicz A. (1998), Outliers – which genuine and which spurious. In: Ph. Nanopulos, P. Garonna, C. Lauro (Eds), *NTTS'98, International Seminar on New Techniques and Technologies for Statistics, Sorrento, Italy, 4-6 November 1998*, Contributed Papers, 19–24.

[4] Daudin, J.J., Duby C., and Trecourt P. (1988), Stability of principal component analysis studied by the bootstrap method. *Statistics* 19, 241–258.

[5] Inselberg A., (1995), Parallel coordinates for visualizing multidimensional geometry. In: *Eurostat Conference Bonn*, 279–288.

[6] Maronna R.A., and Yohai V.J. (1995), The behaviour of the Stahel-Donoho robust multivariate estimator. *JASA* V.90, 330–341.

[7] Rocke D.M, and Woodruff D.L. (1996), Identification of outliers in multivariate data, JASA 91, 1047–1061.

[8] Rousseeuw P.J., Leroy A. (1987), *Robust Regression and Outlier Detection*, Wiley N.Y.

[9] Wegman E.J. (1990), Hyperdimensional data analysis using parallel coordinates. *JASA 85*, 664-675.

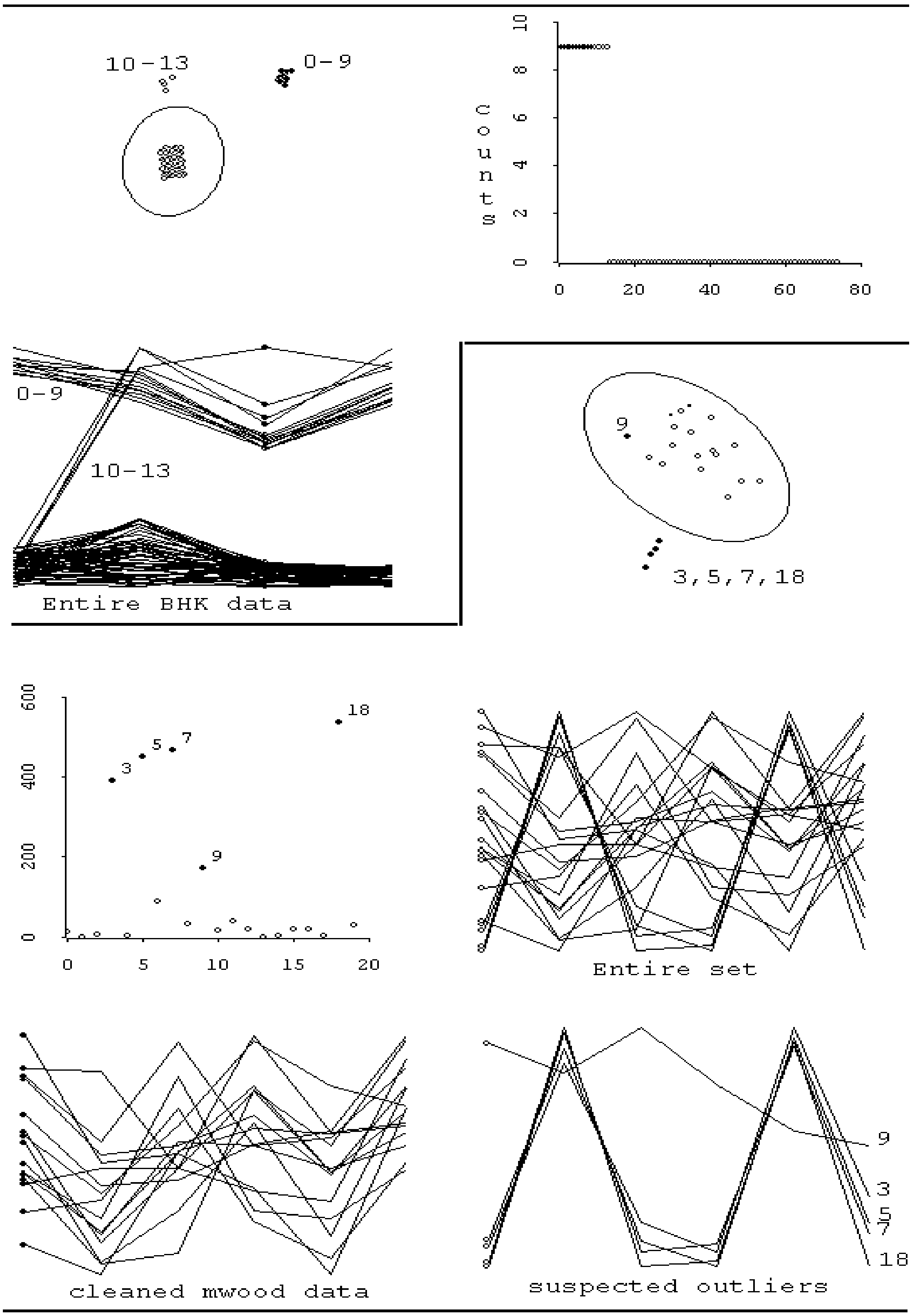

Figure 1. Visualizing multivariate outliers by use of the grand tour and parallel coordinates plots. Top: The Bradu-Hawkins-Kass data, $n = 75$, $p = 4$. Bottom: Modified wood gravity data, $n = 20$, $p = 6$

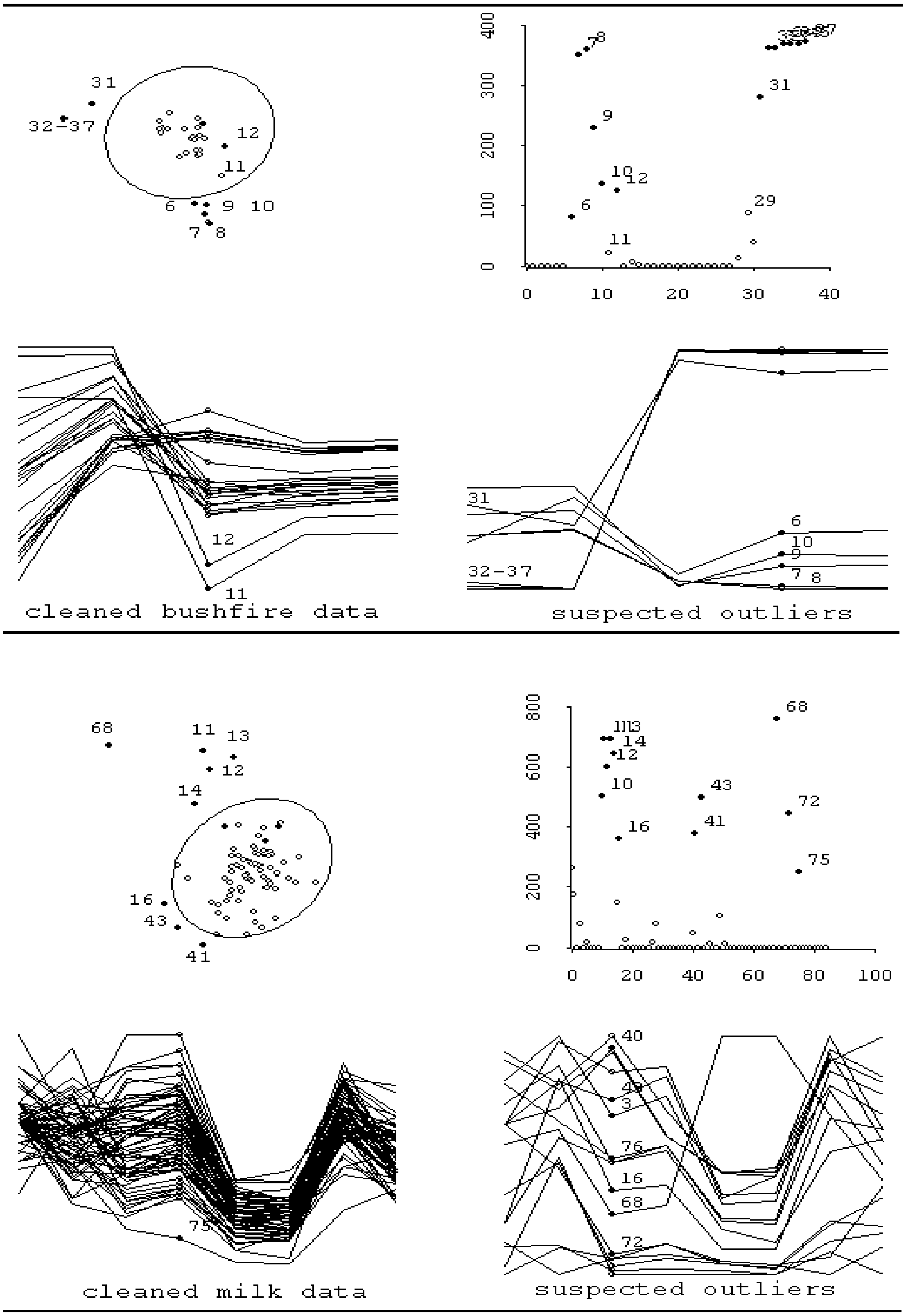

Figure 2. Visualizing multivariate outliers. Top: The Bushfire data $n = 38$, $p = 5$. *Bottom: The Milk data* $n = 85$, $p = 8$.