Visualizing Influential Observations in Dependent Data

Marc G. GENTON and Anne RUIZ-GAZEN

We introduce the hair-plot to visualize influential observations in dependent data. It consists of all trajectories of the value of an estimator when each observation is modified in turn by an additive perturbation. We define two measures of influence: the local influence which describes the rate of departure from the original estimate due to a small perturbation of each observation; and the asymptotic influence which indicates the influence on the original estimate of the most extreme contamination for each observation. The cases of estimators defined as quadratic forms or ratios of quadratic forms are investigated in detail. Sample autocovariances, covariograms, and variograms belong to the first case. Sample autocorrelations, correlograms, and indices of spatial autocorrelation such as Moran's I belong to the second case. We illustrate our approach on various datasets from time series analysis and spatial statistics. This article has supplementary material online.

Key Words: Autocovariance; Moran's I; Outlier; Quadratic form; Robustness; Space; Time; Variogram.

1. INTRODUCTION

The main goal of this article is to propose new plots in order to visualize influential observations in the context of dependent data. For a given estimator, we say that an observation is influential whenever a change in its value leads to a radical change in the estimate. This influence analysis approach, based on the study of the effect of data perturbation on estimators, has been already widely studied in the literature. In the context of linear regression, Cook (1977) proposed several influence analyses according to different types of perturbations; see also the article by Cook and Weisberg (1980). It appears that the most natural perturbation setting for independent observations consists in deleting observations. Statistics such as Cook's distance and its associated plot (Cook's distances versus case numbers) are well-known diagnostics based on the case deletion idea in the

Marc G. Genton is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *genton@stat.tamu.edu*). Anne Ruiz-Gazen is Professor, Toulouse School of Economics, Université des Sciences Sociales, 21 Allée de Brienne, 31000 Toulouse, France (E-mail: *ruiz@cict.fr*).

linear regression context. Influence analysis based on deleting observations has also been studied in the context of linear regression when observations are dependent. Martin (1992) derived several diagnostics including Cook's distance for generalized least squares estimators in a linear model where the error covariance matrix is assumed known or fixed up to a multiplicative scalar. Haining (1994) followed Martin's proposal and calculated diagnostics for the spatial autoregressive model which is well-known in spatial econometrics and corresponds to a particular covariance matrix structure. These authors considered several applications but did not make any proposal for graphical influence diagnostics.

Another tool for detecting influential observations but not focused on linear regression is the Empirical Influence Function (EIF) proposed by Hampel et al. (1986); see also the work of Welsh (1982). There exist two versions of the EIF: one is based on the replacement of an observation while the other consists in adding an observation. The addition of an observation is closely related to the idea of removing an observation. A plot of the EIF versus case numbers is suggested in the work of Hampel et al. (1986) but one must admit that the EIF is not frequently used as a tool for detecting influential observations even in the field of robust statistics. Indeed, in the independent case, the influence function is generally preferred. Based on the definition of an estimator as a functional, it gives a general form of the influence of an infinitesimal contamination and does not depend on a particular sample. But in the case of dependent observations, the definition of an influence function involves the joint distribution function and is not straightforward. In the time series context, for example, there exist several possible definitions (Maronna, Martin, and Yohai 2006).

The use of an empirical influence function as an exploratory tool for influence analysis is therefore of particular interest for dependent data especially if some easy-to-interpret plots are made available to the data analyst. Our goal is to offer simple tools in order to detect and analyze influential cases. If some observations are highly influential, the way to deal with them depends on the context but we may recommend several possible actions. First, influential observations have to be carefully checked in order to be corrected if they are gross-errors. If, however, such observations happen to be correct, we advise the use of robust techniques. For dependent data, we claim that deleting or adding observations is not the most natural perturbation setting anymore. If observations represent countries, as in one of the examples we consider in Section 4 for instance, it is not easy to justify that one country be removed since the truth is that it still exists. It is even less plausible to add an observation which corresponds to the creation of a new country. As stated by Cook and Weisberg (1994, p. 212), deleting cases is only one way of introducing small changes in the data and there are others. In the sequel, we focus on additive perturbations. This perturbation setting permits to vary the amount of perturbation and to display the corresponding changes of the estimate. Therefore, apart from the fact that deleting observations may be questionable in the context of dependent data, the additive perturbation setting we propose gives more insight into the behavior of estimators by considering several possible amounts of perturbation.

Consider an estimator $\widehat{\theta}(\mathbf{Z})$ of a parameter θ based on a data vector $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$. In order to study influential observations on this estimator, we define a perturbation of $\mathbf{Z}$ by $\mathbf{Z}[i, \zeta] = \mathbf{Z} + \zeta \mathbf{e}_i$, where $\mathbf{e}_i$ has a nonzero component only at index i, at which it is 1,

and ζ represents the value of the perturbation. The effect of influential observations can be visualized with a plot of each $\widehat{\theta}(\mathbf{Z}[i, \zeta])$, $i = 1, \ldots, n$, as a function of ζ. We coin such a graphical representation a hair-plot. A hair-plot is a version of the empirical influence function with replacement (see Hampel et al. 1986, p. 93) and with a particular parameterization of the perturbation. The perturbation ζ is added to the original sample so that at $\zeta = 0$, the original value $\widehat{\theta}(\mathbf{Z})$ is recovered for any $i = 1, \ldots, n$. Typically ζ is a real number but for certain applications, such as for a positive variable, the range of ζ can be restricted to ensure sensible values of the variable of interest. Instead of plotting case numbers on the horizontal axis as for Cook's distance plot or the EIF plot, each observation is plotted as a curve in the hair-plot. As stated above, in the case of dependent data, the version of the empirical influence function with replacement of an observation is preferable to the one with addition of an observation. Indeed, in the context of time series or spatial data, there is no obvious way to define a new instant or a new location of observation. This is in stark contrast to the classical case of independent and identically distributed observations where such an issue does not arise.

Associated to the hair-plot, two influential measures are then of interest. First, we define the local influence of the ith observation on the estimator $\widehat{\theta}(\mathbf{Z})$ as

$$\tau_i(\widehat{\theta}, \mathbf{Z}) = \left. \frac{\partial}{\partial \zeta} \widehat{\theta}(\mathbf{Z}[i, \zeta]) \right|_{\zeta=0}. \tag{1.1}$$

It describes the rate of departure from the value $\widehat{\theta}(\mathbf{Z})$ for each observation due to a small perturbation. Hence, the most influential observations correspond to the largest absolute values of $\tau_i(\widehat{\theta}, \mathbf{Z})$. Second, we define the asymptotic influence of the ith observation on the estimator $\widehat{\theta}(\mathbf{Z})$ as

$$\nu_i(\widehat{\theta}, \mathbf{Z}) = \lim_{\zeta \to \infty} \widehat{\theta}(\mathbf{Z}[i, \zeta]). \tag{1.2}$$

It indicates the influence on the value $\widehat{\theta}(\mathbf{Z})$ of the most extreme contamination for each observation. In general the function $\widehat{\theta}(\mathbf{Z}[i, \zeta])$ of ζ depends on i, the exception being the case where $\widehat{\theta}$ is the sample mean or a function of it, as we show next.

Proposition 1. *Let f be a function from $\mathbb{R}^n$ to $\mathbb{R}$ such that for any $z_1, z_2, \ldots, z_n$ in $\mathbb{R}$, any $i, j = 1, \ldots, n$, and ζ in $\mathbb{R}$,*

$$f(z_1, \ldots, z_{i-1}, z_i + \zeta, z_{i+1}, \ldots, z_j, \ldots, z_n)$$
$$= f(z_1, \ldots, z_i, \ldots, z_{j-1}, z_j + \zeta, z_{j+1}, \ldots, z_n).$$

Then f is a function of $\sum_{i=1}^{n} z_i$.

The proof of this result is given in the Appendix.

We have implemented a command `hair.plot(data,thetahat,...)` in R (R Development Core Team 2010) that is available as supplemental material on the *JCGS* website. It produces a hair-plot and allows to identify each observation (i.e., each hair) in the plot. For illustration, we consider a dataset of $n = 91$ monthly interest rates of an Austrian bank. Künsch (1984), Ma and Genton (2000), Azzalini and Genton (2008), and Wang

et al. (2009) have previously studied this dataset in the context of robust time series analysis and noted the presence of three large outliers for the months 18, 28, and 29. Because these authors have argued that an autoregressive model of order 1, AR(1), is appropriate for these data, we focus on the lag-one sample autocorrelation $\widehat{r}(1) = 0.78$ obtained from

$$\widehat{r}(h) = \frac{\sum_{i=1}^{n-h}(Z_{i+h} - \bar{Z})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}, \qquad 0 \leq h \leq n - 1,$$

where $\bar{Z}$ is the sample mean. The top panel of Figure 1 depicts the hair-plot of $\widehat{r}(1)$ for $\zeta \in [-3, 3]$ and allows to identify any curve, that is, any observation. It reveals that the months 17, 18, 19, 27, and 30 are quite influential. The month 76 is also identified to be somewhat influential. The bottom panel of Figure 1 presents a disc-plot for $\zeta > 0$ of the monthly interest rates time series. The radii of the discs are proportional to the rate of departure from $\widehat{r}(1) = 0.78$ due to a small contamination at each observation. Open discs denote an increase of the value $\widehat{r}(1)$ whereas closed discs denote a decrease. The discs with large radii correspond to the most influential observations. The months 4 and 76 are also seen to be somewhat influential.

The article is organized as follows. In Section 2, we study the local and asymptotic influence of estimators that are defined as quadratic forms in the data vector, whereas we concentrate on ratios of quadratic forms in Section 3. Sample autocovariances, covariograms, and variograms belong to the former. Sample autocorrelations, correlograms, and indices of spatial autocorrelation such as Moran's I belong to the latter. In Section 4, we illustrate our approach on applications to pollution data and to African conflict data. We compare the latter with an approach based on deleting observations in turn. We end with a discussion in Section 5 where we propose the use of the hair-plot on robust estimators for dependent data. We also discuss extensions to the case of multiple simultaneous influential observations.

2. INFLUENCE ON QUADRATIC FORMS

We study the effect of an influential observation on estimators defined as quadratic forms in the data vector, that is, $\widehat{\theta}_Q(\mathbf{Z}, \mathbf{A}) = \mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z}$, where $\mathbf{A} = (a_{ij})$ is an $n \times n$ matrix. It follows that, under the contamination scheme $\mathbf{Z}[i, \zeta]$, we have

$$\widehat{\theta}_Q(\mathbf{Z}[i, \zeta], \mathbf{A}) = \mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z} + (\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{Z})\zeta + (\mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{e}_i)\zeta^2.$$

Therefore, the local and asymptotic influences (1.1) and (1.2) of the ith observation are, respectively,

$$\tau_i(\widehat{\theta}_Q, \mathbf{Z}) = \mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{Z},$$

$$v_i(\widehat{\theta}_Q, \mathbf{Z}) = \infty. \tag{2.1}$$

If the matrix $\mathbf{A}$ is symmetric, then the local influence reduces to $\tau_i(\widehat{\theta}_Q, \mathbf{Z}) = 2\mathbf{a}_i^{\mathrm{T}}\mathbf{Z}$ where $\mathbf{a}_i$ is the ith row of the matrix $\mathbf{A}$.
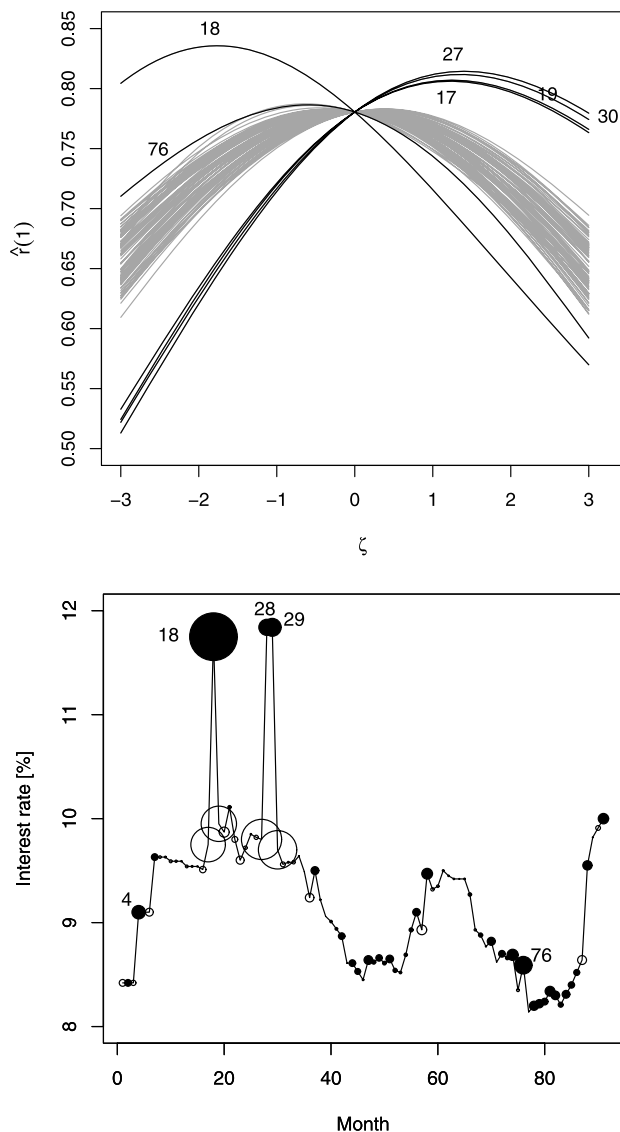
Figure 1.   Hair-plot and disc-plot (for $\zeta > 0$) of the lag-one sample autocorrelation $\widehat{r}(1)$ on the monthly interest rates dataset.

In time series analysis, the sample autocovariance function is an important tool defined at time lag h by

$$\widehat{c}(h) = \frac{1}{n} \sum_{i=1}^{n-h} (Z_{i+h} - \bar{Z})(Z_i - \bar{Z}), \qquad 0 \le h \le n - 1.$$

It can also be expressed as a quadratic form in the vector of time series data. The corresponding matrix is $\mathbf{A} = \frac{1}{n}\mathbf{H}\mathbf{D}(h)\mathbf{H}$ where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is the symmetric and idempotent ($\mathbf{H}^2 = \mathbf{H}$) centering matrix with $\mathbf{I}_n$ the $n \times n$ identity matrix and $\mathbf{1}_n$ the n-vector of ones.

The matrix $\mathbf{D}(h) = \frac{1}{2}(\mathbf{P}(h) + \mathbf{P}(h)^{\mathrm{T}})$ is the temporal design matrix at lag h, where $\mathbf{P}(h)$ is an $n \times n$ matrix with ones on the hth upper diagonal and zeroes elsewhere, $1 \leq h \leq n - 1$, and $\mathbf{P}(0) = \mathbf{I}_n$; see the article by Genton (1999). Therefore, the matrix $\mathbf{A}$ is symmetric and the local influence of the ith observation on the sample autocovariance function takes the form $\tau_i(\widehat{c}(h), \mathbf{Z}) = 2\mathbf{a}_i^{\mathrm{T}}\mathbf{Z}$ with the jth component of the vector $\mathbf{a}_i$ given by

$$a_{ij} = \frac{1}{n}\left\{ d_{ij}(h) - \frac{1}{n}d_{i\cdot}(h) - \frac{1}{n}d_{\cdot j}(h) + \frac{1}{n^2}d_{\cdot\cdot}(h) \right\}, \tag{2.2}$$

where $\mathbf{D}(h) = (d_{ij}(h))$ with $d_{i\cdot}(h) = \sum_{j=1}^{n} d_{ij}(h)$, $d_{\cdot j}(h) = \sum_{i=1}^{n} d_{ij}(h)$, and $d_{\cdot\cdot}(h) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}(h)$. The local influence changes as a function of the temporal lag h. From the particular form of the matrix $\mathbf{D}(h)$, it follows that $d_{\cdot\cdot}(h) = n - h$. Moreover, if $1 \leq h \leq \lfloor \frac{n}{2} \rfloor$, then $d_{i\cdot}(h) = d_{\cdot i}(h) = 1$ for $h < i < n - h$ and $d_{i\cdot}(h) = d_{\cdot i}(h) = \frac{1}{2}$ otherwise. If $\lfloor \frac{n}{2} \rfloor < h \leq n - 1$, then $d_{i\cdot}(h) = d_{\cdot i}(h) = 0$ for $h < i < n - h$ and $d_{i\cdot}(h) = d_{\cdot i}(h) = 1$ otherwise.

Similarly in spatial analysis, the sample covariogram is defined by

$$\widehat{c}(\mathbf{h}) = \frac{1}{n}\sum_{(i,j)\in N_{\mathbf{h}}} (Z_i - \bar{Z})(Z_j - \bar{Z}),$$

where $N_{\mathbf{h}}$ is the set of spatial locations separated by the lag vector $\mathbf{h}$. It can be rewritten as a quadratic form in the vector of data with $\mathbf{A} = \frac{1}{n}\mathbf{H}\mathbf{D}(\mathbf{h})\mathbf{H}$ where now $\mathbf{h}$ is a lag vector in space. Thus, the description of the matrix $\mathbf{D}(\mathbf{h})$ is more involved, but (2.2) still holds and expressions for its terms can be derived. The sample variogram is an alternative tool for measuring spatial dependence and is defined by

$$2\widehat{\gamma}(\mathbf{h}) = \frac{1}{|N_{\mathbf{h}}|}\sum_{(i,j)\in N_{\mathbf{h}}} (Z_i - Z_j)^2,$$

where $|N_{\mathbf{h}}|$ is the cardinality of the set $N_{\mathbf{h}}$. It can also be defined as a quadratic form in the vector of data with $\mathbf{A} = \mathbf{D}^*(\mathbf{h})$. Here, the specific form of $\mathbf{D}^*(\mathbf{h})$ is different from the one of $\mathbf{D}(\mathbf{h})$; see the work of Genton (1998a), Gorsich, Genton, and Strang (2002), and Hillier and Martellosio (2006). The local influence for the sample variogram is still given by (2.1) with $\mathbf{D}(h)$ replaced by $\mathbf{D}^*(\mathbf{h})$ in (2.2).

3. INFLUENCE ON RATIOS OF QUADRATIC FORMS

We investigate the effect of an influential observation on estimators defined as a ratio of two quadratic forms, for example such as the lag-one autocorrelation studied in the Introduction. Traditionally, the asymptotic influence of an observation has been to push the value of the estimator to the edge of the parameter space. However, Genton and Lucas (2003, 2005) and Genton (2003) have shown that this need not be the case for time series and spatial statistics settings. Instead, the estimator is sometimes pushed toward the center of the parameter space by a single outlying value. In Figure 1 we have seen the behavior of the lag-one autocorrelation for perturbations ζ within the interval $[-3, 3]$. How will

an estimator defined as a ratio of two quadratic forms based on the contaminated sample $\mathbf{Z}[i, \zeta]$ be affected when ζ becomes large?

To this end, we investigate the behavior of an estimator defined by

$$\widehat{\theta}_{\mathrm{RQ}}(\mathbf{Z}, \mathbf{A}, \mathbf{B}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z}}{\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{Z}},$$

where $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ are $n \times n$ matrices, under the contamination scheme $\mathbf{Z}[i, \zeta]$. It follows that

$$\widehat{\theta}_{\mathrm{RQ}}(\mathbf{Z}[i, \zeta], \mathbf{A}, \mathbf{B}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z} + (\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{Z})\zeta + (\mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{e}_i)\zeta^2}{\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{Z} + (\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{B}\mathbf{Z})\zeta + (\mathbf{e}_i^{\mathrm{T}}\mathbf{B}\mathbf{e}_i)\zeta^2}.$$

Therefore, the local and asymptotic influence of the ith observation are, respectively,

$$\tau_i(\widehat{\theta}_{\mathrm{RQ}}, \mathbf{Z}) = \frac{(\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{Z})(\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{Z}) - (\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z})(\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{e}_i + \mathbf{e}_i^{\mathrm{T}}\mathbf{B}\mathbf{Z})}{(\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{Z})^2},$$

$$\nu_i(\widehat{\theta}_{\mathrm{RQ}}, \mathbf{Z}) = \frac{a_{ii}}{b_{ii}},$$

provided that $b_{ii} \neq 0$. Hence, the asymptotic influence on the estimator is dictated by the ratio of the iith entries of the matrices $\mathbf{A}$ and $\mathbf{B}$. If the matrices $\mathbf{A}$ and $\mathbf{B}$ are symmetric, then the local influence reduces to

$$\tau_i(\widehat{\theta}_{\mathrm{RQ}}, \mathbf{Z}) = \frac{2}{\mathbf{Z}^{\mathrm{T}}\mathbf{B}\mathbf{Z}}(\mathbf{a}_i - \widehat{\theta}_{\mathrm{RQ}}\mathbf{b}_i)^{\mathrm{T}}\mathbf{Z}, \tag{3.1}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the ith rows of the matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. We study these quantities for various estimators below.

Returning to time series analysis, the sample autocorrelation function studied in the Introduction can be written as a ratio of quadratic forms in the data vector with $\mathbf{A} = \mathbf{H}\mathbf{D}(h)\mathbf{H}$ and $\mathbf{B} = \mathbf{H}$, where $\mathbf{H}$ and $\mathbf{D}(h)$ were defined in Section 2. Hence, its local influence is given by (3.1) where $\mathbf{a}_i$ and $\mathbf{b}_i$ can be derived based on (2.2). Its asymptotic influence is simply

$$\nu_i(\widehat{r}(h), \mathbf{Z}) = \frac{n - h - 2nd_{i\cdot}(h)}{n(n-1)}, \tag{3.2}$$

where $d_{i\cdot}(h)$ is defined after (2.2). Note that $\lim_{n\to\infty} \nu_i(\widehat{r}(h), \mathbf{Z}) = 0$ for any observation i. Similarly in spatial analysis, the sample correlogram can be written as a quadratic form in the vector of data with $\mathbf{A} = \mathbf{H}\mathbf{D}(\mathbf{h})\mathbf{H}$ and $\mathbf{B} = \mathbf{H}$, where now $\mathbf{h}$ is a lag vector in space.

The presence of spatial dependence in data on a lattice is often assessed by means of a statistic such as Moran's I (Moran 1950). If $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ is a spatial sample of dimension n, that is, Z_i represents an observation at the location i on the lattice, then Moran's I is defined by

$$\widehat{I}(\mathbf{Z}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}, \tag{3.3}$$

where the spatial structure matrix $\mathbf{W} = (w_{ij})$ is of dimension $n \times n$ and contains nonnegative weights describing the degree of interaction between neighbor locations in the plane. The spatial structure matrix $\mathbf{W}$ does not need to be symmetric but can be transformed to

symmetry by $(\mathbf{W} + \mathbf{W}^{\mathrm{T}})/2$ without changing the value of Moran's I defined in (3.3). Usually, the matrix $\mathbf{W}$ is obtained from a particular coding scheme in order to stabilize the heterogeneity resulting from different degrees of interaction between the spatial observations. In particular, one can distinguish between the globally standardized C-coding scheme, the row-sum standardized W-coding scheme, and the variance stabilizing S-coding scheme; see, for example, the book by Tiefelsdorf (2000). For all three coding schemes, the resulting matrix $\mathbf{W}$ has the sum of its elements equal to n. The W-coding scheme is often used in practice and implies that $\mathbf{W}\mathbf{1}_n = \mathbf{1}_n$. Under normality of $\mathbf{Z}$, the expectation of Moran's I under the hypothesis of independence is given by $\mathrm{E}(\widehat{I}(\mathbf{Z})) = -1/(n-1)$; see, for example, the book by Cliff and Ord (1981, p. 44).

For the case of Moran's I defined in (3.3), we have again a ratio of quadratic forms in the data vector with $\mathbf{A} = \mathbf{HWH}$ and $\mathbf{B} = \mathbf{H}$. It follows that the local influence on Moran's I can be derived similarly to the correlogram above. Regarding the asymptotic influence, we have $a_{ii} = \frac{1}{n}(w_{..}/n - w_{\cdot i} - w_{i\cdot})$ and $b_{ii} = 1 - 1/n$, where $w_{..} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $w_{\cdot i} = \sum_{j=1}^{n} w_{ji}$, and $w_{i\cdot} = \sum_{j=1}^{n} w_{ij}$. Hence

$$v_i(\widehat{I}, \mathbf{Z}) = \frac{1}{n-1}(w_{..}/n - w_{\cdot i} - w_{i\cdot}). \tag{3.4}$$

Consequently, as $\zeta \to \infty$, we have $\widehat{I}(\mathbf{Z}[i, \zeta]) \to v_i(\widehat{I}, \mathbf{Z})$ whatever the true value of I, that is, whatever the realization of the uncontaminated sample $\mathbf{Z}$. Moran's I no longer conveys any useful information on I and the estimate is totally dictated by the contamination. Therefore, according to Genton and Lucas (2003, 2005), the breakdown-point of Moran's I defined in (3.3) is zero. The precise value of the asymptotic influence $v_i(\widehat{I}, \mathbf{Z})$ given by (3.4) depends on the coding scheme. For all three coding schemes listed above, we have $w_{..} = n$. By noticing that $0 \le w_{\cdot i} + w_{i\cdot} \le n$, we obtain for the C-coding and S-coding schemes that $-1 \le v_i(\widehat{I}, \mathbf{Z}) \le 1/(n-1)$, whereas for the W-coding scheme we have $-1 \le v_i(\widehat{I}, \mathbf{Z}) \le 0$. In the latter case, $v_i(\widehat{I}, \mathbf{Z}) = -w_{\cdot i}/(n-1)$. If the observations are collected on a regular grid and the edge effects are neglected, then approximately we have $w_{\cdot i} = 1$. In this case, Moran's I estimator breaks down to $-1/(n-1)$, the expected value of Moran's I under the null hypothesis of independence, with one extreme contamination. Note that $-1/(n-1)$ is not at the edge of the parameter space. In addition, the aforementioned behavior of Moran's I for large sample size n will break the estimator toward 0. For finite and infinite ζ, there is dependence on $\mathbf{W}$ since ζ will affect several local averages depending on the structure of $\mathbf{W}$. Finally, when $\bar{Z} = 0$, we have $\mathbf{A} = \mathbf{W}$ and $\mathbf{B} = \mathbf{I}_n$ for Moran's I, hence $a_{ii} = w_{ii} = 0$ and $b_{ii} = 1$. Therefore, $v_i(\widehat{I}, \mathbf{Z}) = 0$ in that case for any n.

Li, Calder, and Cressie (2007) have put forward an alternative closed-form measure of spatial autocorrelation defined as an approximate profile likelihood estimator (APLE) of the spatial dependence parameter of a spatial autoregressive (SAR) model. The estimator, a ratio of two quadratic forms, is defined by

$$\widehat{\mathrm{APLE}}(\mathbf{Z}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{H}[(\mathbf{W} + \mathbf{W}^{\mathrm{T}})/2]\mathbf{H}\mathbf{Z}}{\mathbf{Z}^{\mathrm{T}}\mathbf{H}[\mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathrm{tr}(\mathbf{W}^2)\mathbf{I}_n/n]\mathbf{H}\mathbf{Z}}, \tag{3.5}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. Note that (3.5) is a slight extension of the original definition of Li, Calder, and Cressie (2007) who set $\bar{Z} = 0$. We claim also that their

use of $\boldsymbol{\lambda}^T\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is the explicit vector of eigenvalues of $\mathbf{W}$, is unnecessary since $\boldsymbol{\lambda}^T\boldsymbol{\lambda} = \mathrm{tr}(\mathbf{W}^2)$. For the case of APLE defined in (3.5), we have $\mathbf{A} = \mathbf{H}[(\mathbf{W} + \mathbf{W}^T)/2]\mathbf{H}$ and $\mathbf{B} = \mathbf{H}[\mathbf{W}^T\mathbf{W} + \mathrm{tr}(\mathbf{W}^2)\mathbf{I}_n/n]\mathbf{H}$, where again $\mathbf{H}$ is the centering matrix. Therefore, $a_{ii} = \frac{1}{n}(w_{..}/n - w_{.i} - w_{i.})$ and $b_{ii} = \sum_k w_{ki}^2 + \mathrm{tr}(\mathbf{W}^2)(n-1)/n^2 - 4(\sum_k w_{ki}w_{k.})/n - 2(\sum_k w_{k.}^2)/n^2$. Hence

$$v_i(\widehat{\mathrm{APLE}}, \mathbf{Z}) = \frac{w_{..}/n - w_{.i} - w_{i.}}{n\sum_k w_{ki}^2 + \mathrm{tr}(\mathbf{W}^2)(n-1)/n - 4(\sum_k w_{ki}w_{k.}) - 2(\sum_k w_{k.}^2)/n}.$$

Here again, as $\zeta \to \infty$, we have $\widehat{\mathrm{APLE}}(\mathbf{Z}[i,\zeta]) \to v_i(\widehat{\mathrm{APLE}}, \mathbf{Z})$ whatever the true value of APLE, that is, whatever the realization of the uncontaminated sample $\mathbf{Z}$. Therefore, according to Genton and Lucas (2003, 2005), the breakdown-point of the APLE statistic defined in (3.5) is zero. It is more difficult to give bounds for $v_i(\widehat{\mathrm{APLE}}, \mathbf{Z})$ depending on the coding scheme of $\mathbf{W}$, but clearly for large n we have $v_i(\widehat{\mathrm{APLE}}, \mathbf{Z}) \to 0$. When $\bar{Z} = 0$, we have $\mathbf{A} = (\mathbf{W} + \mathbf{W}^T)/2$ and $\mathbf{B} = \mathbf{W}^T\mathbf{W} + \mathrm{tr}(\mathbf{W}^2)\mathbf{I}_n/n$ for the APLE statistic, hence $a_{ii} = 0$ and $b_{ii} = \sum_k w_{ki}^2 + \mathrm{tr}(\mathbf{W}^2)/n$. Therefore, $v_i(\widehat{\mathrm{APLE}}, \mathbf{Z}) = 0$ in that case for any n.

4. APPLICATIONS

4.1 POLLUTION DATA

We consider the application of our methodology to the analysis of pollution levels arising from the pumping of waste material into the English Channel and reported by Haining (1990, p. 217). The dataset consists of reflectance values extracted from an aerial survey and located on a regular 9×9 spatial lattice. High levels of pollution induce high reflectance values. Following a previous study of this pollution data by Haining (1987), a linear trend in the reflectance values is first removed, yielding residuals $Z(s_1, s_2)$, $s_1, s_2 = 1, \ldots, 9$. Our interest is now in modeling the possible dependence structure in these residual values. To this end, we compute the empirical variogram $2\widehat{\gamma}(h)$ at spatial lag distances $h = 1, 2, 3, 4$ assuming isotropy.

Genton and Ronchetti (2003) have noticed several possible outliers in the residual values at locations $(1, 1)$, $(1, 2)$, $(2, 2)$, $(2, 7)$, and $(7, 5)$ on the lattice, and the largest residual takes the value 40 at location $(2, 2)$. For this reason, we investigate the influence of each residual on the variogram estimator. Figure 2 depicts hair-plots of the sample variogram $2\widehat{\gamma}(h)$ on the reflectance residual values for $h = 1, 2, 3, 4$ and $\zeta \in [-40, 40]$. The variogram estimates on the original data at each lag h are identified by a closed black disc. The influence of the largest residual (the observation #17) is identified by the black curve. Notice that the influence changes from one lag to another. For example, for $\zeta > 0$, the observation #17 is the most influential at lag $h = 1$, but not at other lags. This fact has been the motivation of Genton (1998b) for the definition of a spatial breakdown-point of variogram estimators; see also the work of Lark (2008) for recent discussions on this topic. In addition, as can be seen in Figure 2, the effect of a perturbation ζ results in a quadratic departure.

The influence of each observation can be depicted spatially. Figure 3 presents a disc-plot for $\zeta > 0$ of the sample variogram $2\widehat{\gamma}(h)$ on the reflectance residual values for each
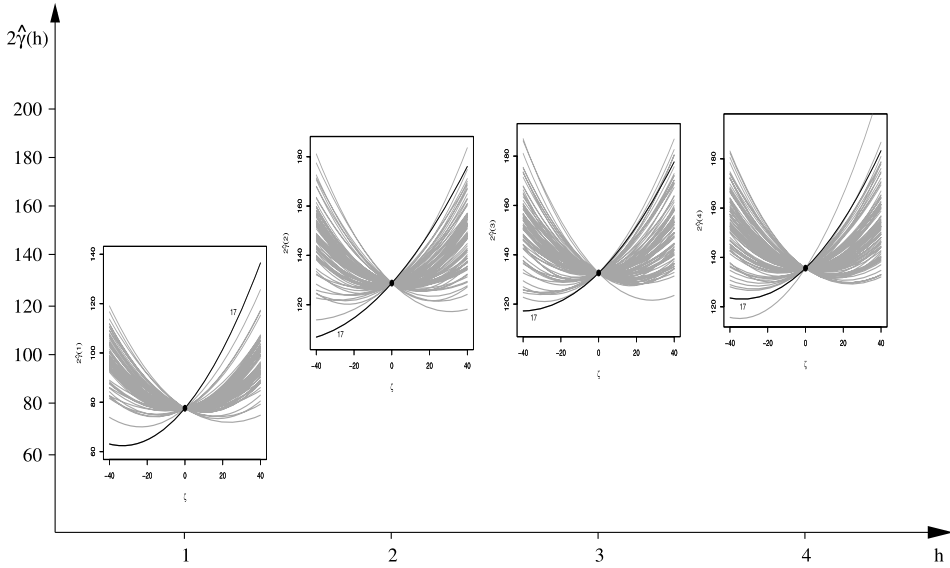
Figure 2.   Hair-plots of the sample variogram $2\widehat{\gamma}(h)$ on the reflectance residual values for spatial lag distance $h = 1, 2, 3, 4$.

$h = 1, 2, 3, 4$. The radii of the discs are proportional to the rate of departure from $2\widehat{\gamma}(h)$ due to a small contamination at each observation. Open discs denote an increase of the value $2\widehat{\gamma}(h)$ whereas closed discs denote a decrease. The discs with large radii correspond to the most influential observations. Here also we can see that the influence changes from one lag to another. However, the observation #17, located at $(2, 2)$, is overall influential. In fact, the influence can be summarized over the various lags, for instance with a mean influence. This information is presented in Figure 4.

When the locations of the observations are irregularly spaced, then the form of the matrix **A** in the quadratic form defining the sample variogram estimator becomes more complex. In that situation, the visualization of influential observations by means of the hair-plot is even more useful.

4.2   AFRICAN CONFLICT DATA

The second application we consider is from the field of economics. The dataset consists of a standardized measure of total conflict for 42 African countries. It was used by Anselin (1995) to study the importance of spatial effects in the statistical analysis of international conflicts. Following Anselin (1995), we consider a W-coding spatial structure matrix **W** based on the first-order contiguity (common border) in order to calculate the Moran index. For this conflict measure, Moran's I equals 0.417 and strongly evidences positive spatial autocorrelation. Anselin (1995) proposed to use a local Moran indicator as a diagnostic for outliers with respect to the global measure of spatial association. These local indicators are such that their sum equals the global Moran index. Such indicators are an alternative to analyze the influence of observations by giving their contribution to the
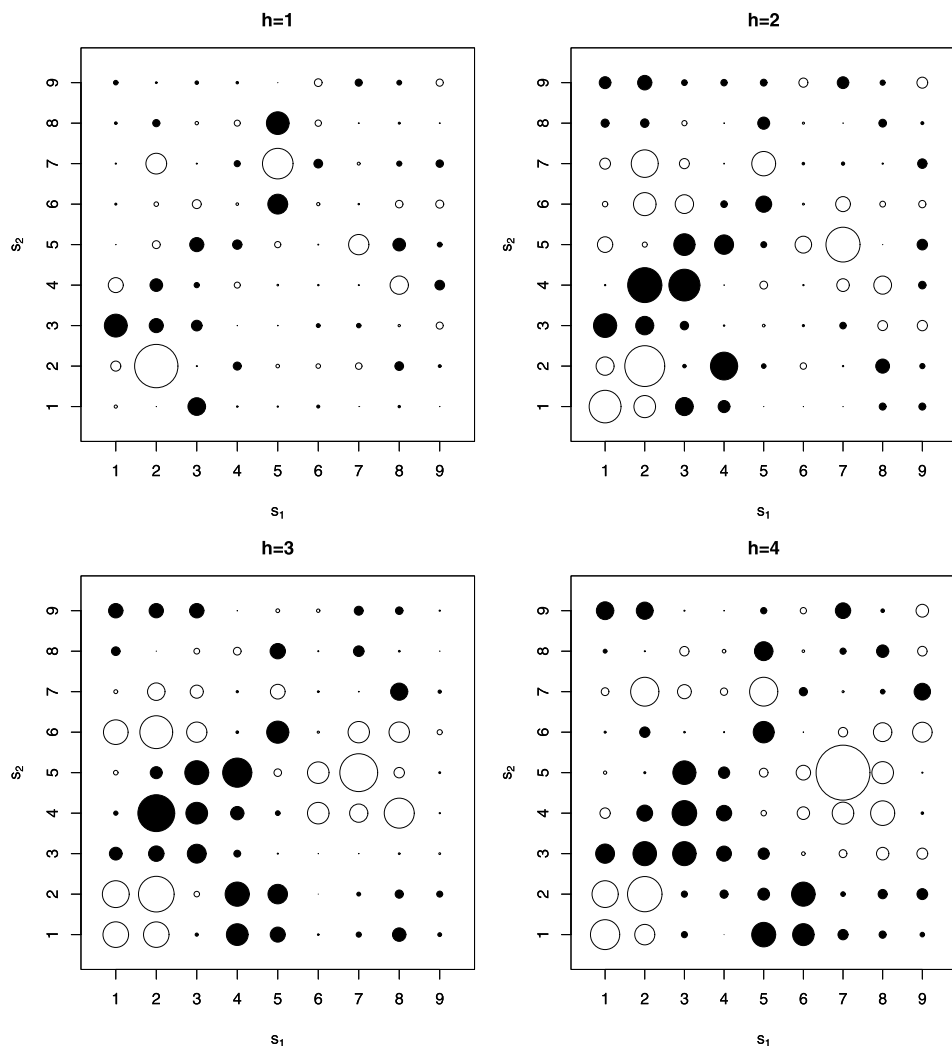
Figure 3.   Disc-plot for $\zeta > 0$ of the sample variogram $2\widehat{\gamma}(h)$ on the reflectance residual values for lag distances $h = 1, 2, 3, 4$.

global measure. In the present example, Sudan and Egypt have highly significant local spatial association. With our proposal, we give further insight into the behavior of the Moran index when adding perturbations to these particular observations below. Figure 5 displays the hair-plot of this autocorrelation index with four countries identified by black curves and labels. South Africa, and to a minor extent Senegal, are among the most influential units for any $\zeta \in [-10, 10]$. More precisely, the hair-plot reveals that a positive contamination ζ, that is, a larger total conflict measure, for these countries leads to a rapid decrease of the autocorrelation index while a low negative contamination leads to an increase of the index. Sudan is also particularly influential when considering large negative ζ values leading to a negative Moran index. Finally, Egypt is not very influential when considering a low level of contamination ζ (positive or negative) but the change in the Moran index for large negative
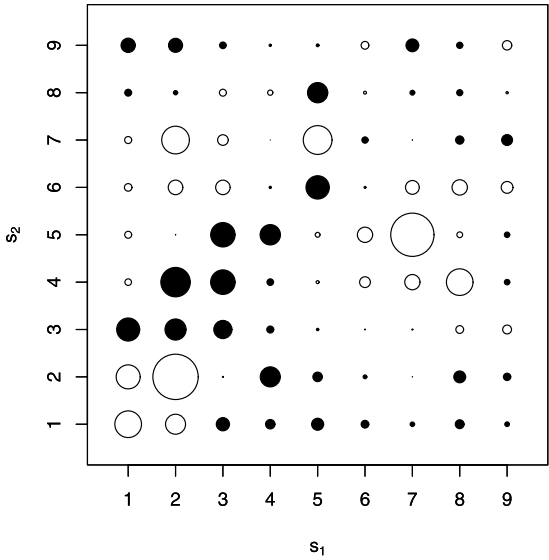
Figure 4. Disc-plot for $\zeta > 0$ of the sample variogram $2\widehat{\gamma}(h)$ on the reflectance residual values, averaged over lag distances $h = 1, 2, 3, 4$.

ζ is very steep and noticeable. Unlike in the first application, the effect of ζ is now a ratio of two quadratics. Furthermore, depending on the specific observation, the maximum of the curve is shifted to the left or to the right leading to different influential behaviors. The
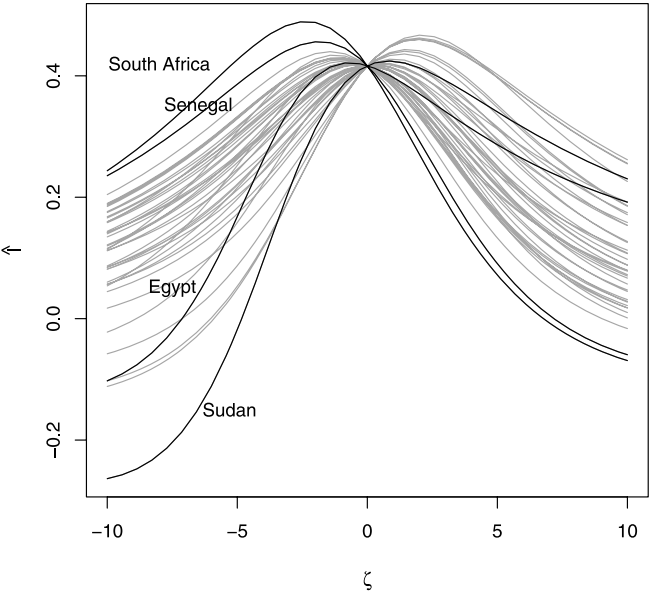

Figure 5. Hair-plot of Moran's I estimator $\widehat{I}$ for the African conflict data.

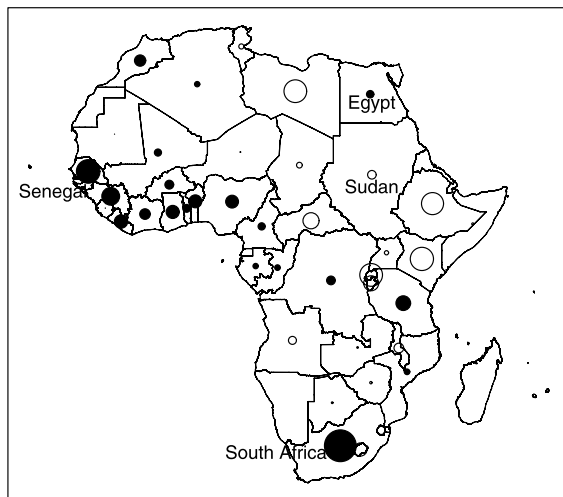**Local Sensitivity to Positive Contamination**

Figure 6. Disc-plot ($\zeta > 0$) of Moran's I estimator $\widehat{I}$ for the African conflict data.

asymptotic influence of the ith African country is $\nu_i(\widehat{I}, \mathbf{Z}) = -w_{\cdot i}/41$ and ranges from -0.066 (South Africa) to -0.003 (Lesotho).

Figure 6 displays a disc-plot of the Moran index overlaid on the Africa map. As in the previous application, open discs denote an increase of the value of Moran's I, whereas closed discs denote a decrease. The discs with large radii correspond to the most influential observations. We recover some of the previous results, namely that South Africa and Senegal are locally influential for positive contamination while Sudan and Egypt are not influential in that sense. Other countries that have not been selected on the hair-plot are associated with large open discs and could also be considered as influential observations.

In Figure 7, we plot the difference between the Moran's I estimator $\widehat{I}$ when deleting observations in turn and the original Moran index. The plot illustrates that the removal of Egypt or Sudan leads to a decrease of the Moran index. This result is similar to the one obtained by Anselin (1995) when looking at local Moran indicators while our proposal gives a different yet complementary insight into the dataset.

5. DISCUSSION

Many simple and natural estimators arising in the context of dependent data are unfortunately sensitive to perturbations of a single observation as we illustrated in this article. This motivates the need for developing robust estimators for time series and spatial data, and a few proposals can be found in the literature. Although classical estimators can often be defined as quadratic forms or ratios of quadratic forms in the data vector, robust estimators are typically more involved and therefore closed-form expressions for their local and asymptotic influence are not available. In that case, the hair-plot becomes very useful as it allows to visualize the sensitivity of each observation on complex estimators.
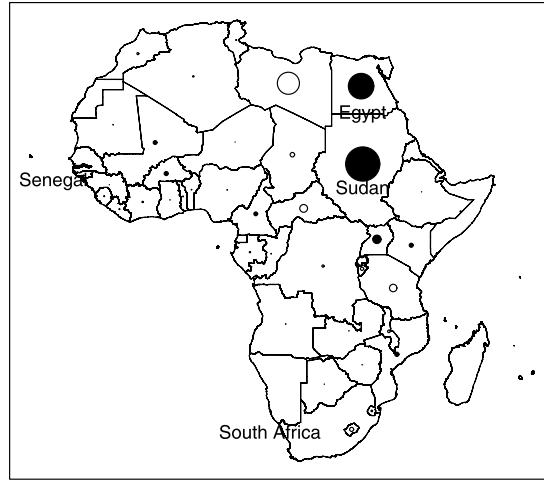
Moran Index Variation when Deleting Observations

Figure 7.   Differences between the Moran's I estimator $\widehat{I}$ when deleting observations in turn and the original index for the African conflict data.

For instance, Ma and Genton (2000) have proposed a highly robust autocorrelation estimator

$$\widehat{r}_{\text{HR}}(h) = \frac{Q_{n-h}^2(\mathbf{U}+\mathbf{V}) - Q_{n-h}^2(\mathbf{U}-\mathbf{V})}{Q_{n-h}^2(\mathbf{U}+\mathbf{V}) + Q_{n-h}^2(\mathbf{U}-\mathbf{V})}, \qquad 0 \le h \le n-1,$$

where Q_{n-h} is a highly robust scale estimator of a sample of size $n - h$ proposed by Rousseeuw and Croux (1993), and $\mathbf{U}$ and $\mathbf{V}$ represent the first and last $n - h$ observations of the data vector $\mathbf{Z}$, respectively. We apply this highly robust autocorrelation estimator to the Austrian bank monthly interest rate data and present the associated hair-plot for $\widehat{r}_{\text{HR}}(1) = 0.966$ in Figure 8. This value is in line with the optimal robust estimate of 0.96 obtained by Künsch (1984), with the value of 0.98 obtained by Azzalini and Genton (2008) based on a skew-t distributional assumption, and with the range of values 0.89–0.96 in the work of Wang et al. (2009). In Figure 8, the patterns of the curves are mainly due to the discreteness of the sample quantiles involved in the computation of Q_{n-h}. The vertical variability is small and the observation for the month 18 is no longer influential on this highly robust autocorrelation estimator.

In this article we have focused on the influential effect of a single observation at a time for computational and visual simplicity. However, the influence of multiple observations, $k > 1$ say, could be studied as well, each with possibly different contamination magnitudes $\zeta_1, \ldots, \zeta_k$. Clearly, this would lead to an explosion of the number of hairs (n choose k) and to dimensional difficulties in their graphical representation. Nevertheless, a particular case of interest is when those k magnitudes are all equal, to ζ say. Then, for an estimator defined as a ratio of two quadratic forms in the data vector as in Section 3, the asymptotic
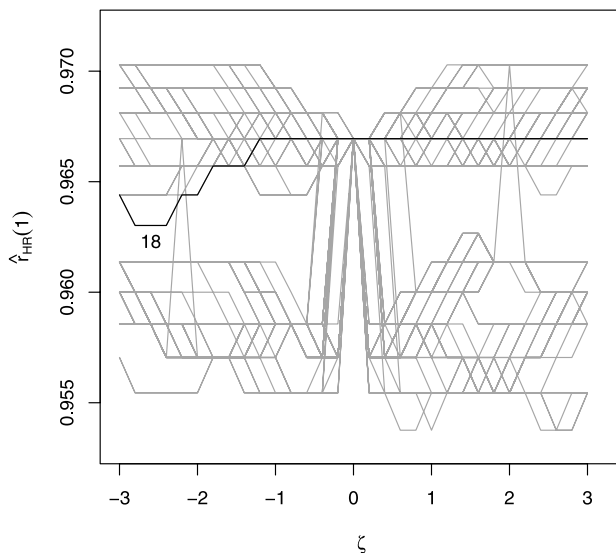
Figure 8. Hair-plot of the highly robust lag-one autocorrelation estimator $\widehat{r}_{\mathrm{HR}}(1)$ on the Austrian bank monthly interest rates dataset.

influence of k observations with indices in the set $\mathcal{I}$ is

$$v_{\mathcal{I}}(\widehat{\theta}_{\mathrm{RQ}}, \mathbf{Z}) = \frac{\sum_{i,j \in \mathcal{I}} a_{ij}}{\sum_{i,j \in \mathcal{I}} b_{ij}}. \tag{5.1}$$

Clearly, $v_{\mathcal{I}}$ depends on the set $\mathcal{I}$ and can be different from the simpler case given by $k = 1$ that we studied earlier.

For illustration, we return to the time series of monthly interest rates of an Austrian bank. We had noticed three possible outliers at months 18, 28, and 29 of about the same magnitude. The influence of those $k = 3$ observations on the lag-one sample autocorrelation $\widehat{r}(1)$ as a function of the common contamination $\zeta \in [-40, 40]$ is represented by the dashed curve in Figure 9. Notice that the dashed curve has a maximum for ζ around -2, which corresponds to bringing the three outlying observations down to the bulk of the data; see Figure 1. In that case the value of $\widehat{r}(1)$ increases, as expected from the highly robust autocorrelation estimator above. From (5.1), the corresponding asymptotic influence for $\mathcal{I} = \{18, 28, 29\}$ is

$$v_{\mathcal{I}}(\widehat{r}(1), \mathbf{Z}) = \frac{n^2 - 9n - 9}{3n(n-3)},$$

and takes the value 0.310 for the sample size $n = 91$. For comparison, the influence of the sole observation at month 18 on the lag-one sample autocorrelation $\widehat{r}(1)$ is represented by the solid curve in Figure 9. From (3.2), its corresponding asymptotic influence is

$$v_{18}(\widehat{r}(1), \mathbf{Z}) = \frac{-n - 1}{n(n-1)},$$

and takes the value -0.011 for the sample size $n = 91$. Hence, those two asymptotic influences are different, as can be seen in Figure 9.
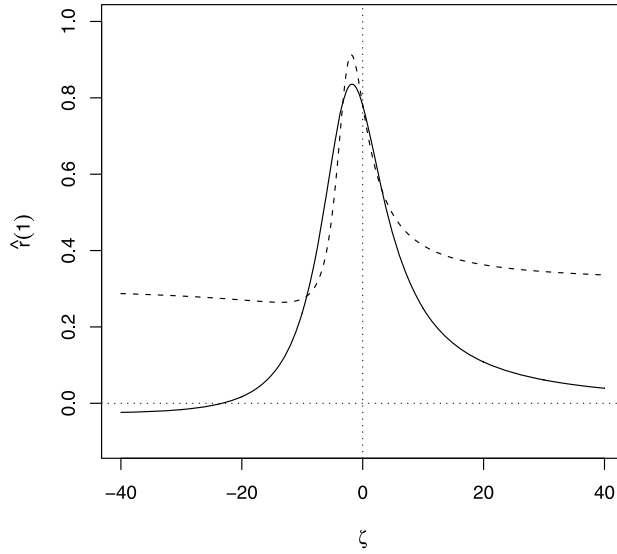
Figure 9.   Hair-plot based on various contaminated observations of the lag-one sample autocorrelation $\widehat{r}(1)$ on the monthly interest rates dataset: month 18 contaminated (solid curve) and months 18, 28, 29 contaminated (dashed curve).

It is also possible, but beyond the scope of the present article, to adapt our methodology to the case of analyzing the influence of observations on parameter estimators for general time series or spatial regression models. In particular, for spatial econometrics models, the local influence measure we propose is linked to the impact measures proposed by LeSage and Pace (2009) and it may be interesting to develop graphical tools in order to facilitate the analysis of such measures.

APPENDIX: PROOF OF PROPOSITION 1

By assumption:

$$f(y_1, y_2 + \zeta, y_3, \ldots, y_n) = f(y_1 + \zeta, y_2, y_3, \ldots, y_n)$$

for any $y_1, \ldots, y_n$. Consequently, for $y_1 = z_1$, $y_2 = 0$, $y_i = z_i$, $i = 3, \ldots, n$, and $\zeta = z_2$, we have

$$f(z_1, z_2, z_3, \ldots, z_n) = f(z_1 + z_2, 0, z_3, \ldots, z_n),$$

and by recurrence,

$$f(z_1, z_2, z_3, \ldots, z_n) = f\left(\sum_{i=1}^{n} z_i, 0, 0, \ldots, 0\right) = g\left(\sum_{i=1}^{n} z_i\right),$$

a function g of $\sum_{i=1}^{n} z_i$ only. This concludes the proof.

SUPPLEMENTAL MATERIALS

R-code for hair-plot: R-code for the command `hair.plot` described in the article (Hair-plot.R) and help file (hair-plot-description.pdf). (R-code.zip)

Data: Monthly interest rates of an Austrian bank dataset described in the article (bank.dat). The pollution data are in the book by Haining (1990) and the Africa data are in the R package *spdep*.

ACKNOWLEDGMENTS

REFERENCES

Anselin, L. (1995), "Local Indicators of Spatial Association—LISA," *Geographical Analysis*, 27, 93–115. [817, 820]

Azzalini, A., and Genton, M. G. (2008), "Robust Likelihood Methods Based on the Skew-*t* and Related Distributions," *International Statistical Review*, 76, 106–129. [810,821]

Cliff, A. D., and Ord, J. K. (1981), *Spatial Processes: Models and Applications*, London: Pion. [815]

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18. [808]

Cook, R. D., and Weisberg, S. (1980), "Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495–508. [808]

——— (1994), *An Introduction to Regression Graphics*, New York: Wiley. [809]

Genton, M. G. (1998a), "Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure," *Mathematical Geology*, 30, 323–345. [813]

——— (1998b), "Spatial Breakdown Point of Variogram Estimators," *Mathematical Geology*, 30, 853–871. [816]

——— (1999), "The Correlation Structure of the Sample Autocovariance Function for a Particular Class of Time Series With Elliptically Contoured Distribution," *Statistics and Probability Letters*, 41, 131–137. [813]

——— (2003), "Breakdown-Point for Spatially and Temporally Correlated Observations," in *Developments in Robust Statistics, International Conference on Robust Statistics 2001*, eds. R. Dutter, P. Filzmoser, U. Gather, and P. J. Rousseeuw, Heidelberg: Springer, pp. 148–159. [813]

Genton, M. G., and Lucas, A. (2003), "Comprehensive Definitions of Breakdown-Point for Independent and Dependent Observations," *Journal of the Royal Statistical Society, Ser. B*, 65, 81–94. [813,815,816]

——— (2005), Comment on "Breakdown and Groups," by L. Davies and U. Gather, *The Annals of Statistics*, 33, 988–993. [813,815,816]

Genton, M. G., and Ronchetti, E. (2003), "Robust Indirect Inference," *Journal of the American Statistical Association*, 98, 67–76. [816]

Gorsich, D. J., Genton, M. G., and Strang, G. (2002), "Eigenstructures of Spatial Design Matrices," *Journal of Multivariate Analysis*, 80, 138–165. [813]

Haining, R. (1987), "Trend-Surface Models With Regional and Local Scales of Variation With an Application to Aerial Survey Data," *Technometrics*, 29, 461–469. [816]

———— (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge, U.K.: Cambridge University Press. [816,824]

———— (1994), "Diagnostics for Regression Modelling in Spatial Econometrics," *Journal of Regional Science*, 34, 325–341. [809]

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics. The Approach Based on Influence Functions*, New York: Wiley. [809,810]

Hillier, G., and Martellosio, F. (2006), "Spatial Design Matrices and Associated Quadratic Form: Structure and Properties," *Journal of Multivariate Analysis*, 97, 1–18. [813]

Künsch, H. (1984), "Infinitesimal Robustness for Autoregressive Processes," *The Annals of Statistics*, 12, 843–863. [810,821]

Lark, R. M. (2008), "Some Results on the Spatial Breakdown Point of Robust Point Estimates of the Variogram," *Mathematical Geoscience*, 40, 729–751. [816]

LeSage, J., and Pace, R. K. (2009), *Introduction to Spatial Econometrics*, Boca Raton: Chapman & Hall/CRC. [823]

Li, H., Calder, C., and Cressie, N. A. C. (2007), "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model," *Geographical Analysis*, 39, 357–375. [815]

Ma, Y., and Genton, M. G. (2000), "Highly Robust Estimation of the Autocovariance Function," *Journal of Time Series Analysis*, 21, 663–684. [810,821]

Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, New York: Wiley. [809]

Martin, R. J. (1992), "Leverage, Influence and Residuals in Regression Models When Observations Are Correlated," *Communications in Statistics: Theory and Methods*, 21, 1183–1212. [809]

Moran, P. A. P. (1950), "Notes on Continuous Stochastic Phenomena," *Biometrika*, 37, 17–23. [814]

R Development Core Team (2010), "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria. Available at *http://www.R-project.org*. [810]

Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283. [821]

Tiefelsdorf, M. (2000), *Modelling Spatial Processes—The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I. Lecture Notes in Earth Sciences*, Vol. 87, Berlin: Springer Verlag. [815]

Wang, Q., Stefanski, L. A., Genton, M. G., and Boos, D. (2009), "Robust Time Series Analysis via Measurement Error Modeling," *Statistica Sinica*, 19, 1263–1280. [811,821]

Welsh, R. E. (1982), "Influence Functions and Regression Diagnostics," in *Modern Data Analysis*, eds. R. L. Launer and A. F. Siegel, New York: Academic Press, pp. 149–169. [809]