An Interactive Tool for Residual Diagnostics for Fitting Spatial Dependencies (with Implementation in R)

Ernst Glatzer
Österreichische Nationalbank

Werner G. Müller
University of Economics and B.A. Vienna

Abstract

There exists a growing number of spatial statistics software packages covering a broad range of methods including the classical technique of kriging following variogram fitting (for a review see e.g. Bivand and Gebhardt, 2000). However, most of these programs do not contain tools for assessing the fit of particular models of the spatial dependencies. Based upon a variogram cloud estimation method proposed in Müller, 1999 we present various techniques for that purpose and their implementations in an R-package called *vardiag*. The paper concentrates on the interactive aspect of the problem worked out in the dissertation of Glatzer, 2002. For the interactive exploration of the fit several plots (a map view, the square-root-differences cloud and two types of residual plots) are linked. The proposed tool allows brushing of single points or sets of points in these plots. The selection of points in one plot is reflected by an automatic selection of the corresponding parts in the other plots. The program is exemplified on a data set of chlorid concentrations in the Südliche Tullnerfeld.

1 Introduction

The characterization of spatial dependencies is an essential component of the analysis of isotropic random fields $Z(s)$, $s \in \mathbb{R}^2$. The so-called square-root-differences cloud (cf. Cressie, 1993 or Ploner, 1999) serves as a convenient display of such dependencies. For all location pairs it plots the $|Z(s_i + h_i) - Z(s_i)|^{\frac{1}{2}} = \gamma_i$ against their distances $\|h_i\|$, where $i = 1, \ldots, \binom{n}{2}$.

Typically this cloud exhibits the following shape: the entries at short distances tend to be low with low variation, whereas with increasing distance the entries and their variation increase. To determine a functional form of the spatial dependency

one usually fits a parametric model $\gamma(\theta)$ to the square-root-differences cloud (see a generic example in Figure 1).
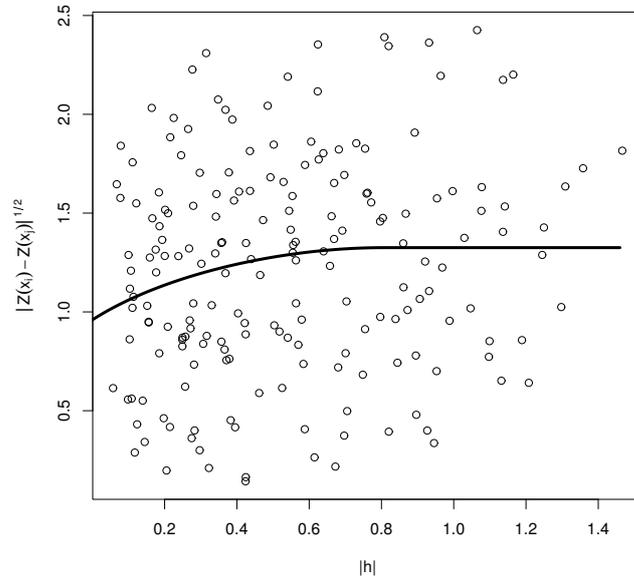

Figure 1: Square-root-differences cloud with a fitted parametric model (solid line).

Following the technique proposed in Müller, 1999 for variogram clouds we will estimate the parameter θ by feasible generalized least squares, i.e.

$$\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\gamma,$$

where $\hat{V}$ is a suitable estimate of the variance-covariance matrix of the residuals $e_i = \gamma_i - \gamma_i(\hat{\theta})$ and $\gamma^T = (\gamma_1, \ldots, \gamma_i, \ldots)$. The matrix X contains the regressors from an adequately linearized model.

2   Residual diagnostics

Studentized square-root-differences cloud

To assess the goodness of the fit of the regression and to identify potential outliers one can – rather than to employ the original residuals e_i – use the studentized residuals

$$r_i = \frac{e_i}{\sqrt{[\hat{V} - X(X'\hat{V}^{-1}X)^{-1}X']_{ii}}}.$$

With these studentized residuals one can construct a studentized version of the square-root-differences cloud. For every pair of observations we plot the sum of the forecasted value and a corresponding (rescaled) studentized residual, i.e.

$$\{h_i, \gamma(\hat{\theta}) + r_i\hat{\sigma}\}.$$

where $\hat{\sigma} = \sqrt{\sum e_i^2 / \left[\binom{n}{2} - p\right]}$ is an estimate for the average standard deviation of the error term.

In case of a correctly specified model and all assumptions met this plot should exhibit an approximately symmetric band around the estimated variogram - the function representing the spatial dependency. One possible deviation would be nonconstant error variance, i.e. varying width of the error band along the horizontal axis. If the functional form of the model does not correspond to the data, the cloud will exhibit a different curvature than the fitted model.

A second feature of this plot is the easy identification of outliers. Studentized residuals should - in case of normal errors - be normally distributed as well. Entries in the cloud that are far from the fitted curve are thus highly suspicious.

Leave-one-out residual plot

Such entries can be double-checked by using a second plot that instead of the conventional residuals employs the so-called leave-one-out residuals

$$e_{[i]} = \gamma_i - \gamma_i(\hat{\theta}_{[i]}).$$

Here $\hat{\theta}_{[i]}$ is the parameter estimate one receives when the i-th observation is dropped. It is clear that these types of residuals are more capable of sensing outliers (as well as influential observations for that matter).

The construction of this second type of residuals is however not entirely unproblematic in our case, since a single entry in the square-root-differences cloud can be affected by leaving out one of a couple of original observations. Now we have for each entry in the square-root-differences cloud one conventional residual and two leave-one-out residuals. We suggest to plot these two sets of leave-one-out residuals against the conventional residuals and compare their positions relative to the first meridian.

Decorrelated residual cloud

Finally we would like to take into account that the entries in the square-root-differences cloud are highly correlated. The model was fitted by generalized least squares. That is equivalent to an ordinary least squares fit of a transformed model

$$\Lambda |Z(s_i + h_i) - Z(s_i)|^{\frac{1}{2}} = \Lambda \gamma_i(\theta) + \Lambda \epsilon,$$

where Λ is the Cholesky decomposition such that $\Lambda'\Lambda = \hat{V}^{-1}$. The transformed residuals $\Lambda\epsilon$ are now uncorrelated and can be plotted against the transformed predictions $\Lambda\gamma_i(\hat{\theta})$ to indicate inadequacies.

Other diagnostic tools that can be employed in this context are described in Haslett *et al.*, 1991 and Barry, 1996.

3   Brushing and linking

As is well disseminated in the literature (see e.g. Bradley and Haslett, 1992 or Buja *et al.*, 1996) the interactive nature in the form of brushing and linking is essential for proper exploratory spatial data analysis. The tools for such an analysis - based

upon the methods described in the previous sections - are part of the R-package *vardiag* which can be downloaded from CRAN:

http://CRAN.R-project.org/src/contrib/PACKAGES.html

We have already mentioned that one entry in the square-root-differences cloud is based on two original observations. When checking suspicious entries in the square-root-differences cloud it is important to know where the two observations are located in physical space. The proposed tool allows selection of single entries or sets of entries. When entries are brushed, the corresponding points in the linked map view are highlighted automatically. This is a convenient way to find out whether a suspicious entry (corresponding to a high square-rooted difference of observations) is based on observations near the border of the region of interest or whether it points to a region of higher variation within the region.

Since a large square-rooted difference of observations can have several reasons, it is important to check other diagnostic quantities as well. This could be for instance the leave-one-out residuals. When an entry in the square-root-differences cloud is brushed, the corresponding leave-one-out residuals are also highlighted in the linked plot. This procedure allows different kinds of views onto the data and especially onto suspicious points.

As an example we use a data set of chlorid concentrations in the Südliche Tull-nerfeld measured daily at 20 locations. The data are already detrended and trans-formed for distributional symmetry. Then an appropriate model $\gamma(\theta)$ for the spatial dependence is estimated. The estimated model for this example together with resid-uals and the variance-covariance matrix of the parameters are also contained in the package *vardiag*. The package further contains the borders of the region of interest as a matrix of coordinates of vertices.

A typical analysis

We begin our analysis by starting R, loading the library *vardiag* and loading the data in form of a variogram object and the region matrix.

```
> library(vardiag)
> data(tulln)
```

Next we display our set of four diagnostic plots as given in Figure 2:

```
> PlotDiag.varobj(vs50,tu1)
```

The argument `vs50` identifies the data object, the argument `tu1` the region object.

Now we look for suspicious entries in the square-root-differences cloud. After having identified one such point we initiate the interactive diagnostics:

```
> interact.varobj(vs50,tu1,"s")
```

After invoking this command the cursor changes to a crosshair signifying that a point can be selected. We place the cursor near the suspicious point and click the left mouse button. The result is that the suspicious point is now colored magenta. The pair of original observations is joined by a magenta line in the map view. One of these observations is colored blue and all entries in the square-root-differences cloud corresponding to this observation are also colored blue. Likewise are all leave-one-out residuals based on this observation colored blue. Analogously the other observation and all corresponding entries are colored red. Equivalently by issuing the command

```
> interact.varobj(vs50,tu1,"m")
```

a pair of points can be marked in the map view.

After repeating this procedure several times for different suspicious entries, we could brush a suspicious subregion in the map view by

```
> interact.varobj(vs50,tu1,"n")
```

and identifying a polygon by a sequence of left mouse-clicks on the desired vertices. Similarly by issuing

```
> interact.varobj(vs50,tu1,"t")
```

a polygon can also be brushed in the square-root-differences cloud.

By issuing the command

```
> interact.varobj(vs50,tu1,"x",pchi=0.05)
```

all points outside a confidence region in the square-root-differences cloud are automatically selected.

Additionally we can select points from the leave-one-out plot by using

```
> interact.varobj(vs50,tu1,"l")
```
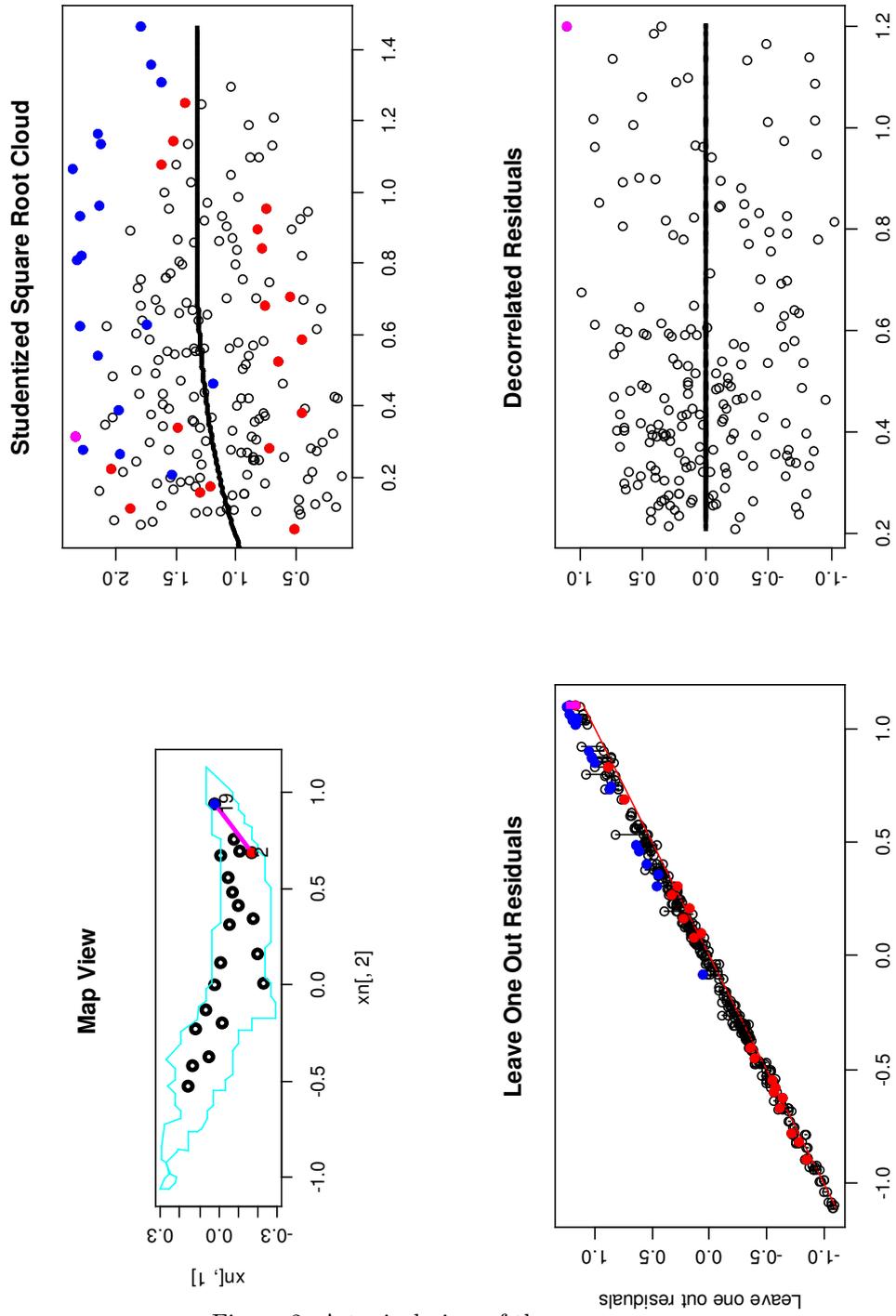
Figure 2: A typical view of the screen.

References

Barry, R.P. (1996), "A Diagnostic to Assess the Fit of a Variogram Model to Spatial Data", *Journal of Statistical Software,* **1** (1), 1-11.

Bivand, R. S. and Gebhardt, A. (2000), "Implementing functions for spatial statistical analysis using the R language", *Journal of Geographical Systems,* 2(3), 307-317.

Bradley, R. und Haslett, J. (1992), "High-interaction diagnostics for geostatistical models of spatially referenced data" *The Statistician,* **41**, 371-380.

Buja, A., Cook D. and Swayne, D. (1996), "Interactive high-dimensional data visualization", *Journal of Computational and Graphical Statistics* 5, 78–99.

Cressie, N.A.C. (1993) *Statistics for Spatial Data.* John Wiley and Sons, New York.

Glatzer, E. (2002), "Residualdiagnostiken für die Variogrammschätzung" (in German), unpublished dissertation at the Vienna University of Economics and Business Administration.

Haslett, J., Bradley, R., Craig, P., Unwin, A., und Wills, G. (1991). Dynamic Graphics for Exploring Spatial Data With Application to Locating Global and Local Anomalies. *The American Statistician,* **45**(3), 234-242.

Müller, W.G. (1999), "Least-squares fitting from the variogram cloud", *Statistics & Probability Letters*, 43, 93-98.

Ploner, A. (1999) "The use of the variogram cloud in geostatistical modelling", *Environmetrics,* 10, 413-437.

Corresponding author

Werner G. Müller
Department of Statistics
University of Economics and Business Administration
Augasse 2–6
A-1090 Vienna
Austria
E-mail: werner.mueller@wu-wien.ac.at