Robustness in compositional data analysis

M Bren

Institute of Mathematics, Physics and Mechanics, 1 000 Ljubljana, Slovenia and Faculty of Organizationial Sciences, University of Maribor, 4000 Kranj, Slovenia, email: matevz.bren@fov.uni-mb.si

R. Tolosana-Delgado

Department of Sedimentology and Environmental Geology, Georg-August-Universität Göttingen, D-37077, Göttingen, Germany, e-mail: raimon.tolosana@geo.uni-goettingen.de

K. G. van den Boogaart

Institute for Mathematics and Informatics, Ernst-Moritz-Arndt-Universität Greifswald, D-17487, Greifswald, Germany, e-mail: boogaart@uni-greifswald.de

Key words: compositions, outlier, robust statistics, ilr

1 Introduction

The aim of this paper is to offer a classification of outliers in compositional data sets. An atypical compositional datum can be produced in many ways. In this paper we consider those generated: (a) due to gross measurement errors in individual components, (b) because of the presence of a small subpopulation with a different centre (*e.g.*, corresponding to another contrasted but lowly-represented facies), or (c) just by chance (individual atypical data with a low probability, the "pure" outliers). The next sections outline these three types of outliers, in a more convenient order. An artifical example containing all of them is used to check our methods (see Figure 1), and their usefulness in real applications is shown with a data set of rutile geochemistry (see Figure 3). Eynatten et al. (2005) present details about this real example.

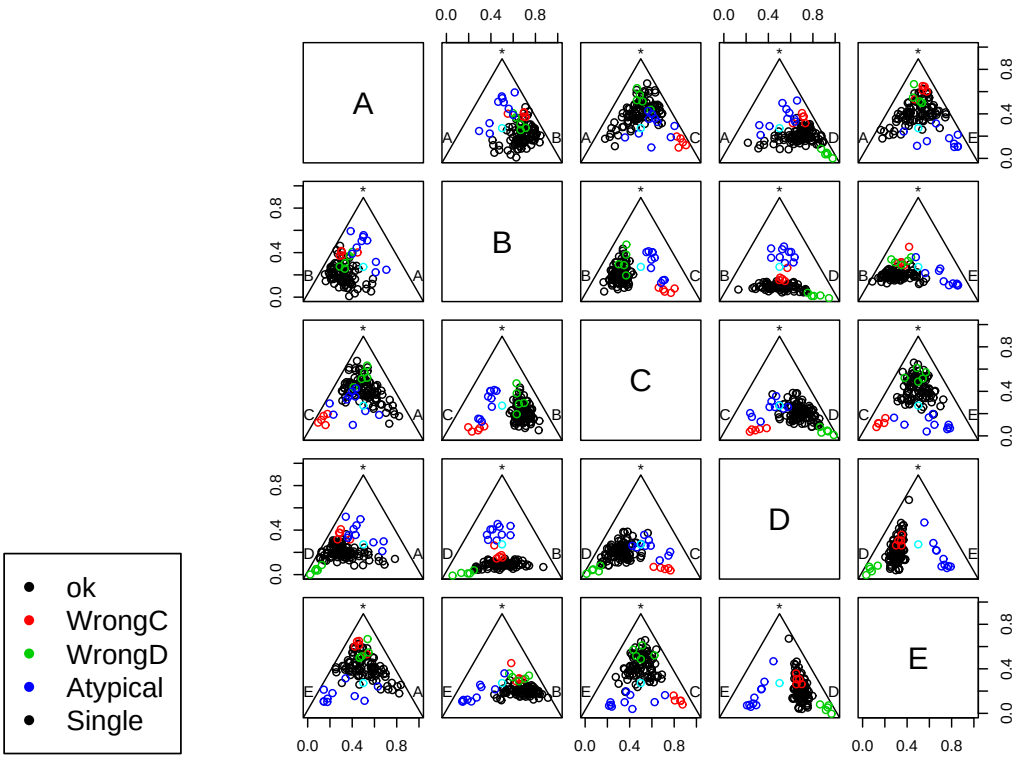

Fig. 1. A simulated example dataset with 3 types of outliers: "ok" are the nonoutlying part of the dataset generated with a multivariate normal distribution on the simplex Mateu-Figueras et al (2003). "WrongC" and "WrongD" represent outliers created by replacing the measurement of component C or D with random values following a lognorm distribution. "Atypical" is a second normal population with the same variance as the main population but with a modified centre. "Single" is a single point precisely in the centre of the main distribution, unrelated to the rest of the data. The data is displayed in a ternary plot matrix using the package "compositions" Boogaart et al. (2006).

2 Detection of outliers

Changing one measured component in a compositional dataset changes all portions. It is thus not possible to judge outliers on the basis of their individual components, because extreme/atypical values in a single component modify the whole composition. However we can use standard outlier detection methods (*e.g.* Rousseeuw and Leroy, 2003, p. 266-270) to detect multivariate outliers in an isometric representation of the data set. This isometric representation is provided by the ilr transform, computed as a set of d log-ratios ($d+1$ is the dimensionality of a composition), and corresponding to the coordinates with respect to an orthonormal basis of the simplex, *a.k.a. ilr basis* (Egozcue et al., 2003; Pawlowksy-Glahn, 2003).

In other words, to classify the atypical individuals in a compositional data set, one can use Mahalanobis distances from a robustly-estimated centre and a robustly-estimated covariance matrix of the ilr-transformed dataset. We used the fast robust MCD-covariance estimator (Rousseeuw and Driessen, 1999) as implemented in Rousseeuw et al. (2006). These distances may be compared to the 0.95-quantile of distribution of the maximum, robustly-estimated, Mahalanobis distance, under the hypothesis of multivariate normality in d dimensions (Hardin and Rocke, 2005, cf.). Then, all points above this limit are considered as outliers. If they cannot be afterwards classified as any of the other two types (as outlined in the next sections), they will be considered of type (c).

3 Detection of directional outliers

Note that, unlike classical multivariate outliers, a directional outlier caused by a measurement error in a single component does not correspond to a canonical direction of the isometric representation. This is so because no canonical basis exists in the simplex, as no ilr basis is one-to-one related to individual components. To decide that an atypical value is explained by a wrong value in a given single component, we check whether it is still an extreme value in the subcomposition without that component. A value is considered extreme if its over the 0.95-quantile of the corresponding χ^2-distribution used for outlier detection in Rousseeuw and Leroy (2003, p.267). If the atypical composition can be explained in this way, it is classified as a single component outlier, or type (a). Table 1(left) shows which individuals of the simulated data set can be explained by a wrong measurement in each single component. Obviously, the components are not always unique, and we select the best-explaining component by choosing the one leading to the lowest Mahalanobis distance of the remaining subcomposition. The result is also reported in table 1(right), and displayed in Figure 2. Although the method is not infallible, it gives a straightforward good classification. Its result with the true data set is reported in Figure 3.

4 Detection of groups

The classification of outliers (Figures 2 and 3) suggests that there might exist groups in the population of outliers, to be taken as type (b). Our proposed way to preliminarily look for them is using cluster analysis of the outliers based on their (robustly-estimated, pairwise) Mahalanobis distances. The obtained dendrogram is given in Figure 4. The dendrogram is cut at a level showing 6

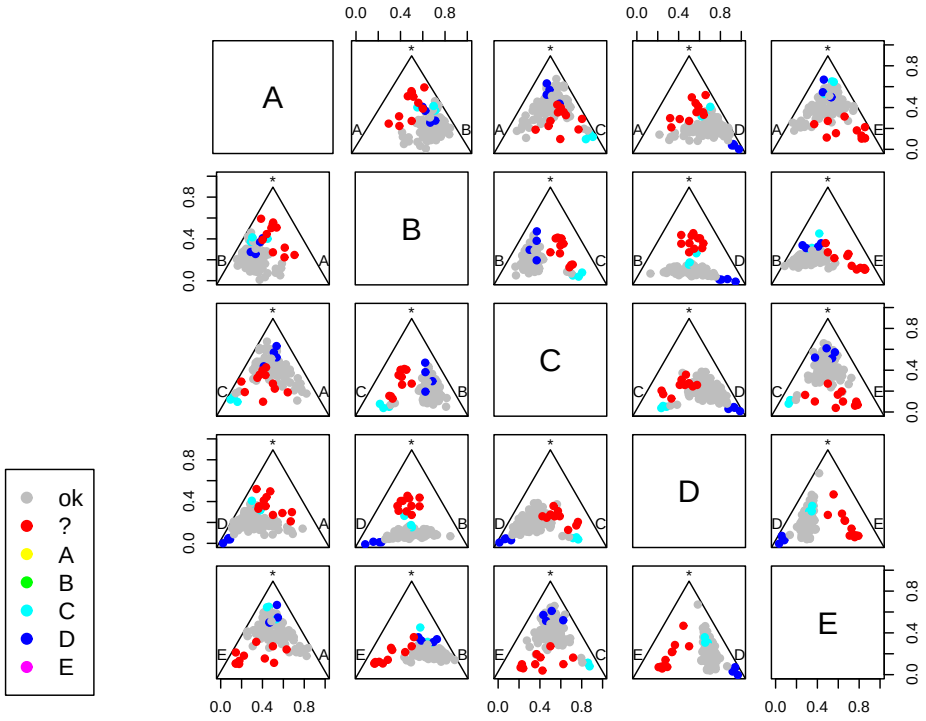

Fig. 2. Classification of simulated data set subgroups according to the final component considered as cause of outlier. See table 1 and section 3 for further explanations.

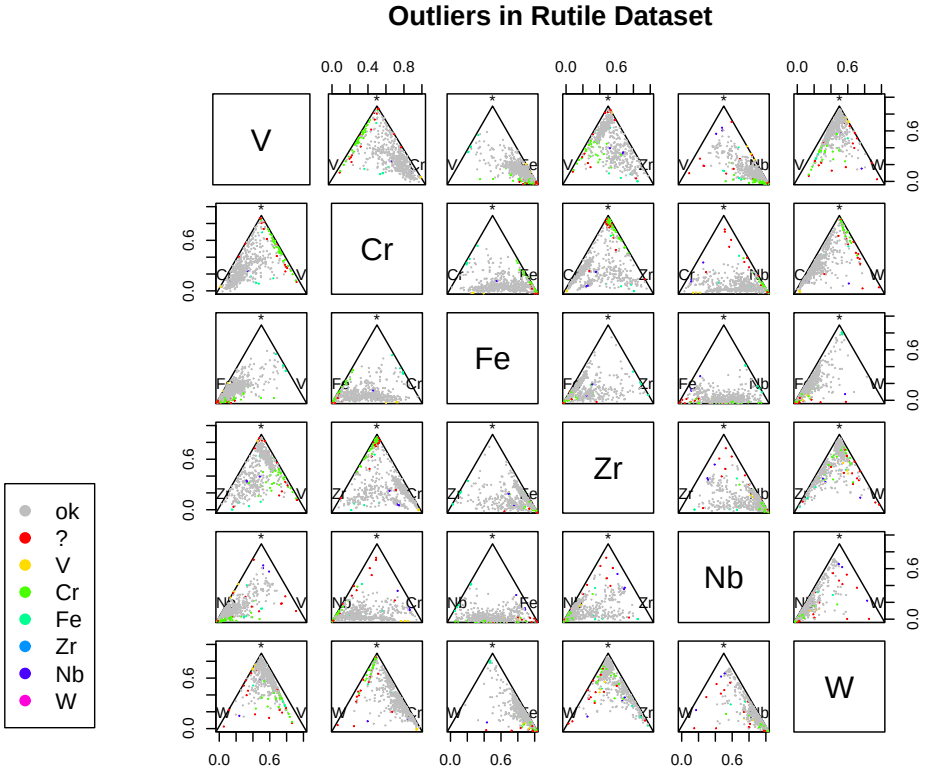

Fig. 3. The outlier classification of section 3 applied to the rutile dataset.

	ok	00000	00010	00100	00110	ok	?	A	B	C	D	E
ok	90	0	0	0	0	90	0	0	0	0	0	0
WrongC	0	0	0	1	4	0	0	0	0	5	0	0
WrongD	3	0	1	0	1	3	0	0	0	0	2	0
Atypical	0	10	0	0	0	0	10	0	0	0	0	0
Single	0	1	0	0	0	0	1	0	0	0	0	0

Table 1
Classification of the simulated data set subgroups according to whether they are explained as single component outliers. The first part of the table shows this as a bitcode, with 0 or 1 for each part in the composition: 1 says that the subcompostion without the given component is typical (thus the atypicality is explained by an outlier in that component). Often multiple components can explain an outlier, *e.g.* some of the outlier in component C could also be explained as an outlier in B. The ouliers with bitcode 00000 cannot be explained by any single component outlier. The second part of the table reports the component finally considered as cause of outlier. Type "?" corresponds to those classified as 00000.

subpopulations of outliers from left to right: two subsets characterised by no single atypical component, a large set of chromium outliers, and three groups respectively characterised by atypical iron, vanadium and niobium values.

5 Conclusions

A combination of classic robust multivariate statistics, isometric transformations of compositions, and conceptual models of "typical" classes of outliers allow a straightforward detection *and classification* of compositional outliers, offering a good overview of the studied data set. Such a study should complement any standard exploratory data analysis, as it provides insights on possible subgroups and erroneous measurements.

References

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279–300.

Eynatten, H.v., R. Tolosana-Delgado, S. Triebold, and T. Zack, 2005. Interactions between grain size and composition of sediments: two examples. In: Mateu-Figueras, G. and C. Barceló-Vidal, eds. 2nd Compositional Data Analysis Workshop – CoDaWork'05, Proceedings. Universitat de Girona. Girona, Spain.

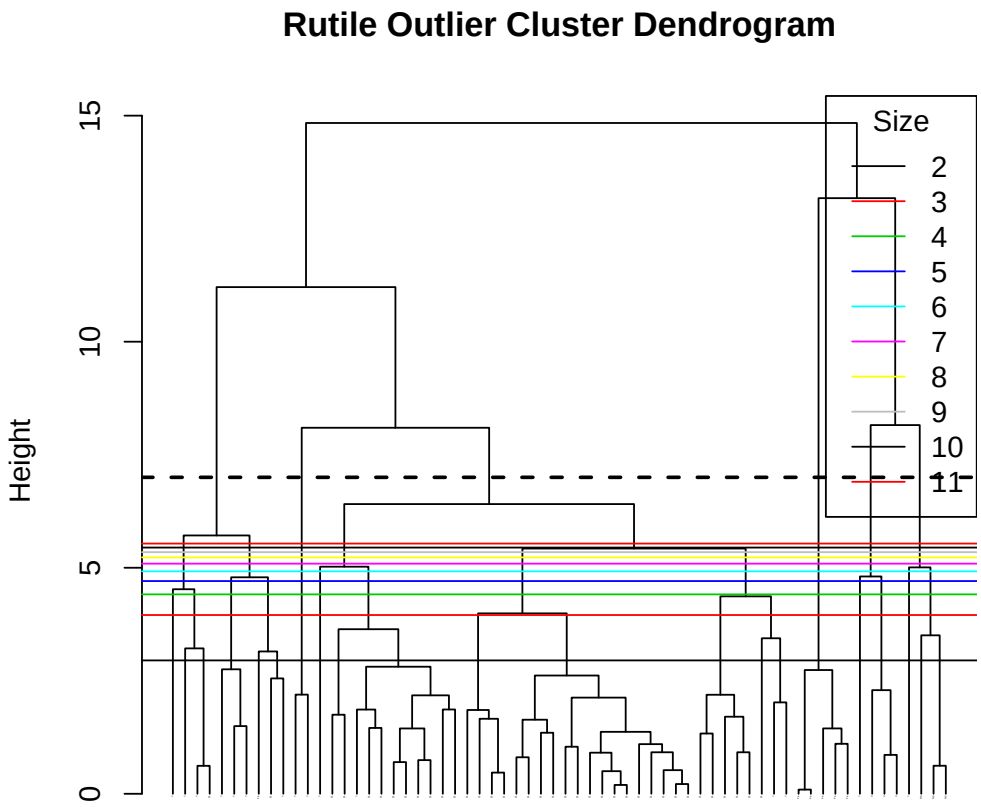

Fig. 4. Dendrogram of a clustering of outliers in the rutile data set using complete linkage and robustly-estimated Mahalanobis distances.

Hardin, J. and D.M. Rocke, 2005. The distribution of robust distances, Journal of Computational and Graphical Statistics, **14**, 928-946

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal, 2003. Distributions on the simplex. In: Thió-Henestrosa, S. and J. A. Martín-Fernández (2003)

Pawlowsky-Glahn, V., 2003. The statistical analysis on coordinates. In: Thió-Henestrosa, S. and J. A. Martín-Fernández (2003).

Rousseeuw, P. J. and K. van Driessen, 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223.

Rousseeuw, P. J. and A. M. Leroy, 2003. Robust Regression and Outlier Detection., Wiley, 329p.

Rousseeuw, P. J., K. van Driessen and V. Todorov, 2006. rrcov: Scalable Robust Estimators with High Breakdown Point, R-package version 0.3-05

Boogaart, K. G. v. d., R.Tolosana-Delgado and M. Bren, 2006. compositions: Compositional Data Analysis, R package version 0.91-6

Thió-Henestrosa, S. and J. A. Martín-Fernández, (Eds.) 1st Compositional Data Analysis Workshop – CoDaWork'03, Proceedings. Universitat de Girona. Girona, Spain.