

# Mat-2.4177

## Seminar on Case Studies in Operations Research

---

Kemira GrowHow: Adjusting estimates of  
locally dependent treatment means by use of  
spatial regression

Project Report  
30.4.2008

Ilkka Anttila  
Mikael Bruun (project manager)  
Antti Ritala  
Olli Rusanen  
Timo Tervola

## Table of contents

Table of contents .....	2
1. Introduction.....	3
1.1. Client.....	3
1.2. Background .....	3
1.3. Limitations.....	3
1.4. Problem statement.....	3
1.5. Structure of the Report.....	4
2. Theoretical Background .....	5
2.1. Descriptive Statistics .....	5
2.2. Analysis of Variance ANOVA .....	6
2.3. Ordinary Least Squares Regression OLS.....	7
2.4. Regression Diagnostics.....	8
2.4.1. Residual Maps and Plots.....	8
2.4.2. Diagnostics for Spatial Autocorrelation .....	10
3. Spatial Regression Model .....	12
3.1. Constructing Spatial Weights.....	12
3.2. Spatial Lag Model – SAR Spatial Autoregressive Model.....	14
3.3. Spatial Error Model - SEM .....	14
4. Our Approach.....	15
4.1. MATLAB Implementation .....	15
5. Results .....	17
6. Discussion .....	19
7. References.....	20
Appendix 1. Indexing and Data sets 1 and 2 .....	21
Appendix 2. Data set 1: 15x7 matrix (greenhouse trials) .....	22
Appendix 3. Data set 2: 4x3 matrix (wheat field trials) .....	23
Appendix 4. Regression diagnostics continued. Data sets 1 and 2.....	24
Appendix 5. Results of SEM and OLS regression for data set 1.....	25
Appendix 6. Moran’s I test for data set 2 .....	26
Appendix 7. Results of SEM and OLS regression for data set 2.....	27

## 1. Introduction

This study is part of the course 'Mat-2.4177 Seminar on Case Studies in Operations Research' held in spring 2008. The study aims to find more accurate estimates for means of field trial treatments by taking into account spatial correlation. The project team consisted of five minors of systems and operations research.

### 1.1. Client

Kemira GrowHow is the second largest producer of fertilizers in Europe. The company is a subsidiary of Norwegian Yara International, which is the largest fertilizer producer in the world. Kemira GrowHow's products are sold in over 100 countries. Their strongest market position is in Europe –especially in Northern Europe.

In 2007 Kemira GrowHow had 2435 employees and net sales were 1,294.7 million Euros. Kemira GrowHow's Crop Cultivation business unit produces fertilizers for agriculture, crop farming, and for gardening.

Kemira GrowHow is active in R&D in the fields of agronomy, organic and inorganic chemistry, and process technology. R&D projects are often long: developing fertilizer products from original ideas can take three to six years. Research is usually done in greenhouses and in test plantations. Experiments are done in different places of the world in order to test potential products in different environmental conditions. (Kemira GrowHow, Annual Report, 2007)

### 1.2. Background

In cultivation experiments, the impact of different treatments on crop yields are examined by dividing treatments to field plots in a specified way. Statistics is used to analyze the experiment results. Most of the commonly used methods of analysis assume that residuals are independent and normally distributed. In practice, residuals are often locally dependent. This is problematic because

1. local systematic error leads to overestimation of natural variation, which makes it more difficult to notice differences between different treatments;
2. replicates of the same treatment can be located in the problematic area, and thus, the mean effect of the treatment will become distorted.

Because of these problems, usual statistical methods cannot be used. Test results must be corrected so that local correlation will be diminished. We present methods both for testing whether results are locally dependent and for correcting test results accordingly.

### 1.3. Limitations

Two spatial regression models and four alternative contiguity weights are discussed. The regression models are the spatial autoregressive and spatial error model. Weights are based on row, column, and so-called Rook and Queen contiguity (see section 3.1).

### 1.4. Problem statement

The background presents the problem setting for the project. The results in cultivation experiments embody spatial correlation and the task of the project is to come up with means how to handle such a situation. The goal setting for the project is double-barreled; firstly the

target is to develop a gauge for measuring the level of spatial autocorrelation. Secondly the target is to develop a systematic approach to handling the spatially autocorrelated data.

### 1.5. Structure of the Report

The report is structured into six sections, a references section, and appendices 1-7. In section 2, the theoretical background is presented. Section 3 continues by introducing concepts important in forming spatial regression models. Our solution is presented in section 4, which includes a brief description of the MATLAB implementation. Section 5 presents the results. The last section discusses the findings.

## 2. Theoretical Background

This section presents the standard methods of statistical analysis used in this study. Data from actual greenhouse and wheat field trials is used in reviewing these methods (see Appendices 1 and 2 respectively). For indexing and basic information about the data, see Appendix 3. The discussion continues in Section 3.

### 2.1. Descriptive Statistics

In statistics, graphical representations are used for obtaining information on properties of data. Two especially useful representations include surface plot and map of observations. Assuming a randomized block design, one should expect observations of high and low-value to scatter randomly. There should not be any clustering of values.

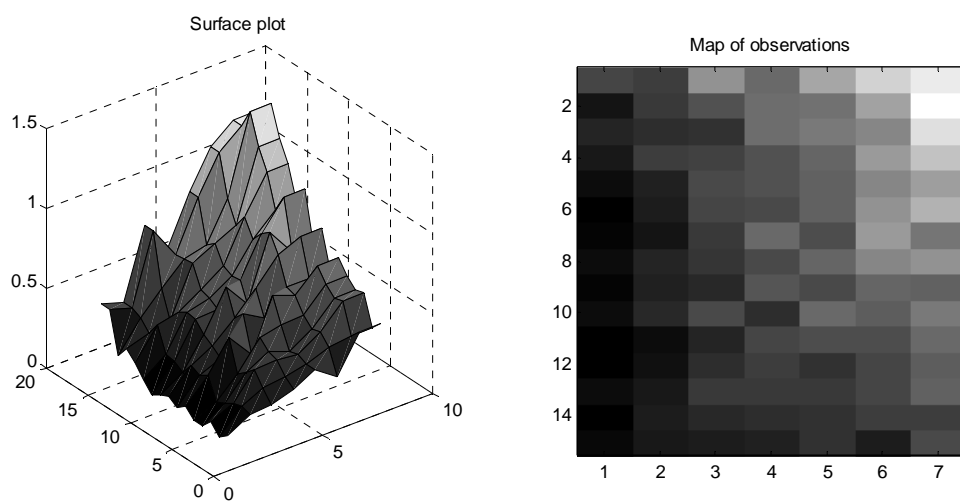


Figure 1. Surface plot & map of observations. Data sample 1 (Appendix 1.)

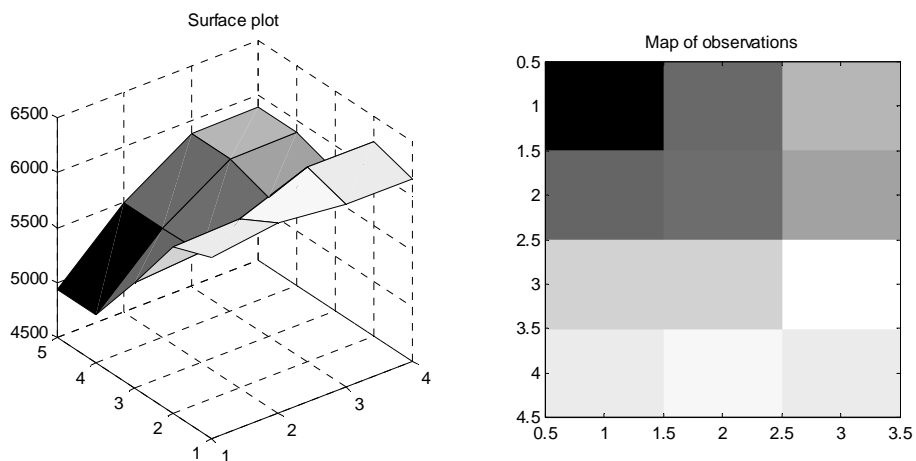


Figure 2. Surface plot & map of observations. Data sample 2 (Appendix 2.)

Figures 1 and 2 for our data samples show just the opposite. Both exhibit areas of high and low values, hills and valleys. Everything seems to be related to everything else, but “closer things more so” (Tobler, 1979).

## 2.2. Analysis of Variance ANOVA

In one-way analysis of variance data is categorized to multiple groups based on some variable. In analysis of variance every observation's variance is assumed to be constant. In addition, every observation is assumed to have the same expected value within the group. Notation: there are k groups. There are  $n_i$  independent observations in every group where i is group's index.  $y_{ji}$  is observation j in group i.  $j = 1, 2, \dots, n_i$  and  $i = 1, 2, \dots, k$ .

Assumptions:

$$E(y_{ji}) = \mu_i$$

$$D^2(y_{ji}) = \sigma^2$$

Group means:

$$\bar{y}_t = \frac{1}{n_t} \sum_{j=1}^{n_t} y_{jt}$$

Combined groups' mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ji} \text{ where } n = n_1 + n_2 + \dots + n_k$$

One-way analysis of variance is based on sum of squares: total sum of squares (SST), group sum of squares (SSG) and residual sum of squares (SSE).

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ji} - \bar{y})^2$$

$$SSG = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ji} - \bar{y}_i)^2$$

ANOVA is a test for equality of expected values. Formally

Null hypothesis  $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

Alternative hypothesis  $H_1: \text{There exist } a \neq b \text{ so that } \mu_a \neq \mu_b$

In our work this test should show that expected values are not the same because of different treatments.

The F-test variable is of the form

$$F = \frac{n - k}{k - 1} \cdot \frac{SSG}{SSE}$$

If the observations are normally distributed and the null hypothesis is valid, the variable F is F-distributed with degrees of freedom k-1 and n-k.

We ran one-way analysis of variance<sup>1</sup> to analyze data sample 2 (Appendix 2) and got the following results:

Source	SS	df	MS	F	Prob>F
Columns	52743.5	2	26371.8	0.14	0.8744
Error	1742349.5	9	193594.4		
Total	1795093	11			

<sup>1</sup> In Matlab:

```
>> X=[6152 60325464 5887; 6222 6267 5507 4935; 6162 6018 5768 5492];
>> anova1(X);
```

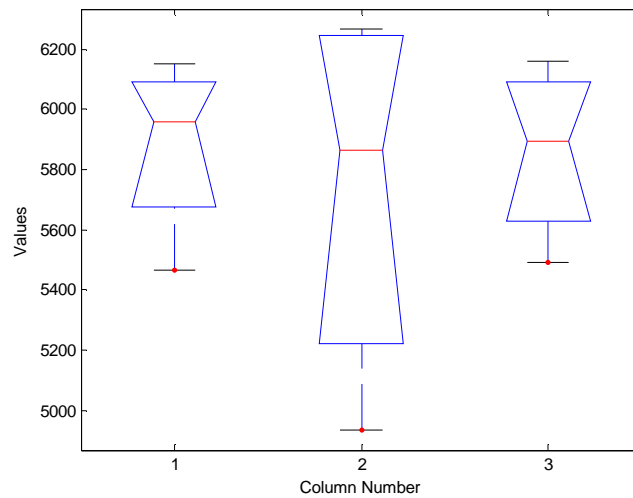


Figure 3. Boxplot of data set 2

Test's F-variable is 0.14 and P-value is 0.8744. Based on this test, the null hypothesis cannot be rejected and all the expected values are equal between groups. This is not satisfactory since different treatments should affect the results. Spatial dependence might cause this result.

### 2.3. Ordinary Least Squares Regression OLS

The method of ordinary least squares is used to achieve an optimal fit between the model and observed numerical data by adjusting the parameters of the model. This is done by minimizing the squared sum of residuals. (Pindyck, R, Rubinfeld, D, 1998)

The general model is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

with the following standard assumptions (Mellin, I. 2007):

- (1) The values  $x_{ij}$  of dependent variable  $x_j$  are fixed non-random constants,  $i=1, 2, \dots, n$ ,  $j=1, 2, \dots, k$
- (2) There are no linear dependencies between dependent variables
- (3)  $E(\varepsilon_i) = 0, i=1, 2, \dots, n$
- (4)  $\text{Var}(\varepsilon_i) = \sigma^2, i=1, 2, \dots, n$
- (5)  $\text{Cor}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
- (6)  $\varepsilon_i \sim N(0, \sigma^2), i=1, 2, \dots, n$

In our research variables  $x_{ij}$  are so-called indicator variables (or dummy variables), and have value equal to 1 if treatment  $j$  is used in observation  $y_i$  is a replication of treatment  $j$ , and 0 if it is not. Appendices 1 and 2 illustrate this for actual greenhouse (21 different treatments) and wheat field trials (three treatments) made in Kotkaniemi, Finland, in 2006.

OLS statistics for the results of the above mentioned field trials is shown in the table below. As only one treatment was used for one field plot, using all treatments as dependent variables leads to a highly multicollinear model. This happens when the model includes a constant. Then, one of the three dummy variables is linearly dependent on the rest:

$$x_{i3} = 1 - x_{i1} - x_{i2}$$

Thus, we have included only two of the three dummy variables to our standard OLS analysis.

Table 1. Standard OLS regression of data sample 2

Statistix 8.0		Kotkaniemi_Wheat_Fie..., 26.4.2008,			
12:36:59					
<b>Unweighted Least Squares Linear Regression of Y</b>					
<b>Predictor</b>					
<b>Variables</b>	<b>Coefficient</b>	<b>Std Error</b>	<b>T</b>	<b>P</b>	<b>VIF</b>
Constant	5860.00	219.997	26.64	0.0000	
X1	23.7500	311.122	0.08	0.9408	1.3
X2	-127.250	311.122	-0.41	0.6921	1.3
R-Squared	0.0294	Resid. Mean Square (MSE)			193594
Adjusted R-Squared	-0.1863	Standard Deviation			439.994
<b>Source</b>	<b>DF</b>	<b>SS</b>	<b>MS</b>	<b>F</b>	<b>P</b>
Regression	2	52744	26372	0.14	0.8744
Residual	9	1742350	193594		
Total	11	1795093			
Cases Included	12	Missing Cases	0		

None of the variables are statistically significant and the model is not statistically significant as a whole. Thus, it is clear that normal OLS regression is inadequate.

## 2.4. Regression Diagnostics

Regression diagnostics is important after the parameters of regression have been estimated. The goal of regression diagnostics is to test the assumptions of OLS-model and check that the model is correct and valid. (Laininen, 2000)

### 2.4.1. Residual Maps and Plots

Figures 4 and 5 illustrate residual analysis plots for data sets 1 and 2 respectively.



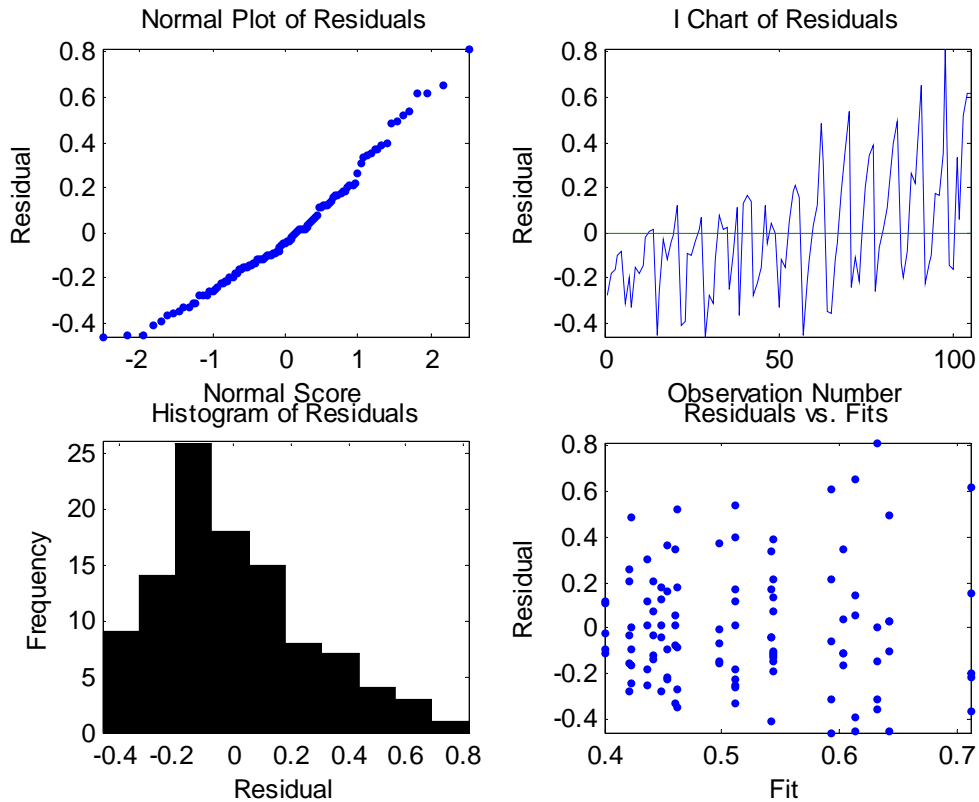


Figure 4. OLS Regression diagnostics for data sample 1 (105 observations)

Normal plot of residuals shows the residuals to be approximately normally distributed. The second chart plots observation number  $i$  against residual  $e_i$ . The chart reveals a clear trend: the closer we get to plot 105, the larger the residuals get and vice versa. Compare this to Figure 1.

Based on the histogram, the mean of the residuals is below zero and the distribution is skewed to the right. The model violates assumptions (3), (5) and (6) of the general linear regression model. The fit  $\hat{y}_i$  against residual  $e_i$  plot implies slight heteroskedasticity (as fit values get larger, residuals diverge more).

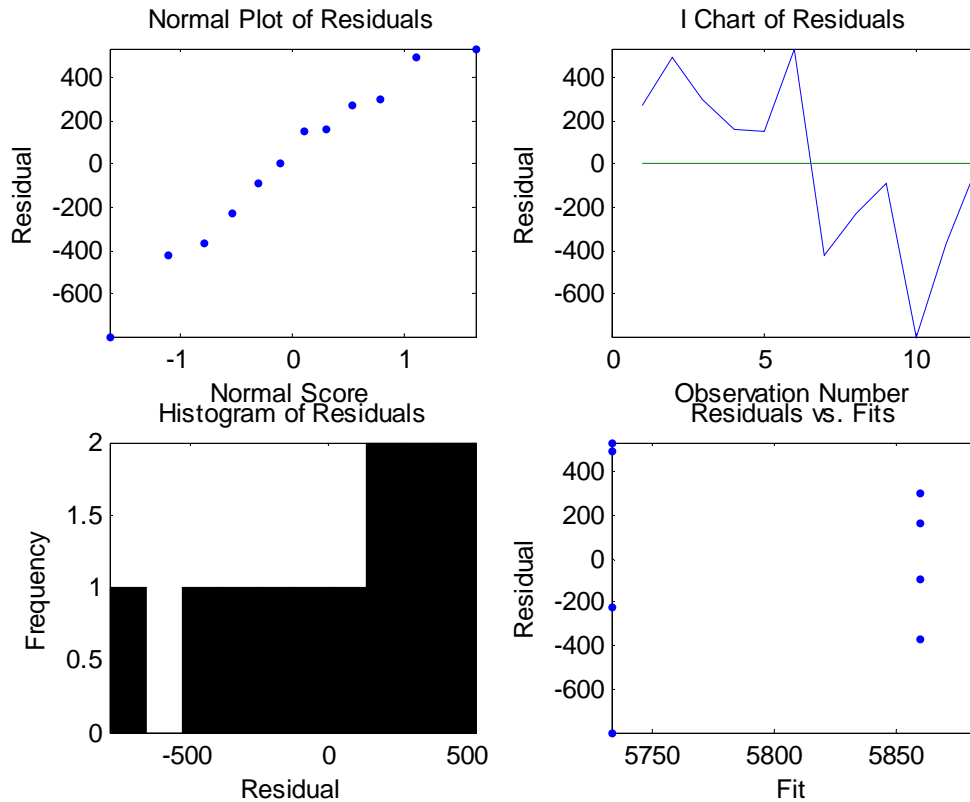


Figure 5. OLS regression diagnostics for data sample 2 (12 observations)

Appendix 4 exhibits more regression diagnostics including a studentized residual map and chart, Cook’s distance<sup>2</sup> and fit vs. observation plot.

### 2.4.2. Diagnostics for Spatial Autocorrelation

Standard assumption (5) guarantees that the residuals of a standard OLS-model do not correlate. Correlation of residuals causes coefficient estimators to be inefficient and wrong. A special form of correlation is autocorrelation where residuals are correlated in some fixed distance (Virtanen 2007). Durbin-Watson test can be used to test for autocorrelation of sequential residuals.

We are not interested in temporal autocorrelation but in spatial autocorrelation. Whereas autocorrelation is about proximity in time, spatial autocorrelation is about proximity in space. Spatial autocorrelation is more interesting since data is spatial. A spatial distribution is spatially autocorrelated “if there is any systematic pattern in the spatial distribution available” (Lembo, 2008).

One measure of the degree of spatial dependence in data is Moran’s I (Moran, 1948), the most commonly used statistic for global spatial autocorrelation.

Moran’s I is defined formally (Moran, 1950) as

<sup>2</sup> When Cook’s distance  $Di > 1$ , the observation  $i$  is a strong outlier candidate

$$I = \frac{1}{S_0 b_2} \sum_{i=1}^m \sum_{j=1}^m w_{ij} (x_i - \bar{x})(x_j - \bar{x})$$

Where  $x_i$  is the variable of interest in region  $i$  in study area  $A$ , which has  $m$  regions. The other terms are defined below.

$$S_0 = \sum_{i=1}^m \sum_{j=1}^m w_{ij}$$

$$\bar{x} = \sum_{i=1}^m \frac{x_i}{m}$$

$$b_k = \sum_{i=1}^m \frac{(x_i - \bar{x})^k}{m}$$

$w_{ij}$  is an element of a spatial weight matrix which measures spatial distance between regions  $i$  and  $j$ . In a tradition Queen or Rook contiguity definition  $w_{ij}$  is 1 if a region  $i$  is connected with  $j$ , and 0 otherwise. It is also possible to use weighting methods that are not binary and that assign non-zero weights to areas that are not directly connected but rather close to each other. In any case Moran's  $I$  statistic can range from -1 to 1. Statistic values close to 1 indicate either high-value clustering or low-value clustering, and values close to -1 indicate that low-values are located next to high-values. When statistic is close to zero values are spatially random. (Moran, 1950)

The pitfall of the Moran's  $I$  statistic is that it can detect only one dominant type of spatial autocorrelation. If high-value clustering and low-value clustering coexist, the method cannot distinguish them. (Tonglin, Lin, 2006). In addition, the method is not helpful in suggesting which alternative specification should be used (Anselin, 2005). To this end, our solution employs a spatial regression model decision rule (see Figure 6 in Section 4).

Despite these difficulties, Moran's  $I$  has considerable power in detecting misspecifications in the model.

### 3. Spatial Regression Model

Spatial econometrics has emerged as a new subfield of econometrics. It is a blanket term for statistical tests and models used to address potential issues introduced by the presence of spatial effects in regression analysis. These include:

- n spatial lag dependence
- n spatial error dependence
- n spatial heterogeneity

Traditional econometrics has largely ignored spatial dependence and heterogeneity, since they violate the traditional Gauss-Markov assumptions used in regression modeling. This gives rise to alternative estimation approaches.

First, we must somehow quantify the locational aspect of our sample data. This can be achieved by defining a spatial weight matrix which maps out the neighbourhood of each observation.

#### 3.1. Constructing Spatial Weights

Weight matrix construction is crucial when using Moran's I. The weight matrix defines the spatial autocorrelation to be analyzed. There are many ways to construct weight matrices that capture the notion of "connectiveness" between regions on the plot.

Figure 4 shows a hypothetical example of five regions as they would appear on a map<sup>3</sup>.

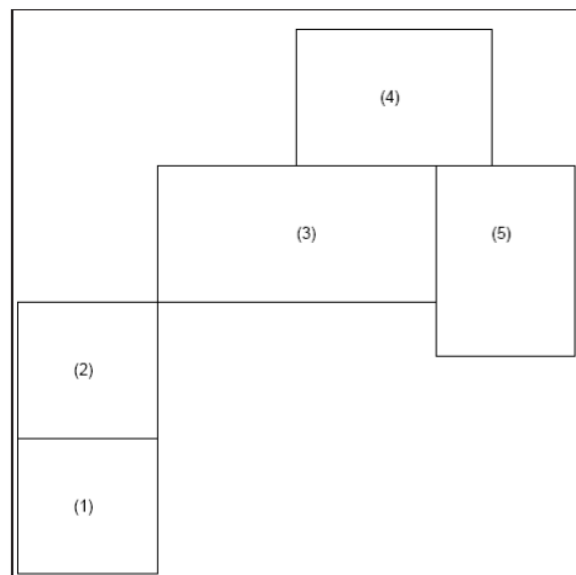


Figure 4. An illustration of contiguity (LeSage, 1998)

We wish to construct a 5 by 5 binary matrix  $W$  containing 25 elements taking values of 0 or 1 capturing the neighbourhood structure of every region on the map. We record in each row of the matrix  $W$  a set of contiguity relations associated with one of the five regions.

<sup>3</sup> This section is largely based on LeSage (1998).

There are many ways to accomplish the task. Below, some of the alternative ways are enumerated.

*Rook contiguity:* Define  $W_{ij} = 1$  for regions that share a common side with the region of interest. For row 1, reflecting region 1's relations we would have  $W_{12} = 1$  with all other row elements equal to zero. As another example, row 3 would record  $W_{34} = 1; W_{35} = 1$  and all other row elements equal to zero.

*Bishop contiguity:* Define  $W_{ij} = 1$  for entities that share a common vertex with the region of interest. For region 2 we would have  $W_{21} = 1$  and all other row elements equal to zero.

*Queen contiguity:* For entities that share a common side or vertex with the region of interest define  $W_{ij} = 1$ . For region 3 we would have:  $W_{32} = 1; W_{34} = 1; W_{35} = 1$  and all other row elements zero.

The matrix  $W$  reflecting first-order rook's contiguity relations for the five regions in Figure 4 is:

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

A transformation often used is to convert the matrix  $W$  to have row-sums of unity. This is referred to as a 'standardized' contiguity matrix  $C$ :

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

If we now multiply  $C$  by a vector of observations on some variable  $y$  associated with the five regions, we get the spatially lagged variable  $Cy$ , used in the SAR model:

$$\begin{pmatrix} y_1^* \\ y_2^* \\ y_3^* \\ y_4^* \\ y_5^* \end{pmatrix} = Cy = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} y_2 \\ y_1 \\ 0.5y_4 + 0.5y_5 \\ 0.5y_3 + 0.5y_5 \\ 0.5y_3 + 0.5y_4 \end{pmatrix}$$

This is one way of quantifying the notion that  $y_i = f(y_j), j \neq i$ .

Cressie (1993) differentiates spatial data structures between point patterns, geostatistical data, and lattice data. Since this study considers data in lattices, the corresponding weight matrices are contiguity-based. In contiguity-based spatial weights, neighbors are defined as cells who share a boundary. In distance-based spatial weights the construction is based on distance between cells, not neighbors. (Anselin, 2005)

Four weight matrices are taken into discussion in this section. They are row, column, Queen and Rook contiguity weight matrices. Let (gray cell) be a neighboring cell to cell 5. In the table on the left all the cells that share a common side or vertex with cell five are neighbors to it (Queen). On the right only cells that share the same column (2 and 8) or row (4 and 6) are neighbors to cell five (Rook). All neighbors get the same weight and non-neighbors get weight

of zero. The resulting weight matrix can be standardized so that the sums of row weights equal 1.

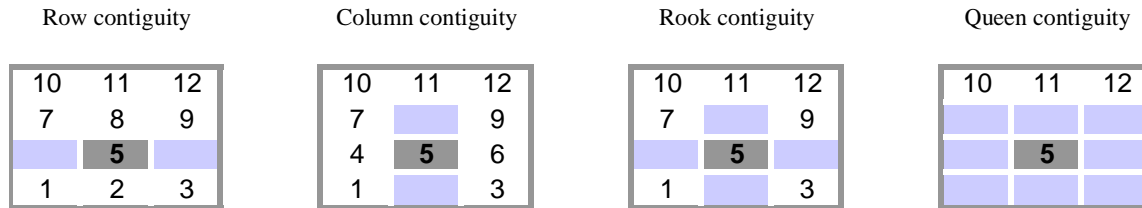


Figure 5. Implemented contiguity weight matrices

We could also use a weight matrix where all the cells are noted when analyzing spatial autocorrelation. Cells that are not neighbors would get a smaller weight than neighboring cells. In this work we do not consider complex weight matrices since our data matrices contain few elements –only the closest neighbors are important.

### 3.2. Spatial Lag Model –SAR Spatial Autoregressive Model

This section considers the estimation by means of maximum likelihood of a spatial regression model that includes a spatially lagged dependent variable. Formally,

$$\begin{aligned}
 \mathbf{y} &= \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \boldsymbol{\varepsilon} &\sim N(0, \sigma^2 \mathbf{I}_n)
 \end{aligned}
 \tag{I}$$

where  $\mathbf{y}$  is a  $N$  by 1 vector of observations on the dependent variable,  $\mathbf{W}\mathbf{y}$  is the corresponding spatially lagged dependent variable for weights matrix  $\mathbf{W}$ ,  $\mathbf{X}$  is a  $N$  by  $K$  matrix of observations on the explanatory variables,  $\boldsymbol{\varepsilon}$  is a  $N$  by 1 vector of error terms,  $\rho$  is the spatial autoregressive parameter, and  $\boldsymbol{\beta}$  is a  $K$  by 1 vector of regression coefficients.

$(\mathbf{W}\mathbf{y})_i$  is always correlated with error term  $\varepsilon_i$ , and thus, an OLS estimator is inconsistent with the model. Thus, maximum likelihood estimators or instrumental variables must be used. Normal OLS regression results are biased and inconsistent when data is spatial. (Anselin, Bera, 1998. p. 246)

Maximum likelihood estimation of this model is based on a concentrated likelihood function. For details on the estimation, the reader is referred to LeSage (1998).

### 3.3. Spatial Error Model - SEM

Spatial Error Model can be defined as,

$$\begin{aligned}
 \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\
 \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \\
 \boldsymbol{\varepsilon} &\sim N(0, \sigma^2 \mathbf{I}_n)
 \end{aligned}
 \tag{II}$$

where the parameter  $\lambda$  is the spatial autoregressive coefficient for the error lag  $\mathbf{W}\mathbf{u}$ , and  $\boldsymbol{\varepsilon}$  is an uncorrelated and homoskedastic error term. The model is implemented in the Econometrics toolbox by LeSage (1998).

Results of applying SEM for data set 1 and 2 can be found in Appendices 10 and 12.

## 4. Our Approach

The variety of test statistics for spatial autocorrelation is great. Since models (I) and (II) both rely heavily on the weight matrix  $W$ , one obvious action is to consider the spatial regression results for different spatial weights. Consequently we run both SAR and SEM with four different spatial weights, namely, the row, column, Rook and Queen contiguity weight matrices.

Figure 6 illustrates a simplified spatial regression model selection decision rule based on Anselin (2005).

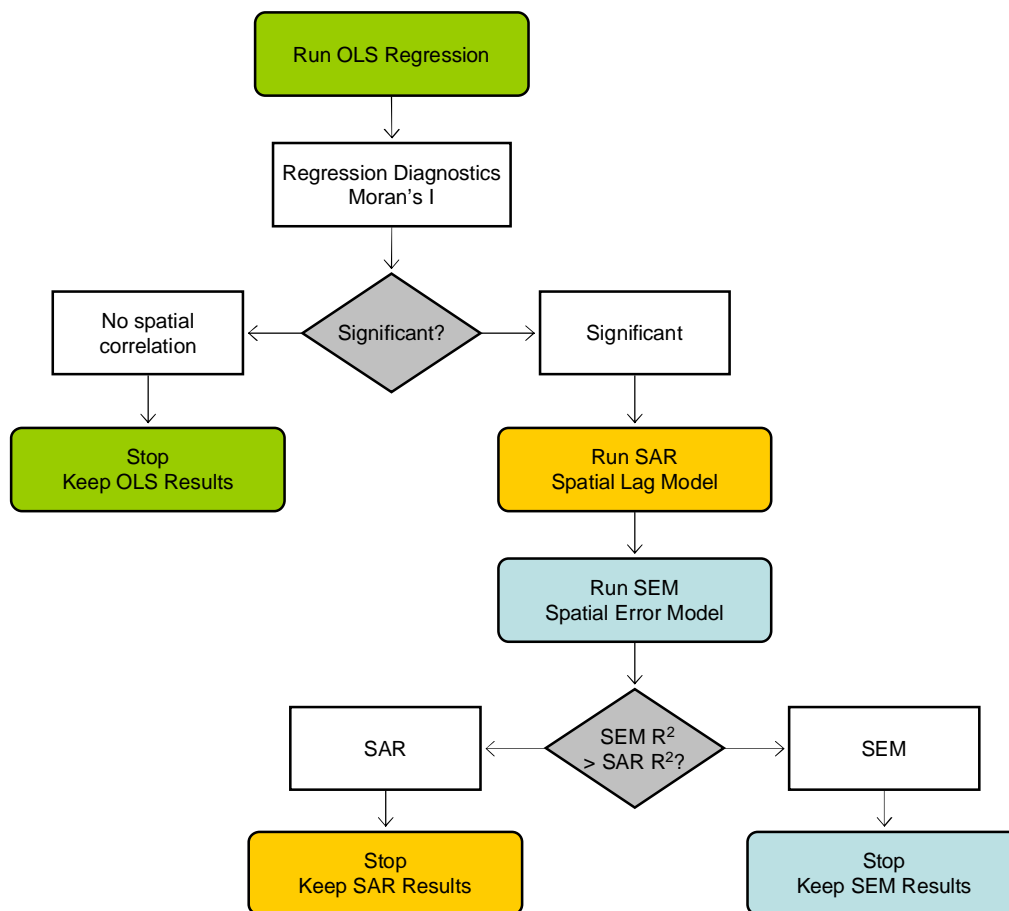


Figure 6. Spatial Regression Decision Process

The process begins at the top of the graph and considers the standard OLS statistics. Moran's I test statistic is calculated with each of the four spatial weights. If one of the I test statistics rejects the null hypothesis of no spatial autocorrelation in residuals, both SAR and SEM are run with each of the weights. The results are compared so as to maximize R-squared. The process ends and reports results of the model that best explains variation in observations

### 4.1. MATLAB Implementation

We used MATLAB and LeSage's Econometrics toolbox (LeSage, 1999) to implement our solution. MATLAB is a high-level language and interactive environment that enables one to perform computationally intensive tasks faster than with traditional programming languages

such as C, C++, and Fortran. The Econometrics toolbox adds over 350 functions to MATLAB's standard set and includes the following models

- n Spatial Autoregressive Model (SAR)-- ML and GMM estimators
- n Spatial Error Model (SEM)-- ML and GMM estimators
- n General Spatial Model (SAC)-- ML and GMM estimators
- n Spatial Durbin Models (SDM)
- n Spatial Error Probit Models (SEMP)
- n SAR model with Fixed Effects (Panel Data)
- n SEM model with Fixed Effects (Panel Data)
- n Bayesian Geographically Weighted Regressions (BGWR)
- n Casetti's Spatial Expansion Model (DARP)

The Spatial Autoregressive (SAR) and Error (SEM) models are considered in this report.

The Econometrics toolbox offers a function, `moran()`, that computes Moran's I-statistic for spatial correlation in the residuals of a regression model. It takes as its parameters a vector of dependent variable observations, matrix of independent variables and a contiguity weight matrix. We implemented a function, `createw()`, that constructs row, column, Queen and Rook contiguity weight matrices for lattice data (see: Appendix 5).

In addition, the toolbox includes functions `ols()`, `diagnose()` and `dfbeta()` that perform ordinary least squares regression, compute regression diagnostic measures, and measures omitting each observation sequentially respectively. Appendix 6 lists 'help ols', 'help diagnose' and 'help dfbeta'.

Functions `sar()` and `sem()` compute spatial autoregressive and spatial error model estimates. Both take three parameters: vector of dependent variable observations, matrix of independent variables and a standardized contiguity weight matrix. Appendices 7 and 8 list 'help sar' and 'help sem'.

The main function of our solution, `fixspat()` (Appendix 9), implements the model selection decision process in Figure 6. The function estimates an OLS regression, calculates Moran's I, and then runs SAR and SEM regression with each weight matrix in turn. `fixspat()` outputs regression diagnostics in two windows, Moran's I test statistic and the best spatial regression model in a separate window.



## 5. Results

The result windows of Moran's I and best model selection for `fixspat('data15x7.xls', false)` are given below.

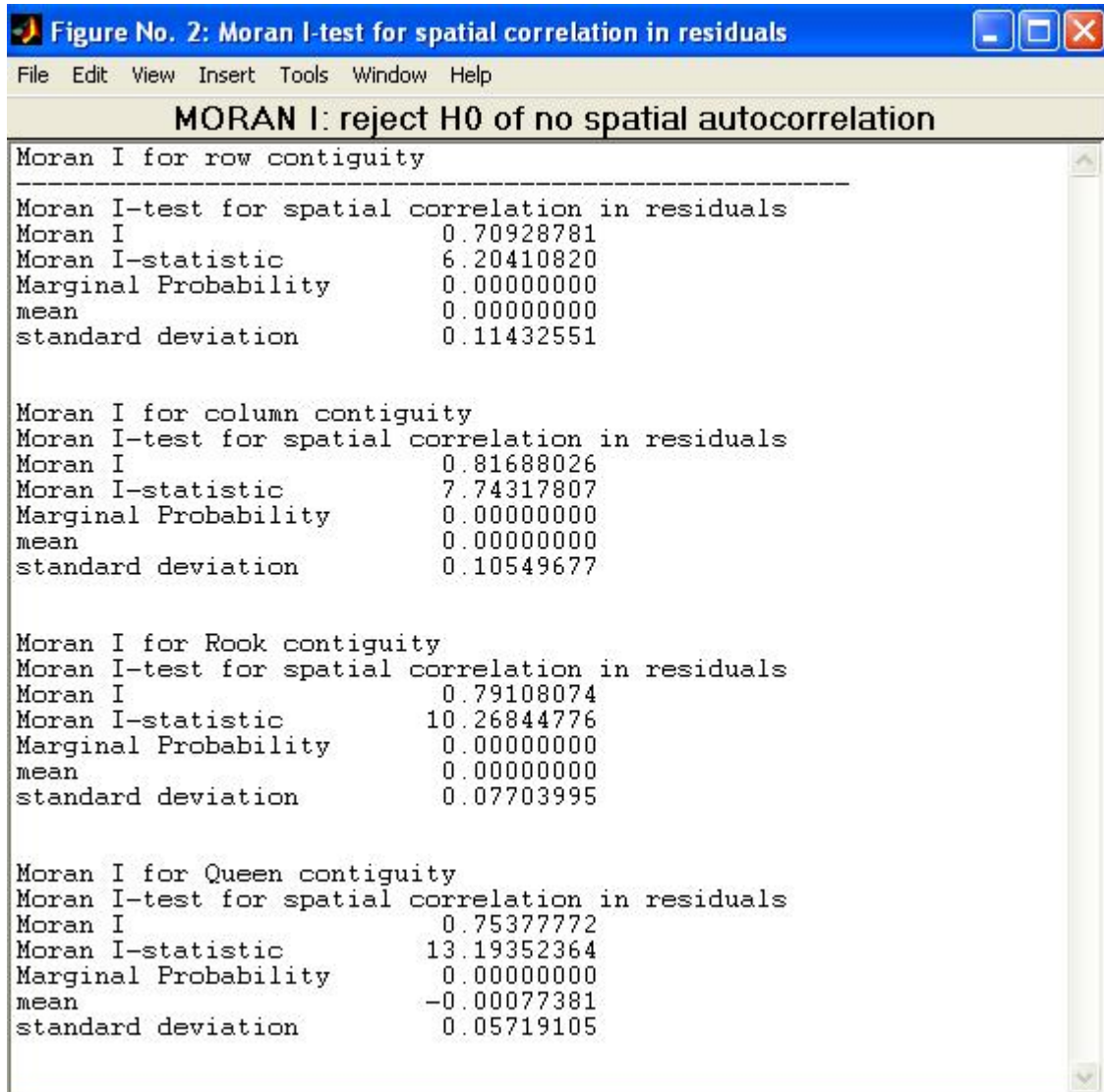


Figure 7. Moran I-test for spatial correlation in residuals. Data set 1

Since maximum I-statistic equals  $13.19 > 1.96$  and marginal probability  $< 0.05$ ,  $H_0$  is rejected. Data set 1 exhibits spatial autocorrelation.

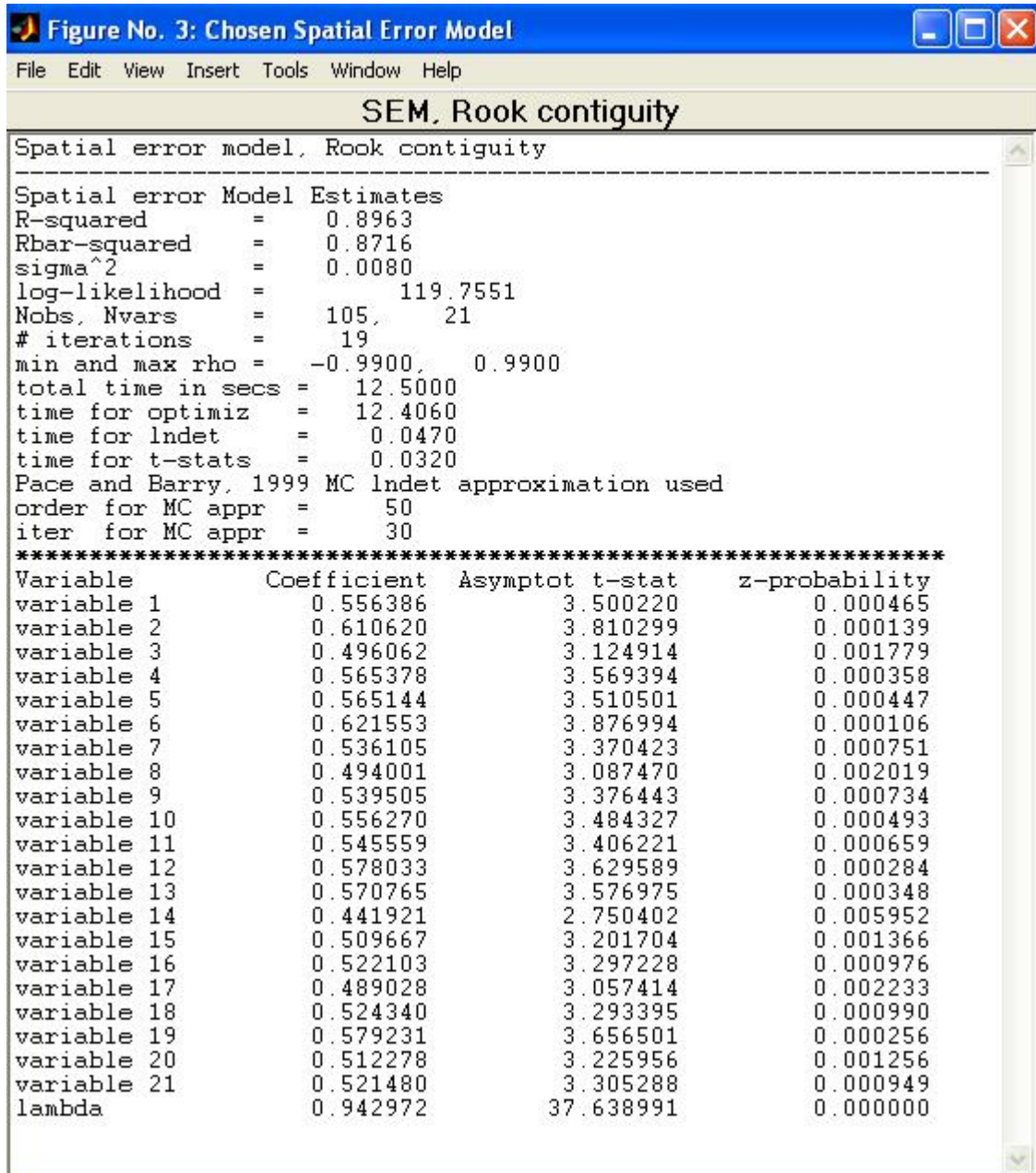


Figure 8. Best spatial regression model for data set 1 is SEM with Rook contiguity

In this case, the best spatial regression model is the spatial error model with Rook contiguity. The model greatly improves R-squared from 9.25% for OLS to 89.63% for SEM. Compare other results with standard OLS regression (Appendix 5).

Appendices 6 and 7 show results for data set 2.

## 6. Discussion

As mentioned in section 4, there are many ways of selecting a spatial regression model. Here, we considered four different contiguity regimes and two different spatial regression models and simply compared results of every combination. To analyze the resulting regression model, we reproduced some of the plots shown before for SEM with Rook contiguity. The plots are illustrated below.

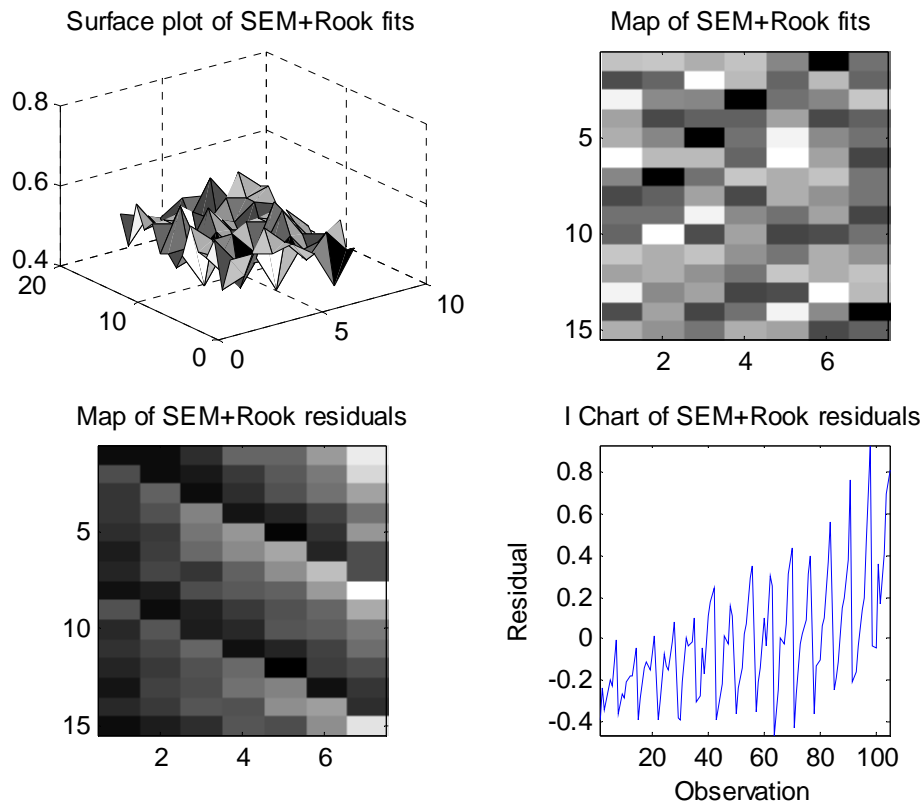


Figure 9. SEM regression diagnostics: fit and residual plots for data set 1. (SEM with Rook contiguity)

Fits for each treatment equal the regression coefficient for that treatment indicator variable. Because of this, the surface plot of fits has less high hills and low valleys. Residuals of the model still exhibit heteroskedasticity. Despite of this, SEM does a better job than standard OLS in finding better values for regression coefficients evidenced by its higher R-squared value, and thus provides better estimates of treatment means. For comparison of regression coefficients between OLS and SEM, see Appendices 5 and 7.

## 7. References

- Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Anselin, L. and Bera, A. 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullah, A. and Giles, D. E., editors, *Handbook of Applied Economic Statistics*, pages 237-289. Marcel Dekker, NY.
- Anselin, L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27:93-115.
- Anselin, L. 2003. *GeoDa 0.9 User's Guide*. Spatial Analysis Laboratory (SAL), <https://www.geoda.uiuc.edu/documentation/manuals>, 30.04.2008
- Anselin, L. 2005. *Exploring Spatial Data with GeoDa: A Workbook*. Spatial Analysis Laboratory (SAL). Department of Geography, University of Illinois, Urbana-Champaign, IL.
- Smirnov, O. and Anselin, L. 2001. Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, 35:301-319.
- Lembo A., 2008, *Spatial Autocorrelation, Lecture 9 notes*, University of Cornell, [www.cornell.edu/academics/docs/Courses\\_of\\_Study\\_0708.pdf](http://www.cornell.edu/academics/docs/Courses_of_Study_0708.pdf), 30.04.2008
- LeSage, J.P. 1998. *Spatial Econometrics*. Department of Economics, University of Toledo.
- LeSage, J.P. 1999. *Applied Econometrics using MATLAB*. Department of Economics, University of Toledo.
- LeSage, J.P. and Pace, R. K. 2004. *Arc\_Mat, a Matlab toolbox for using ArcView Shape files for spatial econometrics and statistics*.
- Heikkinen, J. 2004. *Spatiaalinen tilastotiede. Matematiikan ja tilastotieteen laitos, Helsingin yliopisto*. <http://www.rni.helsinki.fi/~jmh/ss04/>, 30.04.2008
- Kemira GrowHow. 2007. Annual Report. Available: <http://www.yara.fi>, 30.04.2008
- Laininen, P. 2000. *Tilastollisen analyysin perusteet*. Otatieto, Helsinki.
- Cressie, N. 1993. *Statistics for Spatial Data*. Wiley, New York.
- Mellin, I. 2007. Time series and forecasting. Lecture notes, <http://www.sal.tkk.fi/Opinnot/Mat-2.3128/>, 30.04.2008
- Moran, P.A.P. 1948. The interpretation of statistical maps. *J. Roy. Statist. Soc. Ser. B* 10, 243–251.
- Moran, P.A.P. 1950. A Test for the Serial Independence of Residuals. *Biometrika* 1950-06-01 pp. 178-181
- Pindyck, R. and Rubinfeld, D. 1998. *Econometric Models and Economic Forecasts*, Fourth Edition, McGraw & Hill, New York
- Tonglin, Z. and Lin, G. 2007. A decomposition of Moran's I for clustering detection. *Computational Statistics & Data Analysis* 51 (2007) 6123 –6137
- Virtanen K., 2007, *Tilastollisen analyysin perusteet, lecture slides*, <http://www.sal.tkk.fi/Opinnot/Mat-2.2104/>, 30.04.2008


## Appendix 1. Indexing and Data sets 1 and 2

**Data set 1:** greenhouse trials. 15x7 latice (Appendix 2)

Number of observations: 105

Number of treatments: 21

Number of replicates: 5 (A-E)

 heated plots

17C	16C	15C	14C	18E	19E	20E
13C	12C	11C	10C	15E	16E	17E
9C	8C	7C	6C	12E	13E	14E
5C	4C	3C	2C	9E	10E	11E
1C	0C	20B	19B	6E	7E	8E
18B	17B	16B	15B	3E	4E	5E
14B	13B	12B	11B	0E	1E	2E
10B	9B	8B	7B	18D	19D	20D
6B	5B	4B	3B	15D	16D	17D
2B	1B	0B	20A	12D	13D	14D
19A	18A	17A	16A	9D	10D	11D
15A	14A	13A	12A	6D	7D	8D
11A	10A	9A	8A	3D	4D	5D
7A	6A	5A	4A	0D	1D	2D
3A	2A	1A	0A	18C	19C	20C

**Data set 2:** wheat field trials. 4x3 latice (Appendix 3)

Number of observations: 12

Number of treatments: 3

Number of replicates: 4 (A-D)

2D	3D	1D	Block 4
1C	2C	3C	Block 3
3B	1B	2B	Block 2
1A	2A	3A	Block 1

### Indexing\*

	Data set 1			Data set 2			
15 x 7	Col 1	...	Col 7	Col 1	Col 2	Col 3	4 x 3
Row 1	99	...	105	10	11	12	Row 1
:				7	8	9	Row 2
:	:		:	4	5	6	Row 3
Row 15	1	...	7	Plot 1	2	3	Row 4

\*rows and columns according to Matlab matrix notation



## Appendix 3. Data set 2: 4x3 matrix (wheat field trials)

R	C	PLOT	TREAT	REPLICATE	YIELD	TREAT1	TREAT2	TREAT3
4	3	1	1	A	6152	1	0	0
		2	2	A	6222	0	1	0
		3	3	A	6162	0	0	1
		4	3	B	6018	0	0	1
		5	1	B	6032	1	0	0
		6	2	B	6267	0	1	0
		7	1	C	5464	1	0	0
		8	2	C	5507	0	1	0
		9	3	C	5768	0	0	1
		10	2	D	4935	0	1	0
		11	3	D	5492	0	0	1
		12	1	D	5887	1	0	0

Appendix 4. Regression diagnostics continued. Data sets 1 and 2

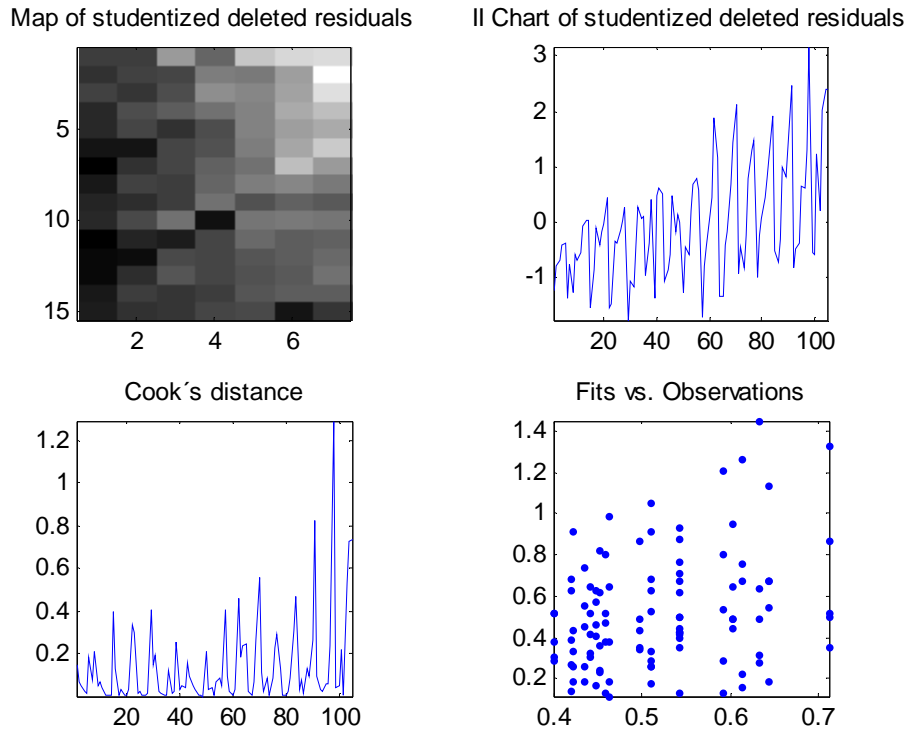


Figure 10. Regression diagnostics for data sample 1, continued.

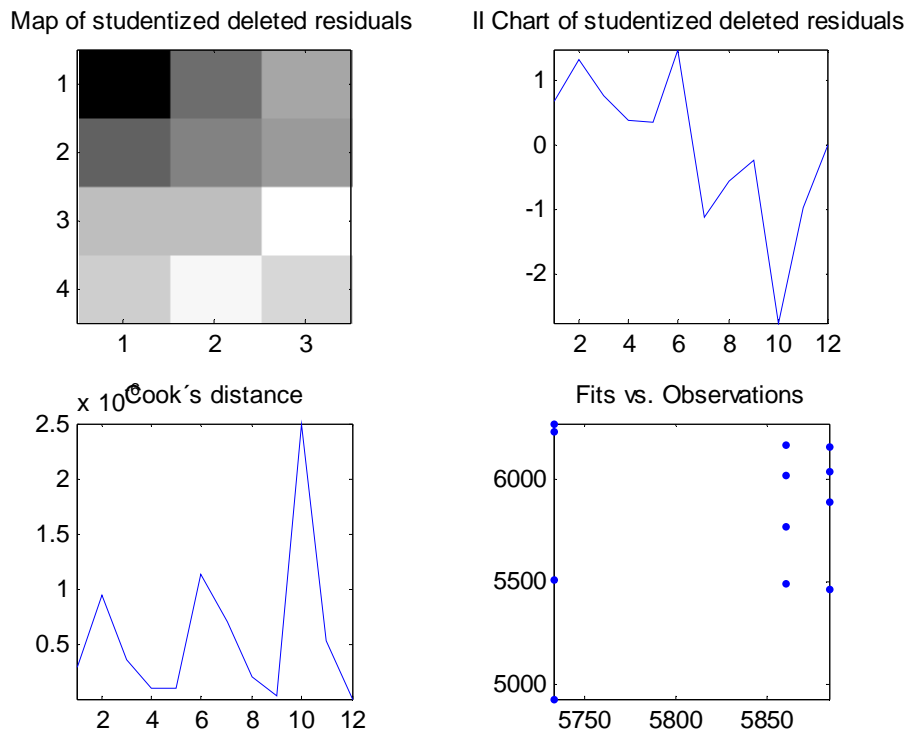


Figure 11. Regression diagnostics for data sample 2, continued.



Appendix 5. Results of SEM and OLS regression for data set 1

```

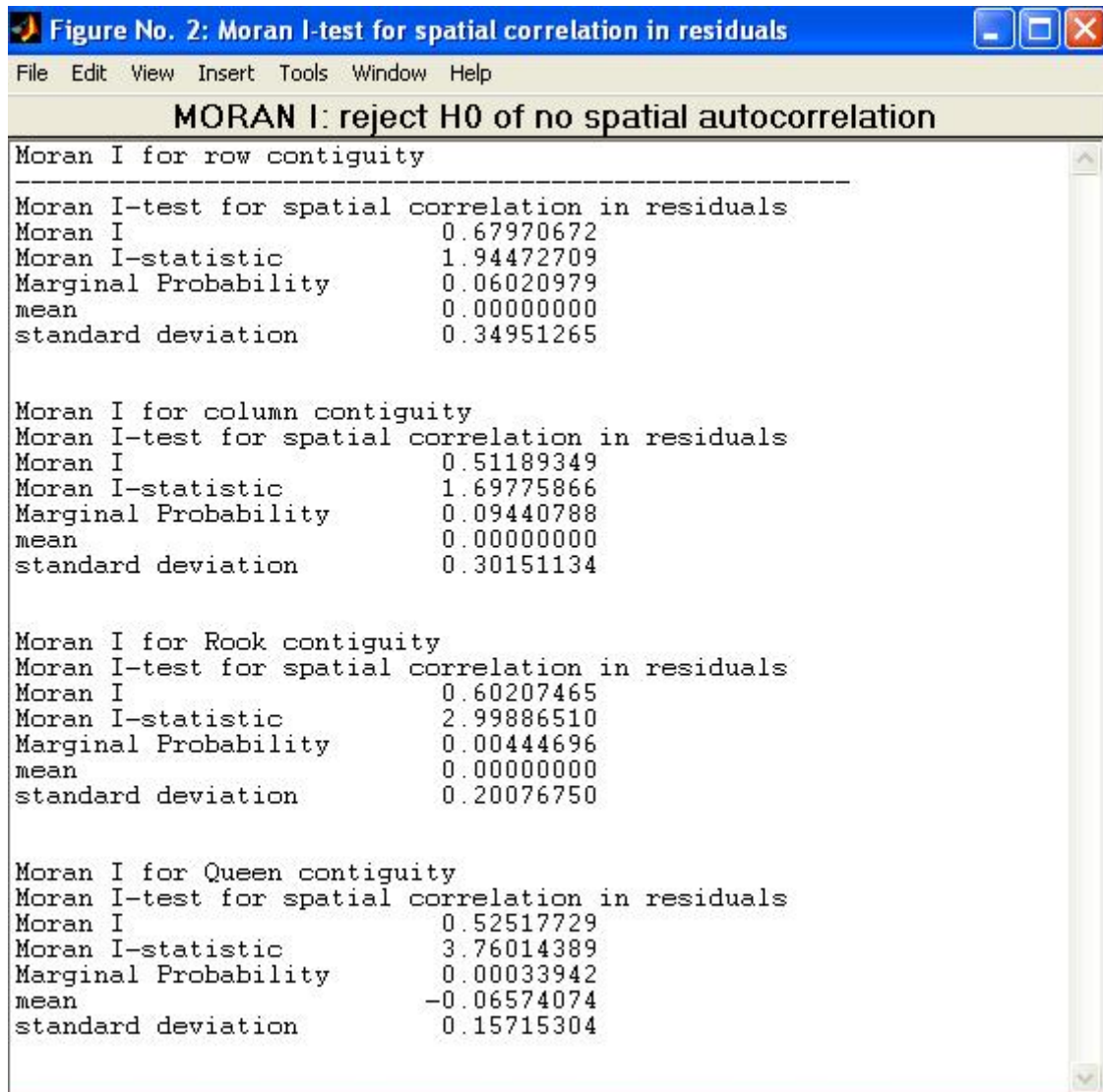
Spatial error Model Estimates
R-squared = 0.8971
Rbar-squared = 0.8726
sigma^2 = 0.0079
log-likelihood = 120.63674
Nobs, Nvars = 105, 21
# iterations = 18
min and max rho = -0.9900, 0.9900
total time in secs = 0.2030
time for optimiz = 0.0630
time for lndet = 0.0470
time for t-stats = 0.0620
Pace and Barry, 1999 MC lndet approximation used
order for MC appr = 50
iter for MC appr = 30
*****

Variable      Coefficient      Asymptot t-stat      z-probability
variable 1    0.559551          3.245282          0.001173
variable 2    0.614160          3.536752          0.000405
variable 3    0.499019          2.897536          0.003761
variable 4    0.568699          3.308500          0.000938
variable 5    0.568590          3.260692          0.001111
variable 6    0.625020          3.597295          0.000322
variable 7    0.539333          3.125088          0.001778
variable 8    0.497256          2.866604          0.004149
variable 9    0.542548          3.131433          0.001740
variable 10   0.559211          3.229725          0.001239
variable 11   0.548452          3.158865          0.001584
variable 12   0.580679          3.361620          0.000775
variable 13   0.573227          3.312342          0.000925
variable 14   0.444000          2.550461          0.010758
variable 15   0.511740          2.964304          0.003034
variable 16   0.524514          3.052066          0.002273
variable 17   0.490936          2.831876          0.004628
variable 18   0.526158          3.048052          0.002303
variable 19   0.582044          3.386391          0.000708
variable 20   0.514620          2.987826          0.002810
variable 21   0.523246          3.056056          0.002243
lambda       0.947959          40.428592          0.000000

Ordinary Least-squares Estimates
R-squared = 0.0925
Rbar-squared = -0.1235
sigma^2 = 0.0875
Durbin-Watson = 1.1005
Nobs, Nvars = 105, 21
*****

Variable      Coefficient      t-statistic      t-probability
variable 1    0.400000          3.023432          0.003314
variable 2    0.423333          3.199798          0.001941
variable 3    0.436333          3.298060          0.001429
variable 4    0.449067          3.394306          0.001052
variable 5    0.497667          3.761653          0.000311
variable 6    0.511000          3.862434          0.000220
variable 7    0.421333          3.184681          0.002034
variable 8    0.459733          3.474931          0.000811
variable 9    0.543000          4.104308          0.000094
variable 10   0.442333          3.343411          0.001238
variable 11   0.510667          3.859914          0.000222
variable 12   0.642667          4.857647          0.000005
variable 13   0.543000          4.104308          0.000094
variable 14   0.453000          3.424036          0.000956
variable 15   0.612800          4.631897          0.000013
variable 16   0.542533          4.100781          0.000095
variable 17   0.602733          4.555808          0.000018
variable 18   0.632167          4.778282          0.000007
variable 19   0.462667          3.497103          0.000754
variable 20   0.592267          4.476694          0.000024
variable 21   0.710933          5.373646          0.000001
    
```

## Appendix 6. Moran's I test for data set 2



## Appendix 7. Results of SEM and OLS regression for data set 2

### Spatial error Model Estimates

```

R-squared      =    0.6660
Rbar-squared   =    0.5918
sigma^2        = 49959.9079
log-likelihood =    -79.6954
Nobs, Nvars    =    12,    3
# iterations   =    20
min and max rho =  -0.9900,  0.9900
total time in secs =  0.1400
time for optimiz =  0.1090
time for lndet  =  0.0160
Pace and Barry, 1999 MC lndet approximation used
order for MC appr =  50
iter for MC appr =  30
    
```

```

*****
Variable      Coefficient  Asymptot t-stat    z-probability
variable 1    5835.947256      15.594030  0.000000
variable 2    5759.748041      15.268620  0.000000
variable 3    5780.306827      15.338746  0.000000
lambda        0.825937      7.102903   0.000000
    
```

### Ordinary Least-squares Estimates

```

R-squared      =    0.0294
Rbar-squared   =   -0.1863
sigma^2        = 193594.3889
Durbin-Watson =    1.1708
Nobs, Nvars    =    12,    3
    
```

```

*****
Variable      Coefficient    t-statistic    t-probability
variable 1    5883.750000      26.744706     0.000000
variable 2    5732.750000      26.058332     0.000000
variable 3    5860.000000      26.636750     0.000000
    
```