Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2012-61

Exploratory Designs for Assessing Spatial Dependence

Agnes Fussl, Werner G. Müller and Juan Rodríguez-Díaz[a]

February 2012

[a]University of Salamanca

Exploratory Designs for Assessing Spatial Dependence

1.1 Introduction

Efficient data acquisition as introduced in the first chapter requires some prior understanding of the process to be observed, ideally in the form of a spatio-temporal model or at least a narrow enough class of such models. If this is not the case as it is frequently at the beginning of a study, the employed design must guarantee that an appropriate model can be identified as data collection progresses. These designs, usually called exploratory designs (cf. chapter 4 in Müller 2007), typically employ minimum assumptions which make random sampling a reasonable choice.

However, random sampling can be awfully inefficient even for simple tasks that arise in the beginning of a study. One of the simplest and most initial (albeit quite important) tasks in a spatio-temporal context is the assessment of whether spatial or temporal dependence is present or not and if yes of which intensity (and form). In this chapter we would like to adress these questions and possible improvements over random sampling in coping with them. As the treatment of the temporal dimension usually involves straightforward regular observations or simple extensions from the spatial case, we will in the following mainly concentrate on the latter.

At a first stage one should then attempt to detect whether there is any spatial dependence in the data or not. Should they be spatially independent, the respective statistical design and analysis usually reduces to the application of a classical and well established toolbox. Thus, the decision of whether one can confine oneself to this well understood body of knowledge or whether one has to resort to the rather freshly developed methodologies of spatial design is a crucial element of any serious spatial investigation. Spending some efforts in efficient designs for testing for spatial dependence could save considerable overall efforts. If, for instance, one detects at an early stage of an investigation that effectively no spatial correlation is present, one can return to the classical, rather simple to construct optimal designs treated e.g. in chapter 3 of Müller (2007). Therefore, in the next sections we will concentrate on the question of how to optimally select coordinates / predictor values to detect the spatial structure, if it is existent, and how to avoid to spuriously detect spatial dependence if there is no such structure. Later we will also consider to evaluate the specific form

of this dependence.

In particular, this chapter deals with statistical modeling of areal data which are observed on polygon entities with defined boundaries. Typical examples for such areal spatial objects are areal entities like states or counties. The aim of the chapter is to give a short overview of how to collect and analyze a data set containing information on areal spatial objects with regard to the following questions:

- How can we define spatial neighbors?
- Which spatial weights should be assigned to the identified neighbor links?
- Are there any spatial dependencies in the data?
- Which statistical models are adequate to the data?
- How are the spatial modeling approaches related to each other, what are the differences between them and what are the consequences on design?

To answer these questions an exemplary spatial data analysis is performed on a data set concerning the grassland usage in Upper Austria (see section 1.2). The data are analyzed using the statistical software R which provides a wide range of packages and functions to work on spatial data; the used R packages and the R code are given in appendix 1.8. For an extensive introduction to spatial data analysis in R see Bivand et al. (2008). We will concentrate our exposition on the lattice type of data for two reasons: the continuous regions can be covered by an arbitrarily fine grid and the continous random field models can be well approximated by lattice based ones such as Gaussian Markov random fields (cf. Rue and Held 2005).

1.2 The data set and its visualization

The data set contains information on the greenland usage in the 445 municipalities of Upper Austria over the years 1995, 1999, 2003, 2005, 2007 and 2008. A municipality's greenland usage is measured by

$$log(\text{area of arable land} + 1) - log(\text{area of grassland} + 1),$$

i.e. the log ratio of these two areas, making it scale free (variables *R95* to *R08*). The value of this log ratio is positive if the area of arable land is larger than the area of grassland and negative if the proportion of arable land compared to the grassland is smaller. Other variables provided in the data set are: *LBBGG* (identification number of municipality), *BEZNR* (identification number of the municipality's district), *Longitude* and *Latitude* (local coordinates of municipality), *FLKM2* (area of the municipality in square kilometers) and *Altitude* (height above sea level in meters). For graphical display of the data a shape file of the boundaries of the municipalities is available.

Before a spatial analysis can be started, both data and shape file must be imported in R and combined to a `SpatialPolygonsDataFrame` object. To make a first check on spatial correlation for the important variables in the data set it is reasonable

to display them in a so-called Choropleth map (Waller and Gotway 2004). For color filling of the maps the R package `RColorBrewer` is used which is based on the web tool *ColorBrewer* (for the latest version see Brewer and Harrower 2010).

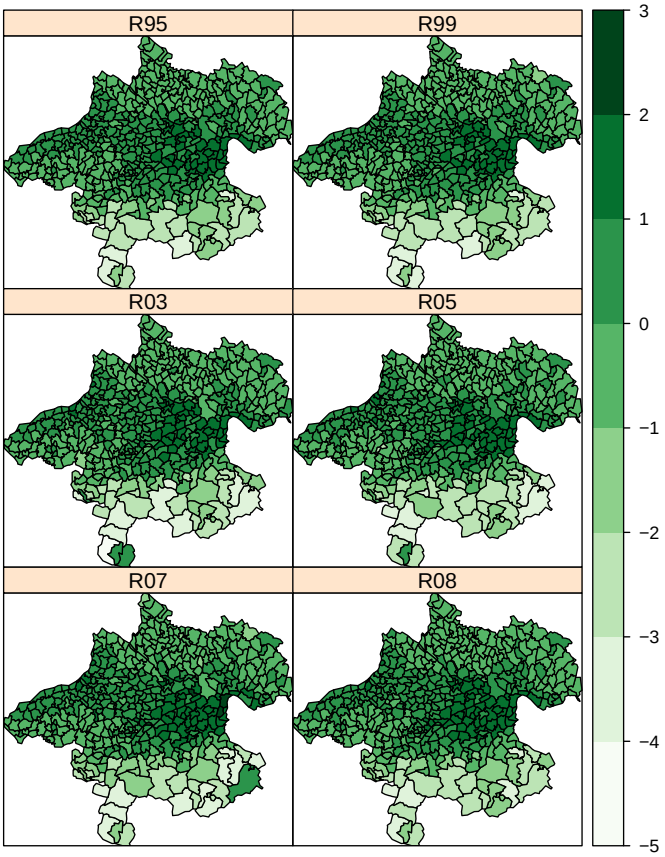

Figure 1.1: Maps of the log ratios *R95* to *R08*

One way to visualize the spatio-temporal development of the grassland usage in Upper Austria is to create a slide show consisting of the series of maps of the log ratios *R95* to *R08*. Alternatively, the function `spplot` can be used to show all the maps at once (Fig. 1.1). According to the maps it is evident to assume some spatial dependencies in the data. Nearly all municipalities in the south of Upper Austria seem to have more grassland than arable land, whereas municipalities in the central east of Upper Austria and along the river Inn tend to have a higher percentage of

arable land in relation to grassland. Examining the development of the log ratios, nearly no change can be identified over the years. Just for a few municipalities the sign of the log ratio switches between the years 1995 and 2008.

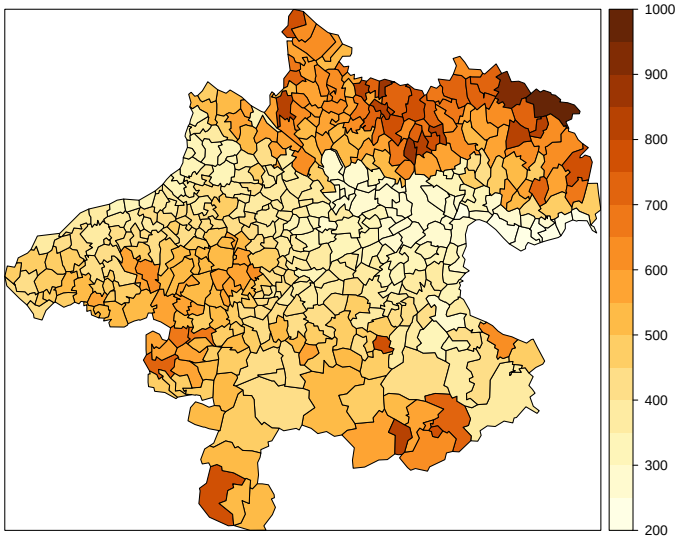

Figure 1.2: Map of the variable *Altitude*

However, it seems to be obvious that the altitude of the municipalities influences the log ratios and therefore also the observable structure in Fig. 1.1. A look at the map of the altitudes (Fig. 1.2) shows that the higher elevated regions are in the northeast and south of Upper Austria, whereas the less elevated and flatter areas can be found in the center of Upper Austria and along the two largest rivers Danube and Inn. These are also the regions where a concentration of more arable land in relation to grassland is present. Possible reasons for this dependency are that it might be easier to cultivate on lowlands than on the slopes of the hillier municipalities and that there is a higher availability of water resources along the two rivers, as well. Thus, the question remains if there are spatial dependencies in the data anyway or if they are only resulting from the different elevations of the municipalities above sea level.

1.3 Spatial links

The sampling design primarily affects the so called spatial link matrices (or spatial weighting matrices), which represent the spatial relationships between observations and are frequently employed in spatial econometrics (for a characterization of this

branch of statistics see Anselin 1988 or more recently Arbia 2006). In general, spatial link matrices represent similarities, e.g. connectivity, neighbourhoods or inverse distances. A spatial link matrix $\mathbf{G}$ is an $n \times n$ matrix (n is the number of observations) with the following properties:

(i) $g_{ij} = 0$ for $i = j$;

(ii) $g_{ij} > 0$ if i and j are spatially connected.

Thus the key concept to analyze and model areal data is to define which sites in the data set are connected and therefore so-called spatial neighbors. Subsequently, spatial weights may be assigned to each of the identified neighbor links. Both steps are essential issues for statistical modeling of areal data because the results of the spatial analysis are crucially dependent on the decisions made in constructing the spatial weights. Hence, the following sections 1.3.1 and 1.3.2 give a short overview of the different approaches to define spatial neighbors and spatial weights. In literature these topics are dealt with e.g. by Waller and Gotway (2004, pp. 223-225), Fortin and Dale (2005, pp. 113-118), O'Sullivan and Unwin (2003), Schabenberger and Gotway (2005, pp. 18-19) and Banerjee et al. (2004, pp. 70-71). Due to the practical relevance for programming in R, the two sections are structured following Bivand et al. (2008).

1.3.1 Spatial neighbors

Neighbor relationships between all objects are usually represented by a $n \times n$ binary connectivity matrix $\mathbf{C}$, where n is the number of observations (Fortin and Dale 2005, pp. 113-118). The component c_{ij} of the connectivity matrix is defined as follows:

$$c_{ij} = \begin{cases} 1, & \text{if there exists a neighbor relationship between two objects } i \text{ and } j \\ 0, & \text{if two objects } i \text{ and } j \text{ are not in a neighbor relationship.} \end{cases}$$

Each component c_{ii} is set to zero since no region is a neighbor to itself. Generally, connectivity matrices can be symmetric or asymmetric, where asymmetry is present when i is a neighbor of j but j is not a neighbor of i or vice versa.
When working with the package `spdep` in R (Bivand et al. 2010), neighbor relationships are represented by a `nb` object (see Bivand et al. 2008, pp. 240). This object consists of a list of length n, where for each observation i an integer vector of index numbers of its neighbors is recorded. To objects without any neighbors an integer zero is assigned. Additionally, the `nb` object gives information about the symmetry, presence of no-neighbor observations, average number of links, link number distribution and least/most connected regions.

To complete the definition of a connectivity matrix the term neighborhood has to be specified more precisely: One of the most commonly used approaches in literature to determine neighbor relationships is to create contiguity neighbors (R function `poly2nb`). Two polygon areas i and j are contiguity neighbors if they share

- at least one point on their boundaries (=Queen-style) or

- at least two points on their boundaries (=Rook-style).

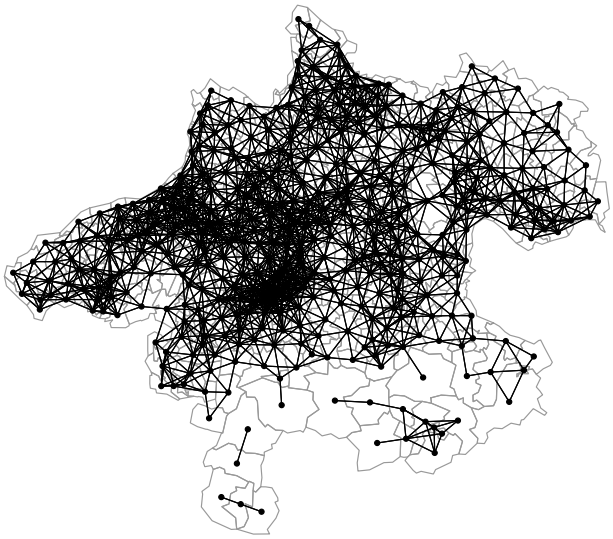

Figure 1.3: Distance-based neighbors within a radius of 10.076 m

To use other neighborhood criteria it is necessary to choose a point to represent each polygon entity, which is often the polygon centroid. Once representative points are available, neighbors can be determined e.g. by means of graph measures like Delauney triangulation neighbors (R function `tri2nb`) or Gabriel graph neighbors (R function `graph2nb`). Another method to create neighbors is to choose the k nearest points as neighbors for each polygon (R function `knearneigh`). In many cases this method leads to an asymmetric connectivity matrix (Banerjee et al. 2004, p. 70), which can be made symmetrical in R using the function `make.sym.nb`. Alternatively, distance-based neighbors can be established by connecting points within an interpoint distance with fixed lower and upper distance bounds (R function `dnearneigh`).

For our example this last method is used as basis for the connectivity matrix. The lower and upper distance bounds are set to 0 and 10.076 m, which is the minimum distance at which all areas have a distance-based neighbour. Thus, it can be guaranteed that all areas in the data set are linked to at least one neighbor (see Fig. 1.3). For more details on creating spatial neighbors in R see Bivand et al. (2008, pp. 242-251).

1.3.2 Spatial weights

After establishing the connectivity matrix $\mathbf{C}$, spatial weights may be assigned to each neighbor relationship. For this purpose a spatial weights matrix $\mathbf{W}$ can be computed. The idea of spatial weights is to assign higher weights w_{ij} to connected areas i and j if area j is (in some sense) closer to area i than other connected areas to i. The definition of proximity may for example be based on:

- distance between two areas
- length of the shared border of two areas or
- relative sizes of the areas.

We could assume, for example, that the strength of neighbor relationships decreases with distance. Therefore, we use the inverse distance between two entities ($1/d_{ij}$) as weight to determine the component w_{ij} of matrix $\mathbf{W}$. If area i is not a neighbor of area j (i.e. if $c_{ij} = 0$), w_{ij} is set to zero. If objects i and j are adjacent component w_{ij} in this example would then be defined as follows:

$$w_{ij} = \frac{c_{ij}}{d_{ij}}$$

Another variant is the exponential weighting

$$w_{ij} = e^{-\theta d_{ij}} - \delta_{ij}, \tag{1.1}$$

where θ is some decay parameter and δ_{ij} is the usual Kronecker δ. Note that this scheme corresponds to a spatial process with a specific version of Matérn variogram $\gamma_M(h, \theta)$.

However, if there is no reason to assume more than the existence and absence of neighbor relationships, a spatial weights matrix $\mathbf{W}$ deviating noticeably from the binary connectivity matrix $\mathbf{C}$ should be avoided. In this case the simplest way to define the spatial weights matrix $\mathbf{W}$ would be to set $\mathbf{W}$ equal to the connectivity matrix $\mathbf{C}$ (i.e. $w_{ij} = c_{ij}$).

Spatial weight matrices are often converted by using coding schemes to cope with the heterogeneity which is induced by the different linkage degrees of the spatial object. A widely used coding scheme is the row-sum standardized coding scheme where the sum of each row of the standardized link matrix is equal to one (Waller and Gotway 2004, p. 225; O'Sullivan and Unwin 2003, p. 42). We simply obtain the components of this row standardized matrix $\mathbf{W}_{std}$ by dividing each w_{ij} by the sum of the neighbor weights for region i:

$$w_{std,ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{1.2}$$

This row standardization is used to take the different numbers of neighbors per unit into account. Other coding-schemes are for instance the globally standardized, and the variance stabilizing coding scheme (Tiefelsdorf 2000). For reasons of

simplification, the spatial link matrix $\mathbf{W}$ is in the following always the row-standardized version.

In R we use the function `nb2listw` to convert a `nb` object into a spatial weights object `listw`. The argument `glist` of this function can be used to pass a list of vectors of weights corresponding to the neighbor relationships. Additionally, there are various different weight styles available to standardize the matrix. Conversion style `W` is the default value and creates a row standardized weights matrix. Style `B` retains a weight of unity for each link, in style `C` the complete set of weights sums to the number of observed entities n and in style `U` all weights together sum to 1 (see Bivand et al. 2008, pp. 251-255).

In our example we set the spatial weights matrix $\mathbf{W}$ equal to the binary connectivity matrix $\mathbf{C}$ and use the row standardized matrix $\mathbf{W}_{std}$ as in (1.2) for further statistical analysis and modeling of the data. Extensions to the spatio-temporal setting are straightforward and can e.g. be found in Dubé and Legros (2011).

1.4 Measures of Spatial Dependence

Several measures for quantifying spatial dependence have been proposed in literature. Generally, they can be classified into two groups: the first ones are based on weighted covariance type expressions analogous to the Durbin-Watson statistic for time series (prototypically the Moran's $\mathfrak{I}$, see Moran 1950). The others are based on weighted averages of squared differences - often called semivariances (prototypically the Geary's $\mathfrak{c}$, see Geary 1954). Most of the available software tools applicable for spatial analysis provide a standard implementation of those measures, see e.g. Rangel et al. (2010) or the many R-packages to be found on `http://cran.r-project.org/web/views/Spatial.html`. To get an overview concerning the relationships between those measures see Dale et al. (2002).

As a measure for the intensity of the spatial dependence and for detecting its potential existence the probably most popular statistic, the Moran's $\mathfrak{I}$, is used here. Therefore, to test for the presence of spatial dependence we employ in this chapter Moran's $\mathfrak{I}$ test as this is perhaps the most common global test for this kind of problem (see e.g. Waller and Gotway 2004, pp. 227; Schabenberger and Gotway 2005, pp. 21; Bivand et al. 2008, pp. 258). Moran's $\mathfrak{I}$ is calculated by dividing the product of the variable of interest and its spatial lag with the cross-product of the variable of interest and adjusting this term for the spatial weights:

$$\mathfrak{I} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i is the ith observation of the variable of interest, $\bar{y}$ is the mean of the variable of interest and w_{ij} is the (standardized) spatial weight of the neighbor relationship between two areas i and j. Moran's $\mathfrak{I}$ will be positive, if neighboring areas tend to have similar values of the variable of interest, and negative, when they tend to have different values. For the testing procedure the observed value is standardized by

subtracting its expected value and dividing this difference by the standard deviation under the null hypothesis of no spatial dependence. Usually, the test is one-sided, testing whether the observed statistic is significantly greater than its expected value. For details concerning the testing procedure see e.g. Schabenberger and Gotway (2005, pp. 21). A different motivation and interpretation of $\mathfrak{I}$ is provided in Dray (2011).

In R the function `moran.test` implemented in the package `spdep` is used to perform the test for spatial autocorrelation. Examining the log ratio of the year 2008 *R08* in our example yields the following result:

```
        Moran's I test under normality

data:  data$R08
weights: ooe_W1

Moran I statistic standard deviate = 36.231, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.795973350      -0.002252252       0.000485391
```

The test results seem to show a significant positive spatial autocorrelation in *R08*, but the interpretation has to be done in a more careful manner:

First, we have to consider the test assumptions of constant mean and variance of the y_i and should be aware of spurious autocorrelation or 'misspecification' (Schabenberger and Gotway 2005, pp. 22). Any autocorrelation in our data may e.g. simply be due to the altitudes of the municipalities and not to any spatial pattern in the log ratio of arable land vs. grassland. To overcome this issue we could for example fit a mean model to the data including additional variables (see section 1.5) and performing the test for spatial autocorrelation again for the residuals of this model using the function `lm.morantest`.

Second, we should keep in mind that the test outcome is also affected by the choice of the spatial weights and the standardizing scheme used for the weights (see the examples in Bivand et al. 2008, pp. 262).

In Fig. 1.4 two approaches to plot the spatial autocorrelation are shown (see Bivand et al. 2008, p. 267). We can find the values of Moran's $\mathfrak{I}$ for sixteen successive lag orders of contiguous neighbors on the left side (function `sp.correlogram`) and for a sequence of distance band neighbors on the right side (function `correlog` in package `pgirmess`). In our example, the first four bands of 0-10 km, 10-20 km, 20-30 km and 30-40 km have significant values of spatial autocorrelation (on the other hand the following significantly negative values may indicate some kind of nonstationarity). Another graphical tool to examine the data for spatial autocorrelation is the Moran scatterplot in Fig. 1.5. It shows the relationship between the variable of interest (x-axis) and the spatially weighted average of neighboring values, also called the spatially lagged values (y-axis). Most of the points in our example appear in the low-low and high-high quadrants representing a positive spatial dependency. Only a few locations can be found in

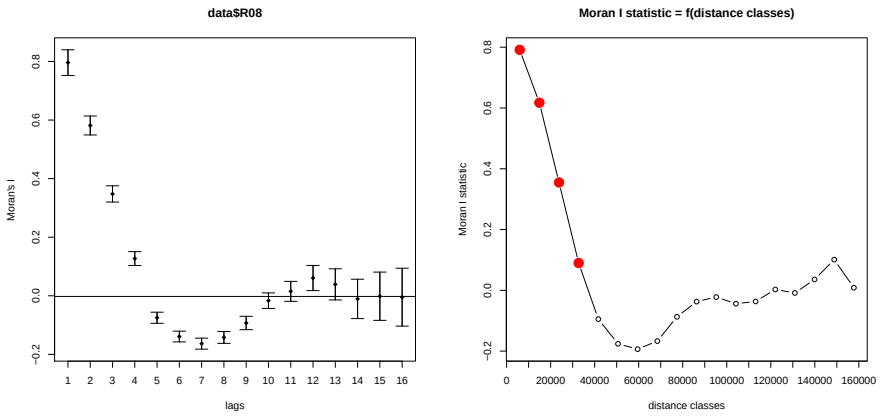

Figure 1.4: Correlograms: values of Moran's $\mathfrak{I}$ for sixteen successive lag orders (*left*); values of Moran's $\mathfrak{I}$ for a sequence of distance band neighbours (*right*)

the low-high and high-low quadrants refering to locations surrounded by dissimilar valued neighbors. The slope of the regression line corresponds to Moran's $\mathfrak{I}$ and represents the linear association between the observed values and the spatially lagged values (see Anselin 1993, Anselin 1995). In R the function `moran.plot` is used to visualize the data in a Moran scatterplot (Bivand et al. 2008, pp. 268).

As already noted several alternative measures to $\mathfrak{I}$ exist that could equivalently be employed for our purposes, most notably the so called contiguity ratio proposed by Geary (1954). Recently, López et al. (2011) provide extensive simulations on four candidate measures including $\mathfrak{I}$ and demonstrate its comparative value in an economic application.

1.5 Models for areal data

For simplification of exposition, we will in the rest of the section again assume that the considered models are linear and the error processes being Gaussian, with the obvious extensions to locally linearized models as in previous chapters. The regression residuals from estimation of the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ under the assumption $\boldsymbol{\varepsilon}$ i.i.d. will be used for the test of spatial dependence. The real data generating process, the true but unknown status of the world, is assumed to be one of the following:

H_0: spaceless;

H_A: spatial;

The distinction between these hypotheses will be made clear in the following subsections. Depending on the two examined cases, one either wants to reject or

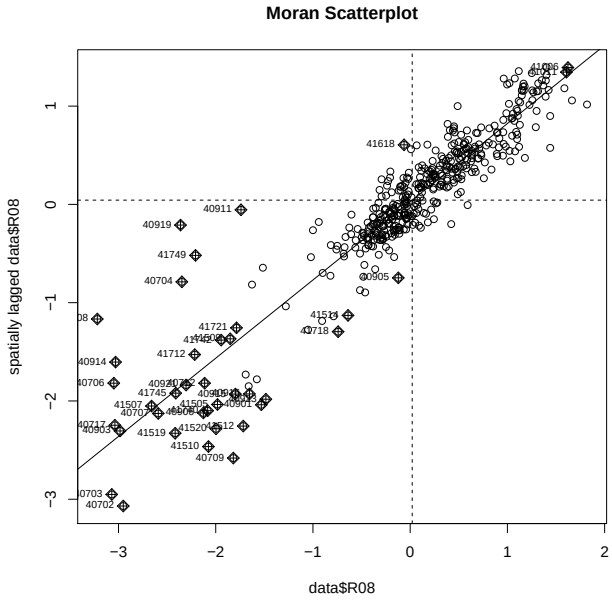

Figure 1.5: Moran Scatterplot

not to reject the null hypothesis of spatial independence of Moran's $\mathfrak{I}$ test to make a correct decision. For more on this issue see e.g. Anselin (1988). Further, one wants to find an optimal or nearly optimal design for a test strategy to receive either nonrejection or rejection of the null hypothesis for derivation of a model that matches the real status of the world.

1.5.1 H_0: a spaceless regression model

We start with the basic linear regression model, which reads:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.3}$$

where $\mathbf{y}$ is the outcome variable of interest, $\mathbf{X}$ a matrix of explanatory variables, $\boldsymbol{\beta}$ the vector of parameters and $\boldsymbol{\varepsilon}$ an error term with errors assumed to be independently distributed.

For a standard regression model it is crucial to know whether the residuals are spatially dependent or not. If there is no spatial dependence in the residuals, one can use standard estimation methods, like OLS, but if the residuals show spatial dependence, one has to use special methods (cf. section 1.5.2). When the OLS estimation method is applied instead, spatial autocorrelation in the error term leads to biased estimates of the residual variance and inefficient estimates of the regression

coefficients. For regression residuals Moran's $\mathfrak{I}$ is defined as scale invariant ratio of quadratic forms in the normally distributed regression residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)'$, i.e.

$$\mathfrak{I} = \frac{\hat{\boldsymbol{\varepsilon}}' \frac{1}{2}(\mathbf{W} + \mathbf{W}')\hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}} \tag{1.4}$$

where $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = n$, see e.g. Tiefelsdorf (2000).

The classical Moran's $\mathfrak{I}$ as the two-dimensional analog of a test for univariate time series correlation is given in e.g. Cliff and Ord (1981). For a random variable Y, measured in each of the n non-overlapping subareas of the whole study area, Moran's $\mathfrak{I}$ is defined from the residuals of an intercept only regression, i.e. $\hat{\boldsymbol{\varepsilon}} = \mathbf{My}$ where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{1}_n$ is an $n \times 1$ vector of ones. In this case, and if the spatial link matrix $\mathbf{W}$ has full rank (i.e. there is no observation completely separated from all others), the expected value of the test statistic $\mathfrak{I}$ under independence is given by $E[\mathfrak{I}|H_0] = -\frac{1}{n-1}$, and the variance of $\mathfrak{I}$ can be given in terms of the eigenvalues γ_i of matrix $\mathbf{K} = \mathbf{M}'\frac{1}{2}(\mathbf{W} + \mathbf{W}')\mathbf{M}$ as $\text{Var}[\mathfrak{I}|H_0] = \frac{2n}{n^2-1}\sum_{i=1}^{n}(\gamma_i - \bar{\gamma})^2 = \frac{2n}{n^2-1}\sigma_\gamma^2$. The test statistic $\mathfrak{I}$ is then asymptotically normally distributed.

The Moran's $\mathfrak{I}$ test is used for parametric hypotheses about the spatial autocorrelation level ρ, i.e. $H_0 : \rho = 0$ against $H_A : \rho > 0$ for positive spatial autocorrelation; or $H_0 : \rho = 0$ against $H_A : \rho < 0$ for negative spatial autocorrelation. Tests for positive correlation are much more relevant in practice, because negative spatial autocorrelation very rarely appears in the real world. Thus, from now on $\rho \geq 0$ will be assumed. The z-transformed Moran's $\mathfrak{I}$ is, for normally distributed regression residuals and well-behaved spatial link matrices under certain regularity conditions (see e.g. Tiefelsdorf 2000), asymptotically standard normally distributed, i.e. $z(\mathfrak{I})$ is defined as

$$z(\mathfrak{I}) = \frac{\mathfrak{I} - E[\mathfrak{I}|H_0]}{\sqrt{\text{Var}[\mathfrak{I}|H_0]}} \sim N(0, 1). \tag{1.5}$$

The exact small sample distribution of Moran's $\mathfrak{I}$ was seemingly independently obtained by Hepple (1998) and Tiefelsdorf and Boots (1995), but does not offer advantages with respect to the design task as shown in Müller et al. (2012). The behaviour under deviations from normality is investigated in Griffith (2010).

Note that it turns out that a special class of spatial objects is relevant especially for design purposes. These are observations that belong to a design but are far apart from all other objects, in the sense that they have no spatial links to other observations. They have been termed far-off objects by Gumprecht (2007) and a discussion of their role in Moran's $\mathfrak{I}$ tests and corresponding designs are given therein.

Conveniently, we also start in our example with estimating the basic linear regression model to see which covariates contribute for explaining the variance in the response variable *R08*. As the spatial autocorrelation that we detected with the tests in section 1.4 could actually be caused by model misspecification we also check if there still remains spatial autocorrelation in the residuals of the model. For this purpose a map of the residuals like in section 1.2 can be plotted and a Moran's $\mathfrak{I}$ test like in section 1.4 is performed for the residuals (in R: function `lm.morantest`).

In a first attempt we include all explanatory variables in the linear regression model except the variable *FLKM2* (area of the municipality in square kilometers), which wouldn't make any sense in explaining the dependent variable *R08*. We call the R function `lm` and obtain the following result:

```
Call:
lm(formula = data$R08 ~ data$HEIGHT + data$R07 + data$R05 +
    data$R03 + data$R99 + data$R95)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75505 -0.04410 -0.00830  0.04535  0.68314

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.602e-01  2.196e-02   7.292 1.44e-12 ***
data$HEIGHT -1.931e-04  4.556e-05  -4.238 2.75e-05 ***
data$R07     8.595e-02  1.463e-02   5.874 8.40e-09 ***
data$R05     4.727e-02  1.839e-02   2.570   0.0105 *
data$R03     8.107e-02  1.744e-02   4.650 4.41e-06 ***
data$R99     4.237e-01  5.116e-02   8.281 1.50e-15 ***
data$R95     3.669e-01  4.825e-02   7.604 1.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1135 on 438 degrees of freedom
Multiple R-squared: 0.9834, Adjusted R-squared: 0.9832
F-statistic:  4324 on 6 and 438 DF,  p-value: < 2.2e-16
```

All covariates seem to have a significant influence on the dependent variable and the Moran's $\mathfrak{I}$ test decides in favor of the hypothesis that there is no more spatial autocorrelation in the residuals (in fact it decides for negative spatial autocorrelation, which indicates some kind of overcorrection):

```
        Global Moran's I for regression residuals

data:
model: lm(formula = data$R08 ~ data$HEIGHT + data$R07 + data$R05 +
data$R03 + data$R99 + data$R95)
weights: ooe_W1


    Moran I statistic standard deviate = -6.4293, p-value = 1
    alternative hypothesis: greater
    sample estimates:
    Observed Moran's I        Expectation           Variance
         -0.1445027577      -0.0061184919       0.0004632761
```

However, if we analyze the explanatory variables in detail, we can find that the ratios *R07* to *R95* are highly correlated (i.e. multicollinearity). According to Fahrmeir et al. (2009, pp. 170), a multicollinearity problem can be identified by computing the variance inflation factor (VIF) for each explanatory variable and a multicollinearity problem is present for VIF > 10. This case arises for two of our

covariates, i.e. *R95* (VIF=56.82) and *R99* (VIF=62.75). Therefore, we drop these two explanatory variables *R95* and *R99* in a second attempt and obtain the following output:

```
Call:
lm(formula = data$R08 ~ data$HEIGHT + data$R07 + data$R05
    + data$R03)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21197 -0.06322  0.00175  0.06983  1.29267

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.774e-01  4.060e-02   4.370 1.55e-05 ***
data$HEIGHT -2.892e-04  8.417e-05  -3.436 0.000646 ***
data$R07     3.292e-01  2.345e-02  14.036  < 2e-16 ***
data$R05     2.980e-01  3.050e-02   9.770  < 2e-16 ***
data$R03     2.633e-01  3.060e-02   8.604  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2109 on 440 degrees of freedom
Multiple R-squared: 0.9424,     Adjusted R-squared: 0.9419
F-statistic:  1800 on 4 and 440 DF,  p-value: < 2.2e-16
```

The regression coefficients of the remaining covariates are all significant again, but the Moran's $\mathfrak{I}$ test for the residuals yields a significant result now, an effect which also occurs for other choices of the spatial weights matrix $\mathbf{W}$:

```
        Global Moran's I for regression residuals

data:
model: lm(formula = data$R08 ~ data$HEIGHT + data$R07 + data$R05 +
data$R03)
weights: ooe_W1

Moran I statistic standard deviate = 3.7525, p-value = 8.752e-05
alternative hypothesis: greater
sample estimates:
Observed Moran's I        Expectation           Variance
      0.0755249376      -0.0057114017       0.0004686499
```

Usually in this case, we fit a spatial regression model instead of the linear regression model to estimate the spatial effect as well. However, we should deal with another possible problem first of all: If the dependent variable $\mathbf{y}$ and one or more of the explanatory variables $\mathbf{X}$ are generated according to a spatial autoregressive process with a positive autoregression parameter ρ (for details see section 1.5.2) and $\mathbf{y}$ is regressed on $\mathbf{X}$, there is a risk of spurious spatial regression (see e.g. Fingleton 1999, Lauridsen and Kosfeld 2006, Beenstock and Felsenstein 2008, Beenstock and Felsenstein 2010). To check whether we have the case of spurious spatial regression, it is necessary to test for spatial nonstationarity. Lauridsen and Kosfeld (2006) and

Beenstock and Felsenstein (2008)/Beenstock and Felsenstein (2010) propose two contrary procedures to test for spatial nonstationarity and it doesn't seem obvious for us up to now which one is the 'correct' approach. Fact is, if we observe spatial nonstationarity, this involves the risk of spurious spatial regression. Moreover, one should also test for spatial cointegration in a second step. Spatial cointegration concerns the case where two or more variables in the regression are nonstationary, while the errors are stationary (see Lauridsen and Kosfeld 2006, p. 9). However, in case all variables are stationary, we can estimate a spatial regression model like in section 1.5.2.

1.5.2 H_A: spatial regression models

As seen in the previous section, we can obviously not assume that the observations (and model errors) are independent of each other when we work with geographically referenced data and suppose correlations between neighboring areas (Gibbons and Overman 2010; Bivand et al. 2008, pp. 273). Therefore in the literature one can find several approaches to include spatial dependencies in the regression equation (see e.g. Gibbons and Overman 2010, Kissling and Carl 2008). These models are called ***spatial simultaneous autoregressive models*** and differ from each other in the assumption where the spatial autoregressive process occurs. The following paragraphs present the ideas of the different spatial regression models and try to point out the connections between them.

In a first step the linear regression model (1.3) is extended by the term $\lambda\mathbf{W}\mathbf{u}$ so that we get the regression equation in (1.6). It is assumed that the errors are no longer independent, but involve the spatial autoregressive process. This model is denoted as the *simultaneous autoregressive (**SAR**)* model (Bivand et al. 2008, pp. 277) or – in the spatial econometrics context – the *spatial error (**SE**)* model (Gibbons and Overman 2010, p. 5):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{with} \quad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{1.6}$$

where $\mathbf{X}$ is again the matrix of explanatory variables and $\boldsymbol{\beta}$ the corresponding parameter vector. $\mathbf{u}$ denotes the spatially dependent error term, $\mathbf{W}$ is the spatial weights matrix as in section 1.3.2 and λ is the spatial autoregression parameter. $\boldsymbol{\varepsilon}$ represents the vector of (spatially) independent residual errors which are normally distributed with zero mean and diagonal covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ with elements $\sigma^2_{\varepsilon_i}$, $i = 1, \ldots, n$ or often a joint variance $\sigma^2_{\varepsilon_i} = \sigma^2_\varepsilon$ (Schabenberger and Gotway 2005, pp. 335; Bivand et al. 2008, pp. 277). The model equation in (1.6) can be easily rewritten in the following way:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\beta} - \lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \end{aligned} \tag{1.7}$$

Some further manipulation of equation (1.7) yields equation (1.8):

$$\begin{aligned} \mathbf{y} - \lambda\mathbf{W}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} - \lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{y} &= (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \end{aligned} \tag{1.8}$$

assuming the invertibility of $\mathbf{I}_n - \lambda\mathbf{W}$. Equations (1.7) and (1.8) clearly show the similarity of the SE model to the linear regression model in equation (1.3) (Schabenberger and Gotway 2005, p. 336). Instead of the uncorrelated errors $\boldsymbol{\varepsilon}$ in (1.3), spatial autocorrelation is induced by introducing the error term $(\mathbf{I}_n - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}$ in (1.8). From the representation in equation (1.7) one can see the two additional terms $\lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ and $\lambda\mathbf{W}\mathbf{y}$ in the regression model compared to model (1.3). These terms are called the spatially lagged explanatory variables ($\lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta}$) and the spatially lagged values of the response variable ($\lambda\mathbf{W}\mathbf{y}$). According to Kissling and Carl (2008, p. 3), the SE model is used if the analyst assumes that the covariates $\mathbf{X}$ do not completely explain the spatial autocorrelation in the data, denoting this case as 'induced spatial dependence'. This case occurs e.g. if important spatially influenced covariates are not included in the analysis. Another motivating reason for this kind of model arises if spatial autocorrelation is an inherent characteristic of the response variable $\mathbf{y}$ itself, denominating this case as 'inherent spatial autocorrelation'.

Omitting the term $-\lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ in equation (1.7) yields the so-called *spatial lag (**SL**)* model (Bivand et al. 2008, p. 291; Kissling and Carl 2008, p. 3) or – equivalently in some other references – the *spatial autoregressive* model (Gibbons and Overman 2010, pp. 4; LeSage and Pace 2009, pp. 8). Apart from the influence of the explanatory variables $\mathbf{X}$, the response variable $\mathbf{y}$ also depends on its spatially lagged values $\rho\mathbf{W}\mathbf{y}$ so that we obtain the regression equation in formula (1.9):

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.9}$$

where ρ is the spatial autoregression parameter, $\boldsymbol{\varepsilon}$ again represents the vector of (spatially) independent residual errors and the other terms are as above. In contrast to the SE model, the SL model assumes that there is only 'inherent spatial autocorrelation' present in the data and therefore the spatial autoregressive process is included in the response variable itself (Kissling and Carl 2008, p. 3). The term $\rho\mathbf{W}\mathbf{y}$ describes the relation between the values of the dependent variable $\mathbf{y}$ and the neighboring values to each observation of $\mathbf{y}$ (LeSage and Pace 2009, p. 9). Rewriting the regression equation in (1.9) leads to the following alternative representation of the SL model in equation (1.10):

$$\begin{aligned} \mathbf{y} - \rho\mathbf{W}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ (\mathbf{I}_n - \rho\mathbf{W})\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{y} &= (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \end{aligned} \tag{1.10}$$

The special case of the SL model (1.9), where the response variable $\mathbf{y}$ only depends on its own spatial lag ($\rho\mathbf{W}\mathbf{y}$) and where no covariates are included in the regression equation, is called *pure spatial lag (**pure SL**)* model:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{1.11}$$

Assuming that the response variable $\mathbf{y}$ does not depend on its own spatial lag $\mathbf{W}\mathbf{y}$, but on the spatial lags of the explanatory variables $\mathbf{W}\mathbf{X}$, yields the *spatial lag of X (**SLX**)* model (Gibbons and Overman 2010, p. 5; LeSage and Pace 2009, p. 30):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{1.12}$$

Finally, the combination of SL (1.9) and SLX (1.12) model leads to the *Spatial Durbin (**SD**)* model (Gibbons and Overman 2010, pp. 5; Kissling and Carl 2008, p. 3):

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{1.13}$$

The SD model in (1.13) presumes that spatial autocorrelation affects both response and explanatory variables, but drops the assumption of spatial dependence in the error process. Apart from nesting the SL and the SLX model one can see that the model equation in (1.13) is also linked to the model equation (1.7) of the SE model.

As the SD model in (1.13) incorporates some other spatial regression models, this model is often estimated first and then tested against the more specific models. Moreover, it is common to compare models with different specifications of the spatial weights matrix and, of course, with different combinations of covariates (Gibbons and Overman 2010, p. 7). A commonly used instrument for comparing models is the Akaike Information Criterion (AIC), accounting both for model fit and model complexity (Kissling and Carl 2008, p. 5).

Some other well-known spatial regression models such as the *spatial conditional autoregressive (**CAR**)* model or the *simultaneous/spatial moving average (**SMA**)* model are not presented in this book chapter. For details concering these models see e.g. Schabenberger and Gotway (2005), Waller and Gotway (2004) and Haining (1993). For a recent review of what is known as spatial econometrics see Anselin (2007).

To fit spatial regression models in R the `spdep` package provides various functions, some of them using different methods to estimate the parameters (see Bivand et al. 2010 and Bivand et al. 2008, pp. 277-296):

- `spautolm`: maximum likelihood estimation for SE model (1.6), CAR model and SMA model; model type can be chosen by the option `family=("SAR","CAR","SMA")`; is based on the function `errorsarlm`
- `lagsarlm`: maximum likelihood estimation for the SL model (1.9) and the SD model (1.13); the default setting of the option `type="lag"` is used for the SL model, for the SD model set `type="mixed"`

- `stsls`: fits also a SL model (1.9) to the data using a two stage least squares procedure in a simultaneous system of equations by using the spatial lags of the covariates as instruments for the spatially lagged dependent variable

- `errorsarlm`: maximum likelihood estimation of the SE model (1.6)

- `GMerrorsar`: a generalized moments estimator for the autoregressive parameter in a SE model (1.6)

For more details on the options of the different R functions please see the individual help files of the routines.

Finally, we present the estimation outputs for some of these spatial regression models. The first output shows the result of the ML estimation for the SE model using the function `spautolm` or, equivalently, the function `errorsarlm`. The regression coefficients estimated in this model are very similar to those estimated in the linear regression model (similar values, same signs, all significant). The spatial dependency in the data is estimated via the spatial autoregression parameter λ:

```
Call: spautolm(formula = data$R08 ~ data$HEIGHT + data$R07 +
    data$R05 + data$R03, data = data, listw = ooe_W1)

Residuals:
       Min         1Q     Median         3Q        Max
-1.1343847 -0.0596381  0.0042328  0.0731940  1.2263243

Coefficients:
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  1.9833e-01  4.6558e-02  4.2599 2.045e-05
data$HEIGHT -3.3582e-04  9.5809e-05 -3.5051 0.0004565
data$R07     3.1967e-01  2.3904e-02 13.3733 < 2.2e-16
data$R05     2.7082e-01  2.9918e-02  9.0521 < 2.2e-16
data$R03     2.7501e-01  3.0169e-02  9.1155 < 2.2e-16

Lambda: 0.23115 LR test value: 7.5206 p-value: 0.0060997

Log likelihood: 67.4818
ML residual variance (sigma squared): 0.042974, (sigma: 0.2073)
Number of observations: 445
Number of parameters estimated: 7
AIC: -120.96
```

If we fit a SL model using the function `lagsarlm`, we once again obtain similar estimation results. However some parameter estimates of the SL model are only half as large as their SE model counterparts, and standard errors are consistently 10-20% smaller. Also, it is worth noting that the SL model fits much better than the SE model, as determined by AIC.

```
Call:lagsarlm(formula = data$R08 ~ data$HEIGHT + data$R07 +
    data$R05 + data$R03, data = data, listw = ooe_W1)

Residuals:
       Min         1Q     Median         3Q        Max
-0.9296895 -0.0713094  0.0025575  0.0701690  1.1748631

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  1.0063e-01  3.6887e-02  2.7281   0.00637
data$HEIGHT -1.6103e-04  7.5936e-05 -2.1206   0.03395
data$R07     2.3497e-01  2.2523e-02 10.4324 < 2.2e-16
data$R05     2.2526e-01  2.7750e-02  8.1174 4.441e-16
data$R03     2.3171e-01  2.7484e-02  8.4307 < 2.2e-16

Rho: 0.28677, LR test value: 96.53, p-value: < 2.22e-16
Asymptotic standard error: 0.02701
    z-value: 10.617, p-value: < 2.22e-16
Wald statistic: 112.72, p-value: < 2.22e-16

Log likelihood: 111.9866 for lag model
ML residual variance (sigma squared): 0.035063, (sigma: 0.18725)
Number of observations: 445
Number of parameters estimated: 7
AIC: -209.97, (AIC for lm: -115.44)
LM test for residual autocorrelation
test value: 1.1163, p-value: 0.29072
```

The SD model includes not only the covariates $\mathbf{X}$ and the spatial lag of the dependent variable $\mathbf{y}$, but also the spatial lags of the explanatory variables. If we estimate such a model using the function `lagsarlm` (option `type="mixed"`), we obtain the following output:

```
Call:lagsarlm(formula = data$R08 ~ data$HEIGHT + data$R07 +
    data$R05 + data$R03, data = data, listw = ooe_W1, type = "mixed")

Residuals:
       Min         1Q     Median         3Q        Max
-0.9252634 -0.0698749  0.0014853  0.0638801  1.1744096

Type: mixed
Coefficients: (asymptotic standard errors)
                  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)      0.07277795  0.04885996  1.4895  0.136350
data$HEIGHT     -0.00035993  0.00013149 -2.7372  0.006196
data$R07         0.23172127  0.02345894  9.8777 < 2.2e-16
data$R05         0.22601075  0.02818171  8.0198 1.110e-15
data$R03         0.22843739  0.02740654  8.3351 < 2.2e-16
lag.data$HEIGHT  0.00025899  0.00017339  1.4937  0.135243
lag.data$R07     0.00923654  0.05165748  0.1788  0.858092
lag.data$R05     0.22267410  0.08011565  2.7794  0.005446
lag.data$R03    -0.23495689  0.08636231 -2.7206  0.006516

Rho: 0.29162, LR test value: 19.153, p-value: 1.2063e-05
Asymptotic standard error: 0.07763
```

```
z-value: 3.7565, p-value: 0.00017229
Wald statistic: 14.112, p-value: 0.00017229

Log likelihood: 117.9214 for mixed model
ML residual variance (sigma squared): 0.034128, (sigma: 0.18474)
Number of observations: 445
Number of parameters estimated: 11
AIC: -213.84, (AIC for lm: -196.69)
LM test for residual autocorrelation
test value: 0.013289, p-value: 0.90822
```

The regression coefficients of the covariates and the spatial autoregressive parameter are still very similar to the estimates of the previous models, while the intercept and the spatial lags of the altitude and the ratio *R07* do not have a significant influence on the dependent variable. However, it is somewhat suspicious that the spatially lagged ratio *R03* has a negative effect on the dependent variable, which should be analyzed more carefully.

Note that the assumed spatial model often not only determines the behaviour under the alternative, but can also govern the choice of the spatial dependence measure. Recently Li et al. (2007) have suggested the APLE statistics, given by

$$\mathfrak{I}_{APLE} = \frac{\hat{\boldsymbol{\varepsilon}}' \frac{1}{2}(\mathbf{W} + \mathbf{W}')\hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}'[\mathbf{W}'\mathbf{W} + \mathrm{tr}(\mathbf{W}^2)\mathbf{I}_n/n]\hat{\boldsymbol{\varepsilon}}}, \tag{1.14}$$

for a better reflection of dependence under a SL alternative. Asymptotic and exact distributions for the APLE (and its potential consequent use in the next chapter) were provided by Reder and Müller (2009) and Li et al. (2010). Also for SL models a theoretical comparison of $\mathfrak{I}$ and Lagrange multiplier tests is given in Baltagi and Yang (2010). A simple regression based formulation can be found in Born and Breitung (2011).

1.6 Design considerations

Let us consider the first case (see section 1.5.1), where we estimate a model under the assumption of spatial independence, and the true model is of the same form. The aim is then not to reject the null hypothesis (spatial independence). For the approximate test we require the moments of Moran's $\mathfrak{I}$, which can be expressed in terms of the eigenvalues of the matrix $\mathbf{K}$ (Tiefelsdorf 2000), with $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denoting the general projection matrix. Since only the moments are of interest, the evaluation of eigenvalues can be by-passed by making use of the trace operator tr. In this case under the assumption of spatial independence, expected value and variance of $\mathfrak{I}$ are then given by

$$E[\mathfrak{I} \mid H_0] = \frac{\mathrm{tr}(\mathbf{K})}{n-k} = \frac{\mathrm{tr}\{\mathbf{M}\frac{1}{2}(\mathbf{W} + \mathbf{W}')\mathbf{M}\}}{n-k} = \frac{\mathrm{tr}(\mathbf{MW})}{n-k} \tag{1.15}$$

and

$$\begin{aligned}\text{Var}[\mathfrak{I} \mid H_0] &= \frac{\text{tr}(\mathbf{MWMW}') + \text{tr}(\mathbf{MW})^2 + \{\text{tr}(\mathbf{MW})\}^2}{(n-k)(n-k+2)} - \{E[\mathfrak{I} \mid H_0]\}^2 \\ &= \frac{2\{(n-k)\text{tr}(\mathbf{K}^2) - \text{tr}(\mathbf{K})^2\}}{(n-k)^2(n-k+2)} \end{aligned} \tag{1.16}$$

respectively, see Henshaw (1966).

An application of the theoretical moments of Moran's $\mathfrak{I}$ is the approximation of the exact distribution of Moran's $\mathfrak{I}$ by well-known simple distributions, that allow fast assessment of the significance of an observed Moran's $\mathfrak{I}$ without numerical evaluation of its exact probability. If the skewness and the kurtosis of Moran's $\mathfrak{I}$ (see Tiefelsdorf 2000) do not differ substantially from their counterparts of the normal distribution, the z-transformation of Moran's $\mathfrak{I}$ can be used to obtain the significance of an observed Moran's $\mathfrak{I}$. However, if there is a marked difference between the skewness and the kurtosis of Moran's $\mathfrak{I}$ to those of the normal distribution, alternative approximation strategies need to be employed.

The null case is the simpler one, there is no spatial effect in the data, data follow an ordinary linear model, the correct model is estimated and the null hypothesis of no spatial dependence should be retained. The intention is to find an optimal design which gives the best locations for the observations in the sense that the rejection of the null hypothesis is minimized.

Under the alternative, the (wrongly) estimated model is still: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}$ i.i.d., but now the true assumed (but unknown) data generating process is e.g. a SAR error process (1.6). The variance-covariance matrix $\boldsymbol{\Omega}(\rho)$ of the error terms is

$$\boldsymbol{\Omega}(\rho) = E[\mathbf{uu}'] = \sigma^2[(\mathbf{I}_n - \rho\mathbf{W})'(\mathbf{I}_n - \rho\mathbf{W})]^{-1} \tag{1.17}$$

To ensure that $\boldsymbol{\Omega}(\rho)$ is positive definite, ρ is restricted to the interval $]\frac{1}{\lambda_{min}}; \frac{1}{\lambda_{max}}[$, where λ_{min} and λ_{max} denote the smallest and largest eigenvalues of $\mathbf{W}$. Note that we are using SAR without being restricted to it, being well aware of its peculiar problems that might effect design considerations as described in Martellosio (2011). In fact, it is only necessary to be able to compute $\boldsymbol{\Omega}$ in the following. For the use of a CAR alternative, see e.g. Müller and Waldl (2011).

The model is estimated via OLS and the residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ are used for the calculation of Moran's $\mathfrak{I}$. If the real data generating process follows a SAR error process, the aim is to reject the null hypothesis of no spatial dependence. The task is to maximize the power of the test, i.e. the probability to reject the null hypothesis given the alternative (spatial dependence). For the normal approximation again only the conditional moments are needed. The conditional expectation of Moran's $\mathfrak{I}$ (cf. Tiefelsdorf 2000) can be evaluated by the improper integral

$$E[\mathfrak{I}|H_A] = \int_0^\infty \prod_{i=1}^{n-k} (1 + 2\lambda_i t)^{-\frac{1}{2}} \cdot \sum_{i=1}^{n-k} \frac{h_{ii}^*}{1 + 2\lambda_i t}\, dt \tag{1.18}$$

where h_{ii}^* are the diagonal elements of matrix $\mathbf{H} = \mathbf{P}'\mathbf{A}\mathbf{P}$ with $\mathbf{A} = \mathbf{\Omega}'^{\frac{1}{2}}\mathbf{M}\frac{1}{2}(\mathbf{W} + \mathbf{W}')\mathbf{M}\mathbf{\Omega}^{\frac{1}{2}}$ and $\mathbf{P}$ is the matrix of the normalized eigenvectors of matrix $\mathbf{B} = \mathbf{\Omega}'^{\frac{1}{2}}\mathbf{M}\mathbf{\Omega}^{\frac{1}{2}}$. The eigenvalues and their associated eigenvectors are re-sequenced so that $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_{n-k}$. The variance of $\mathfrak{I}$ under the alternative is given by

$$\text{Var}[\mathfrak{I}|H_A] = E[\mathfrak{I}^2|H_A] - E[\mathfrak{I}|H_A]^2 \tag{1.19}$$

where

$$E[\mathfrak{I}^2|H_A] = \int_0^\infty \left[\prod_{i=1}^{n-k}(1+2\lambda_i t)^{-\frac{1}{2}}\right] \cdot \left[\sum_{i=1}^{n-k}\sum_{j=1}^{n-k}\frac{h_{ii}^* h_{jj}^* + 2(h_{ij}^*)^2}{(1+2\lambda_i t)(1+2\lambda_j t)}\right] t\, dt$$

and $E[\mathfrak{I}|H_A]$ is given in equation (1.18). The upper truncation points for the integrals can be approximated by a formula from De Gooijer (1980). Following this, we obtain an approximation of the upper bound for the expected value (1.18) of

$$\left[\frac{(n-k)h_{max}}{2\lambda_1^{\frac{n-k}{2}}}\left(\frac{n-k}{2}-1\right)\frac{1}{\epsilon}\right]^{\frac{1}{\frac{n-k}{2}-1}} = \tau_1 \tag{1.20}$$

where h_{max} is the biggest absolute value of the elements of the diagonal of matrix $\mathbf{H}$ and ϵ is a given positive small number less than 1. An approximation of the upper bound for $E[\mathfrak{I}^2|H_A]$ is

$$\left[\frac{3(n-k)^2 h_{max}^{(2)}}{(2\lambda_1)^{\frac{n-k}{2}}}\left(\frac{n-k}{2}-2\right)\frac{1}{\epsilon}\right]^{\frac{1}{\frac{n-k}{2}-2}} = \tau_2 \tag{1.21}$$

with $h_{max}^{(2)}$ denoting the biggest absolute value of the elements of matrix $\mathbf{H}$. Tiefelsdorf (2000) suggests to use $\frac{1}{n-k}\sum_{i=1}^{n-k}\lambda_i$ instead of λ_1. For more details and an implementation of the above in R see Bivand et al. (2009).

1.6.1 A Design Criterion

In both cases, where a linear regression model is estimated and the corresponding residuals are used to calculate Moran's $\mathfrak{I}$ test, the test result, whether to reject or not to reject the null hypothesis of no spatial autocorrelation in the error term, depends on the true data generating process. As the true process is unknown, a general design criterion $\mathcal{J}$ (which does not depend on the knowledge of the true data generating process), is needed. The aim is to minimize the probability that, given the alternative, the Moran's $\mathfrak{I}$ test does not reject the null hypothesis of no spatial autocorrelation. The test statistic $Z = \frac{\mathfrak{I} - E(\mathfrak{I}|H_0)}{\sqrt{\text{Var}(\mathfrak{I}|H_0)}}$ is asymptotically normally distributed and therefore we minimize:

$$\min_{H_A} P\left(\frac{\mathfrak{I} - E(\mathfrak{I}|H_0)}{\sqrt{\text{Var}(\mathfrak{I}|H_0)}} \leq \Phi^{-1}(1-\alpha)\right),$$

where Φ denotes the cdf of the standard normal distribution. This leads to

$$\min_{H_A} P\left(\mathfrak{I} \leq \Phi^{-1}(1-\alpha)\sqrt{\text{Var}(\mathfrak{I}|H_0)} + E(\mathfrak{I}|H_0)\right)$$

Using the z-transformation for $\mathfrak{I}$ under the alternative gives $\frac{\mathfrak{I}-E[\mathfrak{I}|H_A]}{\sqrt{\text{Var}[\mathfrak{I}|H_A]}}$, which is also asymptotically standard normally distributed. The final criterion to be maximized is therefore given by

$$\mathcal{J}_{\mathfrak{I}}(\xi) = 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{\text{Var}[\mathfrak{I}|H_0]} + E[\mathfrak{I}|H_0] - E[\mathfrak{I}|H_A]}{\sqrt{\text{Var}[\mathfrak{I}|H_A]}}\right). \quad (1.22)$$

The maximization of $\mathcal{J}_{\mathfrak{I}}$ over $S_\xi \in \mathcal{X}$ gives the final optimal locations for the observation sites and thus maximizes the power of the Moran's $\mathfrak{I}$ test. To calculate $\mathcal{J}_{\mathfrak{I}}$, the expected value (1.15) and the variance (1.16) of $\mathfrak{I}$ under the null hypothesis, and the expected value (1.18) and the variance (1.19) of $\mathfrak{I}$ under the alternative hypothesis are needed. Unfortunately, the given criterion is not convex and thus we can not employ the sort of equivalence theorems from the well developed optimum design theory (cf. Atkinson et al. 2007) but must resort to alternative algorithmic approaches, as given below. This criterion was first suggested by Gumprecht et al. (2009) and later extended by Müller et al. (2012) for the exact distribution. They show that optimizing the design serves as a regulatory device for the validity of the normal approximation, so it is sufficient to consider the approximation in what follows.

Evidently, the global optimal design can be found by evaluating all possible designs, i.e. in an m-point grid there are $\binom{m}{r}$ possible r-point designs, r goes from $4+k+1$ to m, where k is the number of the regressors in the model. This minimum number of points in a design follows from the approximation of the upper truncation points for the integrals (1.20) and (1.21). The number of possible designs increases very fast with the size of the grid. This leads to a high runtime, as the numerical integration needs a considerable amount of time. From this point of view it is worth to notice that not all possible designs are different in the sense that they have different criterion values. Some of the r-point designs are only rotations, reflections or translations of other r-point designs, and therefore give the same value of the criterion $\mathcal{J}_{\mathfrak{I}}$; let us call the respective designs 'symmetric'. To avoid calculating $\mathcal{J}_{\mathfrak{I}}$ for those designs which are known to be symmetric to others, an appropriate symmetry check can be performed before the computation of $\mathcal{J}_{\mathfrak{I}}$ (see Gumprecht et al. 2009).

For our asymmetric setup, however, there is no hope for that because on our 445-point lattice there are (440-k)! different potential designs and only very few, if any, of them can be considered symmetric. Thus, the complete evaluation of all truly different designs is mostly impractical and can only be performed for very small designs.

Gumprecht et al. (2009) suggest a simple search algorithm for finding a 'nearly' optimal design. This algorithm is much faster than the full enumeration algorithm

as for the r-point design the number of evaluated $(r-1)$-point designs is r. This algorithm can also be performed in an acceptable amount of time for relatively large grids. The procedure is quite simple:

1. Start with an initial design ξ_0 with $S_{\xi_0} = \mathcal{X}$, called 'base' design. Thus in the first iteration the number of points r in ξ_0 is m.
2. Delete each point, one at a time, to get $(r-1)$ designs ξ_e, and compute $\mathcal{J}_{\mathfrak{I}e}$. The symmetries can be checked before the criterion is calculated.
3. Take the best $(r-1)$ design ξ_e, i.e. the design with the largest $\mathcal{J}_{\mathfrak{I}e}$, and put it as new base design.

   Go to step 2.

The algorithm stops if $r = (4 + k + 1)$. The r-point design that gives the largest $\mathcal{J}_{\mathfrak{I}}$ is the 'nearly' optimal one. Note the similarities to the 'coffee-house' procedure given in Müller (2007): the disadvantage of these algorithms is, that once a r-point design is chosen, all smaller $r - i$ point designs are restricted to this set of points. As a result it can happen quite easily that the algorithm is trapped in a local maximum. To avoid this one could alternatively employ methods of stochastic optimization such as in Haines (1987) or more recently Ver Hoef (2011).

Computation of all designs in this section is again what is known as a NP-hard problem and a rigorous search would be prohibitive. Besides, the criteria are not convex and typically exhibit multiple local optima and therefore most of the cited papers employ simulated annealing algorithms (cf. also van Groenigen and Stein 1998). However, by using simple exchange type procedures (cf. Royle 2002) considerable gains in the criteria can be achieved within a few steps, as can be seen from the respective examples in the next section.

1.6.2 Example

In the example of the Upper-Austrian municipalities it would clearly be too demanding to search through the whole 445 item grid of available locations. We employ the spatial link matrix $\mathbf{G}$ implied by Figure 1.3. In the search- and exchange algorithms the corresponding row-standardized spatial weight matrices $\mathbf{W}$ are used. The regression model of the optimal design procedure is an intercept only model. As an exemplary value of the spatial autoregressive parameter $\rho = 0.28677$ (as estimated from the SL model) was used, which is needed for evaluating expressions under the alternative hypothesis (other values give qualitatively similar results though differing designs). Using this parameter and sequential elimination (simple search) leads to the optimal criteria values (maximum power of the test) for the respective n-point designs displayed in Figure 1.6. From this graph it can be seen that the best is the 122-point design with $\mathcal{J}_{\mathfrak{I}} = 0.996311$. The strong decay for larger numbers of observations is another instance of the power-trap described in Krämer (2005) and Martellosio (2010). One should note the following aspects about the optimization:

- In the first part of the optimization procedure there seem to be a lot of problems in the computations, obtaining several NaN's and oscillations in powers. Nobody can be sure that the selection of the nodes to remove is the optimal one (another reason to say that the final design obtained is quasi-optimal).
- Only after reaching the 381-point design the power begins to increase steadily, getting the maximum value for a 122-point design.
- The 108-point design has all the nodes connected in pairs. After reaching this design, zero powers begin to appear when one of the nodes of a pair is removed. From that point on, the results of selecting the best node to remove are no longer very reliable (in most cases the first node of the remaining ones is removed).

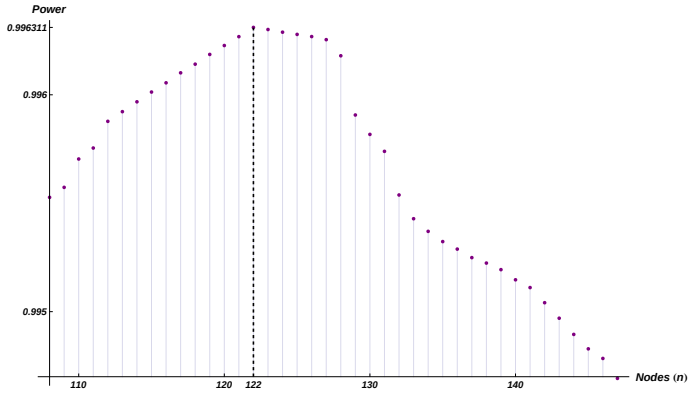

Figure 1.6: Maximum powers of Moran's $\mathfrak{I}$ test for various design sizes n.

Note that we assume here the number of sampling sites to be freely chosen; some considerations on the effectiveness can be found in Griffith (2005) and Griffith (2008). Thus a corresponding 122-point design can be selected exclusively on the basis of the criterion, which is displayed in Figure 1.7. We can observe that this design only consists of connected couples, triples and at maximum quadruples.

1.7 Discussion

To conclude we have to admit that there are several different methods to fit a regression model to areal data, but it is not obvious up to now which one of the presented methods is optimal for the given data set, nor are the implications on a proper design procedure straightforward. Some questions – which go beyond the scope of this chapter – still remain unsettled:

- Should we prefer the linear regression model, where we face the problem of collinearity, but eliminate the spatial effect in the data?
- Do we have the case of spurious spatial regression in the given data? How can we test for spatial nonstationarity?
- If it is feasible to fit a spatial regression model, which one should we use in this context?

To answer these questions further work has to be done and the mentioned topics must be examined in detail.

Note that the techniques given in this section have also some impact on methods for spatial filtering, where the main idea is to separate the regional interdependencies by partitioning the original variable into a filtered non-spatial (so called 'spaceless') variable and a residual spatial variable. Afterwards conventional statistical techniques that are based on the assumption of spatially uncorrelated errors can be used for the filtered part. One of the most common filters is based on an eigenfunction decomposition related to Moran's $\mathfrak{I}$ (cf. Getis and Griffith 2002), and thus may be improved by appropriately selecting supporting sites.

Once spatial dependence in the data is detected the so-called variogram plays a central role in the analysis of spatial data. A valid variogram model is selected and the parameters of this model are estimated before kriging (spatial prediction) is performed. These inference procedures are generally based on the examination of the empirical variogram, which consists of average squared differences of data taken at sites lagged the same distance apart in the same direction. The ability

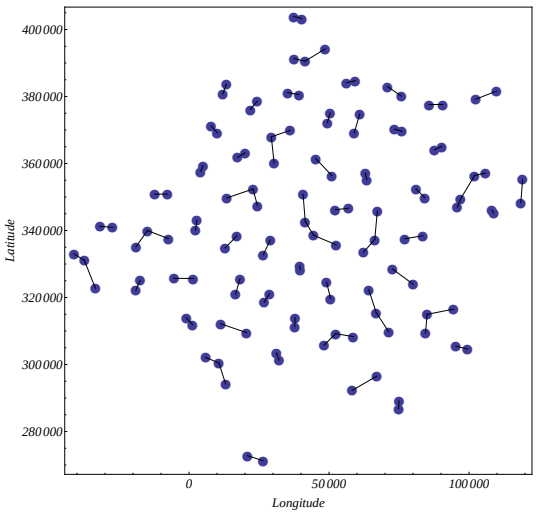

Figure 1.7: A quasi-optimum design for detecting spatial dependence.

of an investigator to estimate the parameters of a variogram model efficiently is again significantly affected by the sampling design, i.e. the locations of the sites $x_1, \ldots, x_n \in \mathcal{X}$ where data $\mathbf{y}$ are observed. For the relationship of variograms to g-ratios and its implications see Bellehumeur and Legendre (1998).

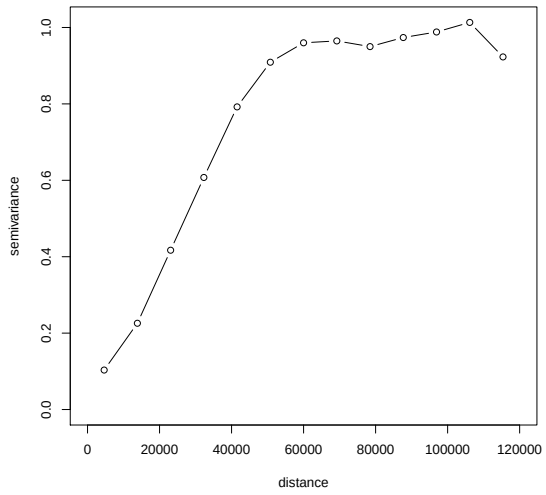

Figure 1.8: Variogram for the greenland data set.

In Müller and Zimmerman (1999) several practical approaches for constructing sampling designs for estimation of the variogram were compared by Monte-Carlo simulations. Those designs could be adopted at the early stages of a sampling program until the variogram is sufficiently well-estimated, after which one could shift to an existing approach that emphasizes prediction. Alternatively, the two objectives could be combined in a compound design, as suggested in Müller and Stehlík (2010).

Instead of directly going after the variograms a number of alternative methods were suggested which allow to ignore correlations. A two stage strategy for instance was suggested by Müller and Zimmerman (1995):

(a) Find the optimal configuration of distances $\xi^*_{\mathcal{L}}$ in the space spanned by all possible point pair distances (the so-called lag space $\mathcal{L}$),

(b) and map this configuration into the original site space $\mathcal{X}$ (find a site space design).

Finding the solution of (b) was also the purpose of Warrick and Myers (1987). The usefulness of such a distance algorithm can only be assessed in two ways: via

simulation, or by comparing it to a technique that directly employs ideas for optimum designs for correlated observations. First results in this direction can be found in Müller and Zimmerman (1999). They conclude that algorithms based on ignoring correlations are much quicker but marginally less efficient with respect to a design criterion than the augmentation procedures from the previous section.

It is traditional practice that the covariance/variogram parameters are estimated in a separate stage. However, if one is willing to make distributional assumptions, it is natural to employ likelihood based estimation techniques. In the following we will thus assume that the errors in our spatial model follow a stationary Gaussian process.

In particular one could now assume that the trend is known and fixed and maximize the log likelihood

$$2L(\boldsymbol{\theta}) = -n\log(2\pi) - \log\det \mathbf{C}(\boldsymbol{\theta}) - \mathbf{y}'\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{y}. \tag{1.23}$$

It is now natural to base a design criterion on the information matrix associated with the corresponding estimate of the parameter $\boldsymbol{\theta}$, which is given by (note that it depends upon a design ξ via $\mathbf{C}$)

$$\mathbf{M}''(\xi,\boldsymbol{\theta})_{jj'} = \frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_{j'}}\right\}, \tag{1.24}$$

where the $\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_j}$ are $n \times n$ matrices with entries $\frac{\partial c(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})}{\partial \theta_j}$, $\mathbf{x}, \mathbf{x}' \in \xi$.

Designs maximizing the determinant of $\mathbf{M}''$ have been suggested by Zhu and Stein (2006) (they also employ a minimax and Bayesian criterion to avoid undesirable effects due to the linearizations) and Zimmerman (2006), who calls them CP (covariance parameter) optimal. Both demonstrate the behaviour of the criteria for numerous artificial and real examples. A related discussion in this volume is given by Zimmerman (2012); for a Bayesian adaptive approach see Marchant and Lark (2006).

1.8 Appendix: R Code

```
# Packages
require(maptools)
require(maps)
require(spdep)
require(RColorBrewer)
require(pgirmess)
require(HH)
require(lmtest)
require(sandwich)

# Map of Upper Austria
data=data.frame(read.csv("greenlandmunratio_cs.csv",header=TRUE,
sep=";",dec=".")[,1:12])
years=names(data[,7:12])
row.names(data)=data[,3]
getinfo.shape("ooegemeinden.shp")
ooe <- readShapePoly("ooegemeinden.shp", IDvar="LBBGG")
plot(ooe, border="blue", axes=TRUE, las=1)
Greenland=SpatialPolygonsDataFrame(ooe,data=data)
```

```
# Video
dir.create("movieGreenland")
max(data[,7:12])
min(data[,7:12])
at_green=pretty(as.vector(unlist(data[,7:12])),n=10)
cols_green=colorRampPalette(rev(brewer.pal(10,"RdYlBu")))
(length(at_green)-1)   # rev() re-sorts
for(i in years){
    png(file=paste("movieGreenland/pic",i,".png",sep=""),
     width=960,height=600)
    print(spplot(Greenland,i,col.regions=cols_green,
     at=at_green, main=paste("log Ratio: area of arable
     land and area of greenland",i,sep=" ")))
    dev.off()
}

# Plot of log ratios R95 to R08
pdf("Plots.pdf")
at_green=pretty(as.vector(unlist(data[,7:12])),n=9)
cols_green=colorRampPalette(brewer.pal(9,"Greens"))(length(at_green)-1)
spplot(Greenland,years,col.regions=cols_green,at=at_green,as.table=T)
dev.off()

# Influence of Altitude
pdf("Altitude.pdf")
at_altitude=pretty(as.vector(unlist(data[,6])),n=11)
cols_altitude=colorRampPalette(brewer.pal(9,"YlOrBr"))(length(at_altitude)-1)
altitude=names(data[6])
spplot(Greenland,altitude,col.regions=cols_altitude,at=at_altitude,as.table=T)
dev.off()



# Creating Neighbours

pdf("Neighbours.pdf")
par(mfrow=c(1,2))
ooe_nb1=poly2nb(Greenland)          # Queen-style contiguities
plot(Greenland,border="grey60")
plot(ooe_nb1,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
title(main="Queen-style contiguities")

ooe_nb2=poly2nb(Greenland, queen=FALSE)       # Rook-style contiguities
plot(Greenland,border="grey60")
plot(ooe_nb2,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
title(main="Rook-style contiguities")

coords=coordinates(Greenland)
IDs=row.names(as(Greenland,"data.frame"))
ooe_nb3=tri2nb(coords,row.names=IDs)          # Delauney triangulation
plot(Greenland,border="grey60")
plot(ooe_nb3,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
title(main="Delauney triangulation neighbours")

ooe_nb4=graph2nb(gabrielneigh(coords),row.names=IDs)        # Gabriel graph
plot(Greenland,border="grey60")
plot(ooe_nb4,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
title(main="Gabriel graph neighbours")

ooe_nb5=knn2nb(knearneigh(coords,k=1),row.names=IDs)        # k=1 neighbors
plot(Greenland,border="grey60")
plot(ooe_nb5,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
title(main="All areas k=1 neighbours")
dev.off()
```

```
pdf("Distbased.pdf")
dists=unlist(nbdists(ooe_nb5,coords))
summary(dists)
max_dist=max(dists)
# Distance based neighbors within 1*max_dist
ooe_nb6=dnearneigh(coords,d1=0,d2=1*max_dist,row.names=IDs)
is.symmetric.nb(ooe_nb6)
summary(ooe_nb6)
plot(Greenland,border="grey60")
plot(ooe_nb6,coordinates(Greenland),pch=19,cex=0.6,add=TRUE)
#title(main="Distance based neighbours within 1*max_dist")
dev.off()



# Spatial Weights

# available styles: W, B, C, U
# error when given an nb argument with areas with no neighbors = default

# row standardized weights matrix --> sums of weights in each row = 1
ooe_W1=nb2listw(ooe_nb6,style="W")

# weight of unity for each neighbor relationship
ooe_W2=nb2listw(ooe_nb6,style="B")

# equal weights for all links --> complete set of weights sums to number
  of areas
ooe_W3=nb2listw(ooe_nb6,style="C")

# equal weights for all links --> complete set of weights sums to 1
ooe_W4=nb2listw(ooe_nb6,style="U")

# Connectivity Matrix/Spatial Lag Matrix
W1=listw2mat(ooe_W1)
W2=listw2mat(ooe_W2)
W3=listw2mat(ooe_W3)
W4=listw2mat(ooe_W4)



# Spatial autocorrelation tests

# Moran's I
moran=moran.test(data$R08,listw=ooe_W1)     # randomisation=F
moran
moran1=moran.test(data$R08,listw=ooe_W1,randomisation=T)
moran1

# Spatial Correlogram
correlo=sp.correlogram(neighbours=ooe_nb6,var=data$R08,order=16,method="I",
           style="W",zero.policy=TRUE)
correlo
dist_correlo=correlog(coords,data$R08,method="Moran")
dist_correlo

pdf("Correlogram R08.pdf")
plot(correlo)
plot(dist_correlo)
# Moran Scatterplot (different for different spatial weight styles)
moran.plot(data$R08,listw=ooe_W1)
title(main="Moran Scatterplot")
dev.off()
```

```
# Linear Regression Models

# Scatterplot-Matrix (covariates)
pairs(data[,5:11])
cor(data[,5:11])

# Linear regression model
linreg=lm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03+data$R99+data$R95)
linreg=lm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03)
summary(linreg)

# Collinearity - Variance inflation factor
vif(linreg)
# Plot of the residuals
Greenland$lmresid=residuals(linreg)
at_res=pretty(as.vector(unlist(Greenland$lmresid)),n=9)
cols_res=colorRampPalette(brewer.pal(9,"Greens"))(length(at_res)-1)
spplot(Greenland,"lmresid",col.regions=cols_res,at=at_res,as.table=T)

# Moran's I test for residuals
moran_res=moran.test(Greenland$lmresid,listw=ooe_W1,randomisation=F)
# under randomisation/under normality --> randomisation=T/F
moran_res

lm.morantest(linreg,ooe_W1)



# Spatial regression models

# SAR
sar=spautolm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,listw=ooe_W1)
summary(sar)

# CAR
car=spautolm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,family="CAR",listw=ooe_W1)
summary(car)

# SMA
sma=spautolm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,family="SMA",listw=ooe_W1)
summary(sma)

# Spatial lag model
lag=lagsarlm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,listw=ooe_W1)
summary(lag)

# Plot of the residuals
Greenland$lagresid=residuals(lag)
at_res1=pretty(as.vector(unlist(Greenland$lagresid)),n=9)
cols_res1=colorRampPalette(brewer.pal(9,"Greens"))(length(at_res1)-1)
spplot(Greenland,"lmresid",col.regions=cols_res1,at=at_res1,as.table=T)

# Spatial Durbin model (spatially lagged explanatory variables)
mix=lagsarlm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,listw=ooe_W1,type="mixed")
summary(mix)
anova(lag,mix)     # AIC

# Spatial error model
er=errorsarlm(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,
     data=data,listw=ooe_W1)
summary(er)
```

```
# Alternatives
# two stage least squares procedure
stsls=stsls(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,data=data,
       listw=ooe_W1)
summary(stsls)
stslsR=stsls(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,data=data,
       listw=ooe_W1,robust=TRUE)
summary(stslsR)
#Generalized Moments estimator
GMerr=GMerrorsar(data$R08~data$HEIGHT+data$R07+data$R05+data$R03,data=data,
       listw=ooe_W1)
summary(GMerr)
```

1.9 Acknowledgement

We are grateful to an attentive referee, whose comments lead to an improvement of this contribution. Some work on this contribution was financed by Acciones Integradas 2008-2009 (Project Nr. ES 18/2008).

References/Further Reading and Bibliography

Anselin, L 1988 *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Anselin, L 1993 The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association In *Spatial analytical perspectives on GIS* (ed. Fischer MM, Scholten HJ and Unwin D) Taylor and Francis London pp. 111–125.

Anselin, L 1995 Local Indicators of Spatial Association – LISA.*Geographical Analysis*, **27**, 93–115.

Anselin L 2007 Spatial econometrics In *Palgrave Handbook of Econometrics: Volume 1: Econometric Theory* (ed. Patterson K and Mills TC) first edition edn Palgrave Macmillan pp. 310–330.

Atkinson A, Donev A and Tobias R 2007 *Optimum Experimental Designs, with SAS (Oxford Statistical Science Series)*. Oxford University Press, USA.

Arbia, G 2006 *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Springer Verlag, Berlin Heidelberg.

Baltagi BH and Yang Z 2010 Standardized LM Tests for Spatial Error Dependence in Linear or Panel Regressions. Technical Report 11-2010, Singapore Management University, School of Economics.

Banerjee S, Carlin BP and Gelfand AE 2004 *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, London.

Beenstock M and Felsenstein D 2008 Testing Spatial Stationarity and Spatial Cointegration. *Working Paper*.

Beenstock M and Felsenstein D 2010 Testing for Unit Roots and Cointegration in Spatial Cross Section Data. *Working Paper*.

Bellehumeur C and Legendre P 1998 Multiscale sources of variation in ecological variables: modeling spatial dispersion, elaborating sampling designs. *Landscape Ecology* **13**(1), 15–25.

Bivand RS, Pebesma EJ and Gómez-Rubio V 2008 *Applied Spatial Data Analysis with R*. Springer, New York.

Bivand RS, Müller WG and Reder M 2009 Power Calculations for Global and Local Moran's I. *Computational Statistics & Data Analysis* **53**, 2859–2872.

Bivand RS et al. 2010 *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-26. `http://CRAN.R-project.org/package=spdep`

Born B and Breitung J 2011 Simple regression-based tests for spatial dependence. *The Econometrics Journal* **14**(2), 330–342.

Brewer CA and Harrower M 2009 *ColorBrewer 2.0*. `http://www.colorbrewer2.org`, [accessed: January 25, 2012].

Brimkulov UN, Krug GK and Savanov VL 1980 Numerical construction of exact experimental designs when the measurements are correlated (in Russian). *Zavodskaya Laboratoria (Industrial Laboratory)* **36**, 435–442.

Cliff AD and Ord JK 1981 *Spatial Processes. Models & Applications* Pion Limited, London.

Dale MRT, Dixon P, Fortin MJ, Legendre P, Myers DE and Rosenberg MS 2002 Conceptual and mathematical relationships among methods for spatial analysis. *Ecography* **25**(5), 558–577.

De Gooijer JG 1980 Exact moments of the sample autocorrelations from series generated by general ARIMA processes of order (p,d,q), d=0 or 1. *Journal of Econometrics*, **14**, 365–379.

Dray S 2011 A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a Linear Regression Problem. *Geographical Analysis* **43**(2), 127–141.

Dubé J and Legros D 2011 A spatio-temporal measure of spatial dependence: An example using real estate data. *Papers in Regional Science*, doi: 10.1111/j.1435-5957.2011.00402.x

Fahrmeir L, Kneib T and Lang S 2009 *Regression. Modelle, Methoden und Anwendungen*. Springer, Berlin Heidelberg.
Fingleton B 1999 Spurious Spatial Regression: Some Monte Carlo Results with a Spatial Unit Root and Spatial Cointegration. *Journal of Regional Science* **39**, 1–19.
Fortin M-J and Dale M 2005 *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, Cambridge.
Geary RC 1954 The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*.
Getis A and Griffith D 2002 Comparative spatial filtering in regression analysis. *Geographical Analysis* **34**, 130–140.
Gibbons S and Overman HG 2010 Mostly Pointless Spatial Econometrics? *SERC Discussion Papers* **61**.
Griffith DA 2005 Effective Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers* **95**(4), 740–760.
Griffith DA 2008 Geographic sampling of urban soils for contaminant mapping: how many samples and from where.. *Environmental geochemistry and health* **30**(6), 495–509.
Griffith DA 2010 The Moran coefficient for non-normal data. *Journal of Statistical Planning and Inference* **140**(11), 2980–2990.
Gumprecht D 2007 Treatment of far-off objects in Moran's $\mathfrak{I}$ test. *Research Report Series, Vienna University of Economics and Business Administration* **46**.
Gumprecht D, Müller WG and Rodríguez-Díaz JM 2009 Designs for Detecting Spatial Dependence. *Geographical Analysis* **41**(2), 127–143.
Haines LM 1987 The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. *Technometrics* **29**, 439–447.
Haining RP 1993 *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge.
Henshaw RC 1966 Testing single-equation least squares regression models for autocorrelated disturbances. *Econometrica* **34**, 646–660.
Hepple LW 1998 Exact testing for spatial autocorrelation among regression residuals. *Environment and Planning A* **30**(1), 85–108.
Kissling WD and Carl G 2008 Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* **17**(1), 59–71.
Krämer W 2005 Finite sample power of Cliff-Ord-type tests for spatial disturbance correlation in linear regression. *Journal of Statistical Planning and Inference* **128**(2), 489–496.
Lauridsen J and Kosfeld R 2006 Spurious Spatial Regression, Spatial Cointegration and Heteroscedasticity. *Working Paper, Southern University of Denmark, Odense*.
LeSage J and Pace RK 2009 *Introduction to Spatial Econometrics*. Chapman & Hall/CRC.
Li H, Calder CA and Cressie N 2007 Beyond Moran's ¡i¿I¡/i¿: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. *Geographical Analysis* **39**(4), 357–375.
Li H, Calder CA and Cressie N 2010 One-step estimation of spatial dependence parameters: Properties and extensions of the APLE statistic. Technical Report 846, Department of Statistics, The Ohio State University.
López FA, Matilla-García M, Mur J and Marín MR 2011 Four tests of independence in spatiotemporal data. *Papers in Regional Science* **90**(3), 663–685.
Marchant BP and Lark RM 2006 Adaptive sampling and reconnaissance surveys for geostatistical mapping of the soil. *European Journal of Soil Science* **57**(6), 831–845.
Martellosio F 2010 Power properties of invariant tests for spatial autocorrelation in linear regression. *Econometric Theory* **26**(01), 152–186.
Martellosio F 2011 Nontestability of equal weights spatial dependence. *Econometric Theory* **27**(06), 1369–1375.
Moran PAP 1950 A test for the serial dependence of residuals. *Biometrika* **37**, 178–181.

Müller WG 2007 *Collecting Spatial Data: Optimum Design of Experiments for Random Fields* 3rd rev. and extended edn. Springer, Heidelberg.

Müller WG and Stehlík M 2010 Compound optimal spatial designs. *Environmetrics* **21**(3-4), 354–364.

Müller WG and Waldl H 2011 Discussion of lindgren, rue and lindström "an explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society - Series B* **73**(4), 483–485.

Müller WG and Zimmerman DL 1995 An algorithm for sampling optimization for semivariogram estimation In *Model-Oriented Data Analysis 4* (ed. Kitsos CP and Müller WG) Physica Heidelberg pp. 173–178.

Müller WG and Zimmerman DL 1999 Optimal design for variogram estimation. *Environmetrics* **10**, 23–37.

Müller WG, Rodríguez-Díaz JM and Rivas López MJ 2012 Optimal design for detecting dependencies with an application in spatial ecology. *Environmetrics* **23**(1), 37–45.

O'Sullivan D and Unwin DJ 2003 *Geographic Information Analysis*. Wiley & Sons, Hoboken, New Jersey.

Rangel TF, Diniz-Filho JA and Bini LM 2010 SAM: a comprehensive application for Spatial Analysis in Macroecology. *Ecography* **33**(1), 46–50.

Reder M and Müller W 2009 More on the APLE statistic. *Mathematica Slovaca* **59**(5), 565–578.

Royle JA 2002 Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference* **100**(2), 121–134.

Rue H and Held L 2005 *Gaussian Markov Random Fields: Theory and Applications (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)* 1 edn. Chapman and Hall/CRC.

Schabenberger O and Gotway CA 2005 *Statistical Methods for Spatial Data Analysis*. Chapman & Hall, London.

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)*. Springer.

Tiefelsdorf M and Boots B 1995 The exact distribution of Moran's I. *Environment and Planning A* **27**(6), 985–999.

van Groenigen JW and Stein A 1998 Constrained optimization of spatial sampling using continuous simulated annealing. *J Environ Qual* **27**(5), 1078–1086.

Ver Hoef JM 2011 Practical considerations for experimental designs of spatially autocorrelated data using computer intensive methods. *Statistical Methodology*.

Waller LA and Gotway CA 2004 *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, New York.

Warrick AW and Myers DE 1987 Optimization of sampling locations for variogram calculations. *Water Resources Research* **23**, 496–500.

Zhu Z and Stein ML 2006 Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **11**(1), 24–44.

Zimmerman DL 2006 Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17**(6), 635–652.

Zimmerman D 2012 Model-based frequentist design for univariate and multivariate geostatistics In *Spatio-temporal Design* (ed. Mateu J and Müller WG) Wiley & Sons New York pp. ?–?