SPARD

Spatial Analysis of Rural Development Measures
Contract No. 244944

Work Package No. 4

August 2011

D4.1

Manual for the tests of spatial econometric model

Authors: Vincent Linderhof, Peter Nowicki, Eveline van Leeuwen, Stijn Reinhard and Martijn Smit

Document status

Internal use	x
Confidential use	x
Draft No. 2	31-7-2011
Final	29-3-2013
Submitted for internal review	22-3-2013

COOPERATION

Table of contents

Figures

Abbreviations

CMEF	Common Monitoring and Evaluation Framework
EAFRD	European Agricultural Fund for Rural Development
EU	European Union
LDI	Logical Diagram of Impact
NUTS	French abbreviation for *nomenclature des unités territoriales statistiques*, a geocode standard for referencing the subdivisions of EU countries for statistical purposes
RD	Rural Development
RDP	Rural Development Plan

Summary

The SPARD project aims at developing tools to analyze to what extent EU rural development measures impact a number of economic, social and environmental objectives that they are designed to target. This report describes the econometric test to select the appropriate spatial econometric model to estimate the effect of RDP measures on their intended effect (the impact indicators). First the method to identify relations between a dependent and potential explanatory variables is presented. Econometric test are given to select the relevant variables for the model. This model is tested whether spatial correlation is present or not and if so the type of correlation. Finally the econometric specification is tested using standard tests.

This is document is intended to present the general approach within SPARD to specify, estimate and test spatial econometric models. It will be used in WP4 and WP5, also by less econometrically experienced researchers. For them this document can be used as a manual how to perform spatial econometrics in the context of SPARD. This SPARD document 4.1 is a living document to enable adaptations to the SPARD methodology as it will develop during the project. This living document can then be used as a manual throughout the project.

1 Introduction

1.1 Objective of WP4.1

The SPARD project aims at developing tools to analyze to what extent EU rural development measures impact a number of economic, social and environmental objectives that they are designed to target. One important obstacle to the proposed spatial econometric analysis is data availability. This is due to two aspects: The first obstacle applies to all impact assessment problems, the difficulty to construct a counterfactual situation (what would have happened without the policy). The second obstacle is related to the Common Monitoring and Evaluation Framework (CMEF), which SPARD is supposed to base its analyses on.

The CMEF is a relatively new instrument and still under development. Following types of indicators are included: baseline indicators (objective- and context-related), input indicators (expenditures), output (physical), result (physical and successful) and impact. Baseline indicators describe the socio-economic, environmental and farm structure related situation of a region, while the other indicators are related to budget, implementation and impact of rural development measures. There are still many data gaps and the data delivered by the authorities in the member states has not been sufficiently checked yet. In addition, the indicators gathered by the framework refer to different spatial units. Baseline indicators, for example, are available at NUTS2 level, while input, output, result and impact indicators are measured at the programming level. Input, output, and result indicators are available for the single RDP measures, while impact indicators measure the outcome of an entire program (consisting of a number of RDP measures).

In SPARD we enable policy analysis to look at causal relationships between characteristics, needs, expenditures and results of rural development measures in a spatial dimension. We analyse to what extent a spatial econometric approach will be useful to provide information on the effect of the RDP measures, and whether the aim reflected by the impact indicators will be reached. In WP4 Task 4.1 is the definition of the econometric test to assess the impact of RDPs. This follows from the work in WP2 to select relevant variables and the work in WP3 on the design of logical diagrams and the identification of relations that have to be tested (the identification of causal relationships). Task 4.2 proceeds with an analysis of the database for spatial patterns. This is followed by Task 4.3, which is the identification and estimation of the model at NUTS0 level. In order to prepare for the case study analyses in WP5, the next step is Task 4.4, which is the specification of the model to be used at the NUTS2 and NUTS3 levels.

Task 4.5 brings together the knowledge gained in WP4 through a description of a general methodology for the use of spatial econometrics in Rural Development Programmes.

This report describes the analytical framework used by SPARD. Based on the available literature and the expertise of the SPARD researchers, the theoretical assumptions followed by SPARD are outlined. Secondly, the expected impacts of EU rural development measures are derived both from previous studies and the available literature. Thirdly, under consideration of the available data from the CMEF, the theoretical assumptions and expected impacts are operationalized for three EU rural development measures, namely modernization of agricultural holdings (121), agri-environment measures (214) and diversification into non-agricultural activities (311). These measures were selected to begin the analysis with. Step-wise the analysis will be extended to other measures.

The spatial econometric analysis will be built upon ex-post analysis, i.e. mainly based on the input, output and result indicators provided by the RDPs themselves and the baseline indicators if available. The objective of the spatial econometric analysis is to estimate to what extent the measured values for the impact indicators can be ascribed to the RDP measure being examined.

This econometric analysis starts with a (theoretical) model that describes the causal relationships. We build upon the SPARD 3.1 Report (Report on analytical framework – conceptual model, data sources, and implications for spatial econometric modeling). The spatial scale of the tool will be both NUTS0 and NUTS2. The principal one will be the scale of RD programming. In some Member States it is the National scale, in others Federal States and for certain RDP measures also the regional scale. To set up the model applicable for the regional scale is crucial, since this will provide insight into how spatial heterogeneity within a country affects the impact of an RDP measure. Moreover, in many countries it is at the regional level that the RDPs are planned and managed. However, the impact indicators need to be aggregated to the national (NUTS0) level as well, so that the member state can assess the overall effectiveness of its RDP. Lower spatial scales (lower than NUTS2) will be used for validation of the model in the case studies. These data will be collected based on information available on local RDPs (and their evaluations).

1.2 Objective of this document

The main objective of this document is to support the spatial econometric analysis in WP4 and WP5. In the latter work package, the spatial econometric model has to be developed for the specific case study areas. To stimulate a standard methodology over WP4 and these case studies and to support researchers that have any experience in econometrics, but not yet up-to-date knowledge of spatial econometrics, this document can be regarded as a manual for the spatial econometric analyses within the SPARD project. It will be a living document to enable updates in the future to capture the development of the SPARD methodology during the project. Being a living document it can truly be a manual for WP4 and WP5 researchers.

Prerequisites for using this document:

- Basic Stata knowledge and experience.
- Basic econometric knowledge (at least decent knowledge of OLS)

This report will describe the estimation procedure for all SPARD models regardless the aggregation level of the analyses and the dependent variable.

1.3 General methodology of spatial econometric analysis

Spatial econometric analysis for a RDP measure can be roughly divided in six steps:

a. Select the RDP measure and its relevant indicator for the assessment. Note that the RDP measure might have more than one impact and result indicator;
b. Check economic theory with respect to the measure (and indicator)
c. Specifying deterministic relationships (conceptual framework)
d. Identifying dependent and independent variables and a functional form
e. Testing the variables and relations
f. Estimation of the model.

The conceptual framework of the spatial econometric analysis is given in SPARD document 3.1 (Uthes et al., 2011). CMEF is the basis for our analysis, describing roughly the relation between a RDP measure and the intended effect; the baseline variables. These baseline variables are for instance growth in labour productivity, increase in gross value added in agriculture etcetera. In WP4 we analyse econometric literature for an economic theory and corresponding (spatially) econometric models. These truly econometric models are linked to the conceptual models presented in WP3. In paragraph 2.1 procedure for the selection of variables is presented. The first two steps are

1.4 Logical Diagrams of Impact (LDI) and measures

The Logical Diagrams of Impact (LDI) present the relevant deterministic relationships per measure. They depict all factors that affect the base line indicator(s), including of course the measure itself, the result and the output indicators. For measure 121 and 311 the LDI is presented. These measures will be elaborated upon in WP4 to demonstrate the possibilities of spatial econometrics. The econometric literature and the LDI provide suitable variables and functional form for the spatial analysis. Given the availability of suitable data the model can be estimated. The next paragraph describes the procedure to come to the best empirical model.

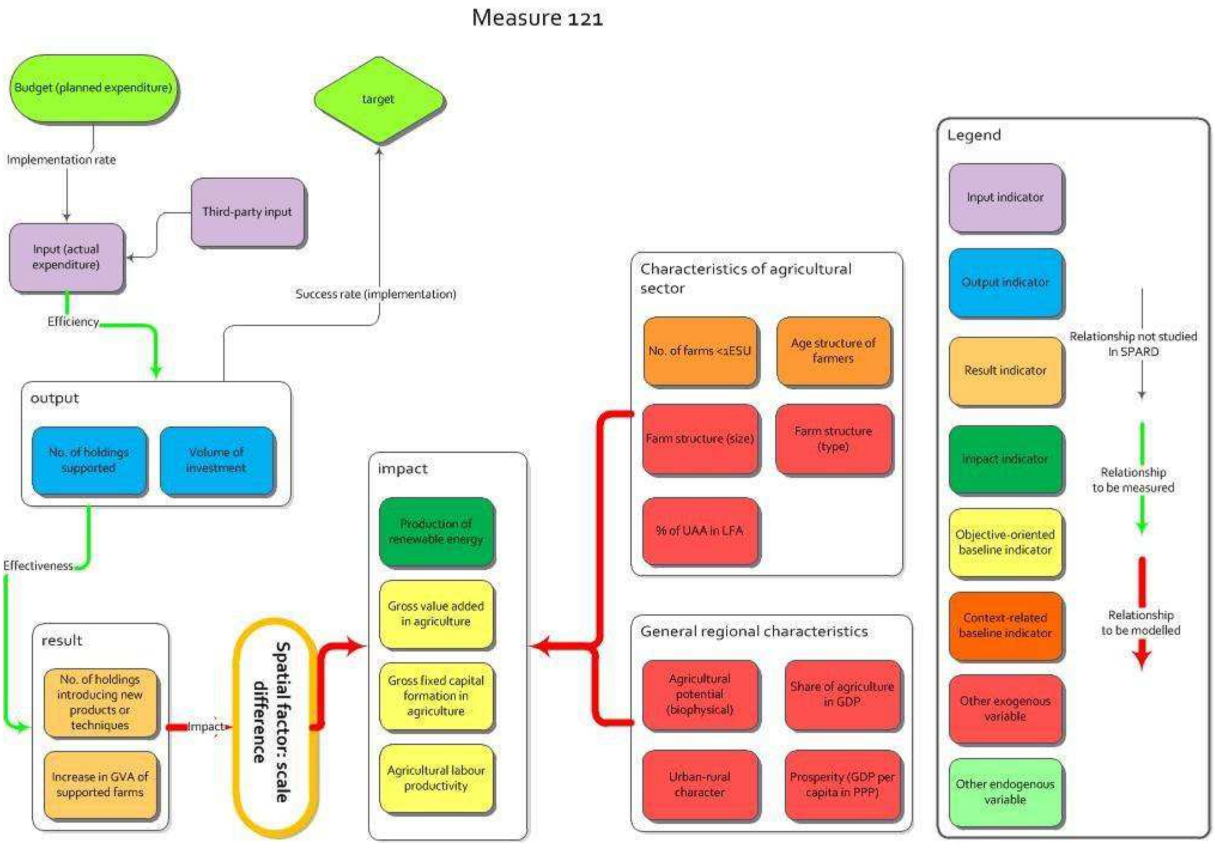

Measure 121
Budget (planned expenditure)
target
Implementation rate
Third-party input
Input (actual expenditure)
Success rate (implementation)
Efficiency
output
No. of holdings supported
Volume of investment
Effectiveness
result
No. of holdings introducing new products or techniques
Increase in GVA of supported farms
Impact
Spatial factor: scale difference
impact
Production of renewable energy
Gross value added in agriculture
Gross fixed capital formation in agriculture
Agricultural labour productivity
Characteristics of agricultural sector
No. of farms <1ESU
Age structure of farmers
Farm structure (size)
Farm structure (type)
% of UAA in LFA
General regional characteristics
Agricultural potential (biophysical)
Share of agriculture in GDP
Urban-rural character
Prosperity (GDP per capita in PPP)
Legend
Input indicator
Output indicator
Result indicator
Impact indicator
Objective-oriented baseline indicator
Context-related baseline indicator
Other exogenous variable
Other endogenous variable
Relationship not studied in SPARD
Relationship to be measured
Relationship to be modelled

Figure 1 Logical Diagramme of Impact for measure 121 modernisation

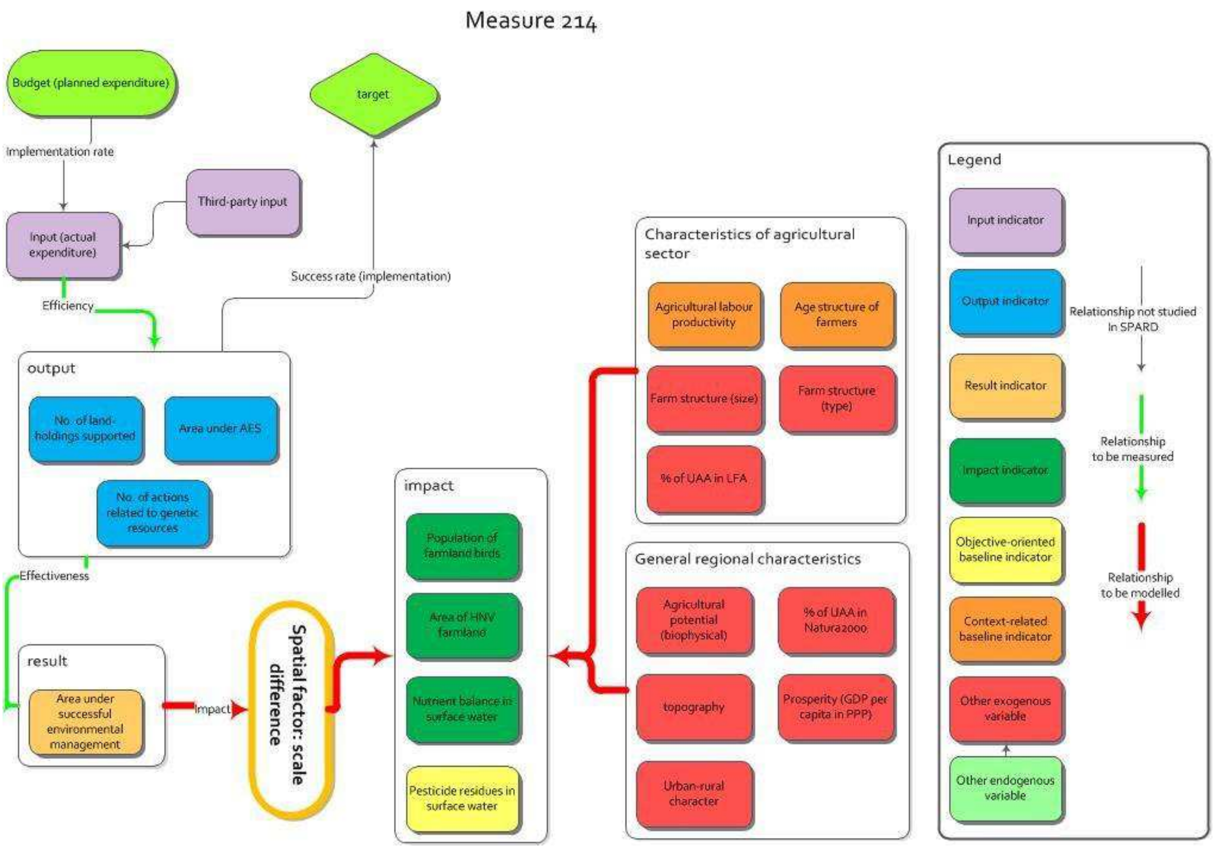

Measure 214
Budget (planned expenditure)
target
Implementation rate
Third-party input
Input (actual expenditure)
Success rate (implementation)
Efficiency
output
No. of land-holdings supported
Area under AES
No. of actions related to genetic resources
Effectiveness
result
Area under successful environmental management
impact
Spatial factor: scale difference
impact
Population of farmland birds
Area of HNV farmland
Nutrient balance in surface water
Pesticide residues in surface water
Characteristics of agricultural sector
Agricultural labour productivity
Age structure of farmers
Farm structure (size)
Farm structure (type)
% of UAA in LFA
General regional characteristics
Agricultural potential (biophysical)
% of UAA in Natura2000
topography
Prosperity (GDP per capita in PPP)
Urban-rural character
Legend
Input indicator
Output indicator
Result indicator
Impact indicator
Objective-oriented baseline indicator
Context-related baseline indicator
Other exogenous variable
Other endogenous variable
Relationship not studied in SPARD
Relationship to be measured
Relationship to be modelled

Figure 2 Logical Diagramme of Impact for measure 214

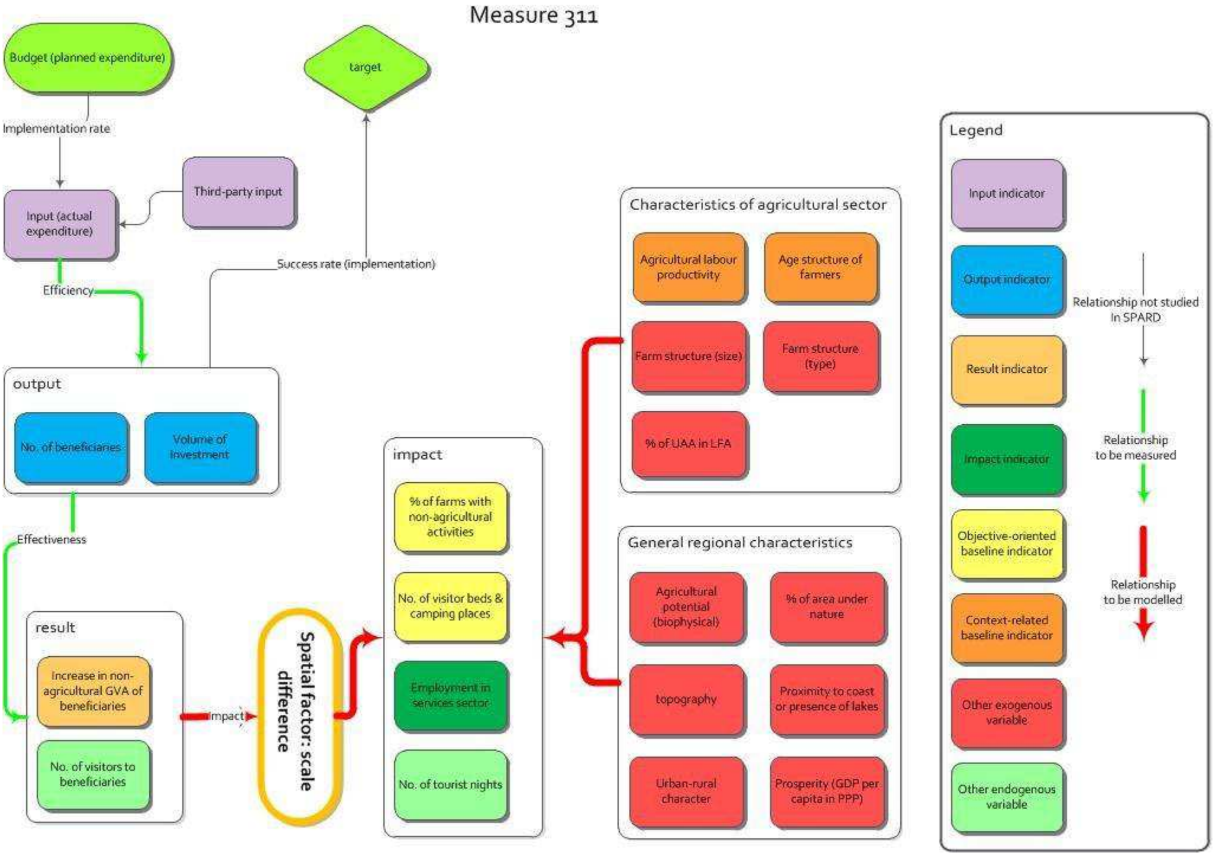

Measure 311
Budget (planned expenditure)
target
Implementation rate
Third-party input
Input (actual expenditure)
Success rate (implementation)
Efficiency
output
No. of beneficiaries
Volume of Investment
Effectiveness
result
Increase in non-agricultural GVA of beneficiaries
No. of visitors to beneficiaries
impact
Spatial factor: scale difference
impact
% of farms with non-agricultural activities
No. of visitor beds & camping places
Employment in services sector
No. of tourist nights
Characteristics of agricultural sector
Agricultural labour productivity
Age structure of farmers
Farm structure (size)
Farm structure (type)
% of UAA in LFA
General regional characteristics
Agricultural potential (biophysical)
% of area under nature
topography
Proximity to coast or presence of lakes
Urban-rural character
Prosperity (GDP per capita in PPP)
Legend
Input indicator
Output indicator
Result indicator
Impact indicator
Objective-oriented baseline indicator
Context-related baseline indicator
Other exogenous variable
Other endogenous variable
Relationship not studied in SPARD
Relationship to be measured
Relationship to be modelled

Figure 3 Logical Diagramme of Impact for measure 311

1.5 Selection of variables, functional form and tests

The empirical investigation provides estimates of unknown parameters in the model and often attempts to measure the validity of the propositions against the behaviour of observable data. The next sections describe a number of techniques used in this context.

1.6 Outline of the report

The selection of the dependent and independent variables is given in chapter 2, where also tests are presented to obtain the best specification based on the available data. In chapter 3 the spatial aspects are elaborated and tests are provided to select the best spatial functional form. Tests for the econometric specification are given in chapter 4.

The report is conceptualized as a 'living document' with possible changes during the operation of SPARD in order to keep this central document up to date with progress in the data availability and results from the spatial econometric analysis. It will also be updated based on the feedback from WP5.

The software used to illustrate the tests is Stata (see the Stata on line help and Cameron and Trivedi, 2010) and GeoDa (see the GeoDa website and Anselin, 2005). GeoDa is a free software package that conducts spatial data analysis, geovisualization, spatial autocorrelation and spatial modeling. With GeoDa comes a free workbook entitled *Exploring Spatial Data with GeoDa: A Workbook* (Anselin, 2005)

2 Selection procedure of variables

2.1 Selection of variables based on the economic theory

The selection and preparation of the data can be quite labour intensive. Therefore, this chapter proposes a procedure with criteria how to select the relevant variables for the spatial and data analysis. The procedure of selecting variables for the econometric analysis is specific for each RDP measure. First of all, each measure has its own impact and result indicator as indicated by the Logical Diagrams of Impact (see Figure 1 to Figure 3 in paragraph 1.4). Secondly, each RDP measure is affected by different developments. Finally, the impact of the RDP measures differs across the regions.

The composition of a database of the gross list of variables requires a number of stages:

1. Select the RDP measure for the assessment. Note that the RDP measure might have more than one impact and result indicator;
2. Check the economic literature for relevant indicators that might have an impact on the impact indicator of the RDP measure to be assessed (see stage 1). In this stage, one can also consider the relevance of time- or space lagged variables to be included;
3. Check the availability of data (OECD; CMEF database; Metabase; Cambridge Econometrics, for instance, for NUTS 2 or 3 levels; or other databases for NUTS 5 level). Take into account the opportunities to construct spatially and time-lagged variables;
4. Compare the variables required from a theoretical perspective (stage 2) and the data available (stage 3). Note that the inclusion of similar variables is advocated at this stage. The selection of variables to be included in the actual regression equations will be discussed in the next paragraphs and in chapter 3 on the spatial data analysis. Identify the omitted variables as well;
5. Compose the Stata database and GeoDa database

In the next paragraph, we will check the correlation coefficients between dependent and independent variables.

2.2 Check for correlation with dependent variable

Econometric estimation techniques rely on the correlation between the independent and dependent variables. The list of dependent and independent variables resulting from the selection procedure in paragraph 2.1 might be quite long. The spatial data analysis and preparation of the data for the econometric regression analysis might be labour intensive. One can consider to exclude the variables that do not, or hardly, correlate to any of the dependent variables from the database in order to save time in the spatial data analysis or the econometric regression analysis. Variables that are important from a theoretical perspective should be included in the model with respect to the interpretation of the results. Preferably, an independent variable correlates to dependent variables but does not correlate to other independent variables.

2.3 Tests for correlations between independent and dependent variables.

Stata command for running correlation matrix

```
correlate depvar1 depvar2 var1 var1b var2 var3
```

Or

```
pwcorr depvar1 depvar2 var1 var1b var2 var3, sig
```

The `sig` option provides the statistical significant levels of the correlation coefficients. For more information see Cameron and Trivedi (2010, p. 86) or Stata help on `correlate` or `pwcorr`.

The resulting correlation matrix C of the variables is a diagonal matrix. This means that element c_{ij} of matrix C is equal to c_{ji}. The maximum value of the elements is 1, and the minimum value is -1. Variables with values of correlation coefficients close to zero have no correlation. We propose the following rules of thumb for the selection of variables to be included in the econometric analysis:

Rule 1: An independent variable that has no or insignificant correlation coefficient with a dependent variable can be considered for exclusion for further analysis. If the independent variable is included because of theoretical reasons, one could maintain it for further analysis.

Rule 2: An independent variable that has a significant correlation coefficient with a dependent variables will be included for further analysis.

In addition to Rule 2, we propose to check the correlation coefficients between independent variables that have significant correlation coefficients with the dependent variable. An additional rule has to be considered: If independent variables are close to perfectly collinear (statistically significant correlation coefficients), then numerical instability (of the estimated parameter) may cause problems. The parameters may be estimated very imprecisely.

One can consider to include the temporally and/or spatially lagged variables in this checking procedure as well.

3 Spatial analyses of the data

3.1 Choice of weight matrix

An important difference between spatial and traditional (a-spatial) statistics is that spatial statistics integrate space and spatial relationships directly into their mathematics. Consequently, the conceptualization of spatial relationships prior to analysis is very important (Anselin et al., 2008). Weight matrices are a necessity when studying the relationships between regions. Whereas for relationships over time the distance in time can be measured in different quantities (days, weeks, years) – but these are always related to each other – distance in space is less clear. Is the distance measured from border to border, or from centre to centre, in a straight line or following transport lines? Do distances across other regions or across water bodies also count?

Weight matrices are used to model the spatial relation between observations. Binary weight matrices contain information for every 'region A'-'region B' combination whether they are to be considered neighbours or not (0 or 1). This means that it is assumed that spatial autocorrelation in the region under study only occurs between nearest neighbouring spatial units, whatever is their size and shape. Alternatively, weight matrices made up of weights representing various types of spatial connections can be used to represent the nuances of spatial associations in real-world circumstances, trying to solve the problem of topological invariance (Getis, 2009; Harris et al., 2011). In such cases, a weight matrix generally consists of weights between 0 and 1 for every A-B combination; those weights then sum to 1 by row and/or column. However, for the Exploratory Spatial Data Analysis a binary approach is most appropriate.

Four types of binary weight matrices are commonly used, namely nearest neighbour distance cut-off, rook contiguity and queen contiguity, and they are offered by the free GeoDa software. However, not all of these four types are equally useful.

Nearest neighbours

This analysis renders a robust type of matrix, as it always assigns neighbours to a region, whether they actually share borders or not. The number of neighbours is the same for all regions, and it is identified by a number *k*. Depending on the size and number of regions, settings vary; 10 is tractable in the NUTS2 setting. The robustness of this matrix lies in the fact that islands pose no problems. However, a disadvantage is that distances between ‘neighbours’ can vary widely across the map (e.g. North Sweden vs. the Netherlands). If regions are of equal size, this is the most convenient choice.

Distance cut-off

A distance cut-off works in a way similar to the nearest neighbours approach, except that here all regions within a certain distance range are considered neighbours. Some regions that are far off (Cyprus, Azores, Iceland) may end up without neighbours, which often leads to problems in software for spatial analyses. If population densities and travel times are homogenous across all regions, this is a very realistic choice, but islands can create problems.

Rook and Queen contiguity

Pure contiguity matrices are the most basic concept: whoever touches your region is considered a neighbour. This renders islands neighbourless, and therefore some models will not work with this type, including LISA analyses (see Anselin, 2005: 140). Rook contiguity differs from Queen contiguity in that corner contacts are not counted in rook contiguity. However, in a European context these are rare anyway, although they do occur in the United States and Africa. These are the most commonly used types of weight matrix outside LISA analyses. Yet the fact that the shape of regions decides which regions are neighbours can lead to strange results if two regions share a narrow border but otherwise extend away from each other.

3.1 ESDA analysis: Life-long learning example

The Exploratory Spatial Data Analysis is a first step to check whether spatial patterns exist, or, in other words, whether high and low values are suspiciously sorted in space. We show how the procedure works with two examples using the freely available tool GeoDa (http://geodacenter.asu.edu/). The analysis is done for one specific year, and uses a weight matrix of choice; in this case, we use k-10 nearest neighbours.

In a panel setting, the analysis can be repeated for all available years. However, if spatial patterns exist for one year, that is already enough to merit the inclusion of spatial econometrics in a model. Performing an analysis for the first and last available year can however be relevant in order to visually estimate whether spatial concentration increases or decreases over time.

We choose two variables which will be used as dependents in the analyses performed within the project. However, LISA (Local Indicators of Spatial Autocorrelation) can and should also be used to investigate independent variables to look for possible sources of bias (see Anselin, 2005:140). Finally, it can be extremely useful to perform a LISA analysis on the residuals from a regression, to see whether any spatial pattern has remained undetected by the variables already in the regression.

Life-long learning is important for human capital formation in an area, which is important both for the technological level, productivity and innovativeness of current activities, and for the attractiveness of a location for new activities. Moreover, the variable may have an impact on the participation rate in new training initiatives offered through or with the help of European funding.

We measure life-long learning as the percentage of the labour force (i.e., persons aged 25-64) participating in education and training. Among the European NUTS-2 regions in 2009, this percentage ranges from 0.5% to over 35% (see figure below). Especially Denmark, Norway, Sweden, the Netherlands, and the UK stand out in a positive sense; France and Spain show a mixture of higher and lower values, with the two highest values in France close to other high-scoring regions in Spain and Germany.

The map in *Figure 4* shows the descriptives. Regions outside the EU are not displayed in *Figure 4*. The Canaries and French overseas territories are left out for mapping regions, because spatial econometrics make no sense for these outlying regions at NUTS2 level. The map in Figure 5 then shows the actual clusters, defined by Local Indicators of Spatial Autocorrelation (LISA): we see clusters of high values grouped together (e.g., the UK; High-High means high values where the neighbours have high values too), and likewise for low values (e.g., Poland); and we see which regions have a low value in or next to a cluster of high values (e.g., North-western France; high-low). In this case, these regions would be areas where life-long learning is taken up very well in the vicinity, but the region itself lags. Theoretically, high ‘spikes’ in an area of low values could also exist, and they possibly do at a lower spatial scale if we would distinguish cities from their hinterlands.

The conclusion from these maps is that spatial patterns exist, and that it might be meaningful to perform a spatial analysis on them to see whether there are actual spill-overs, i.e. influences, between regions.

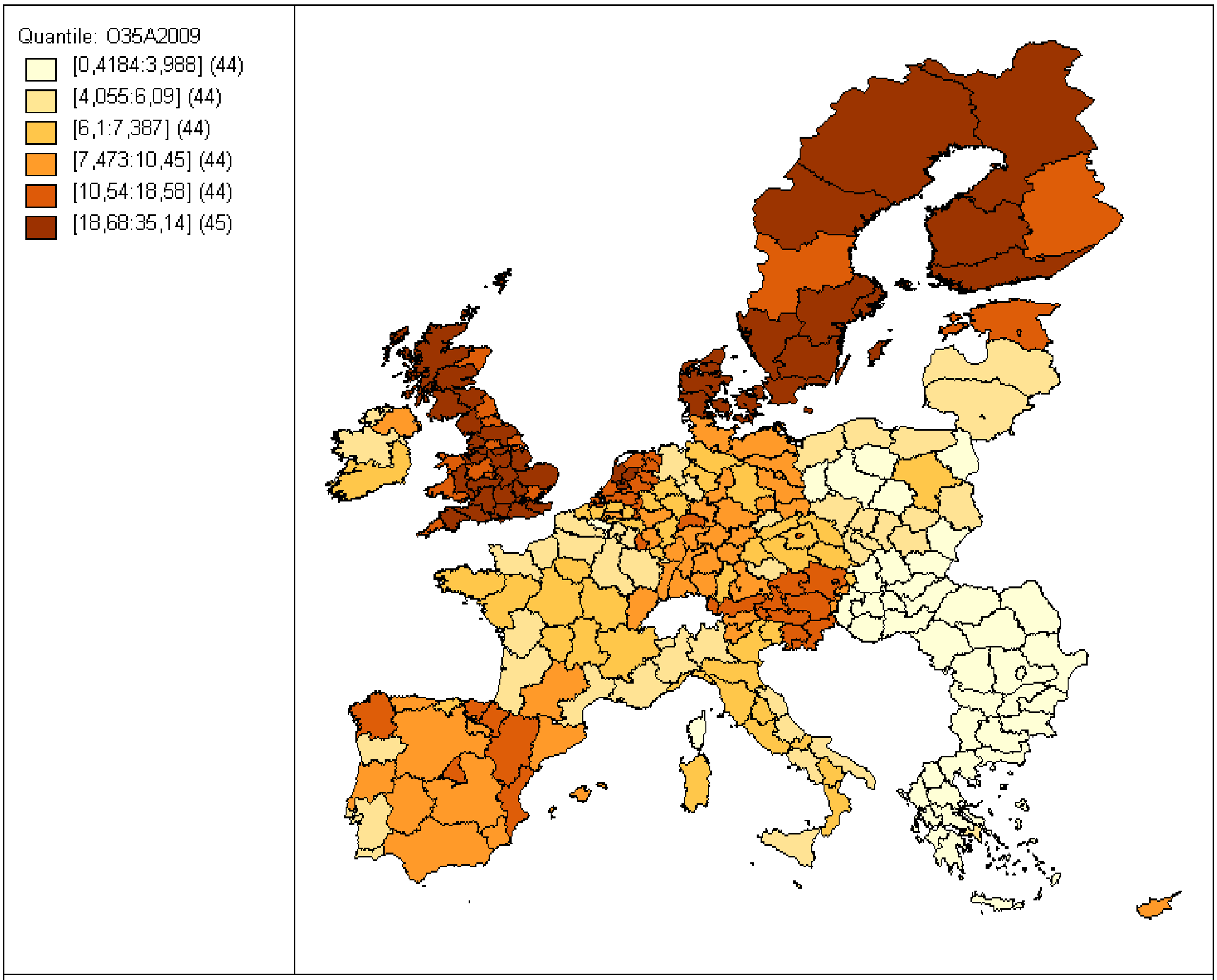


Figure 4 Life-long learning in Europe, by NUTS2 region, for 2009.

The variable is measured as the % of 25 to 64 year olds participating in education and training.

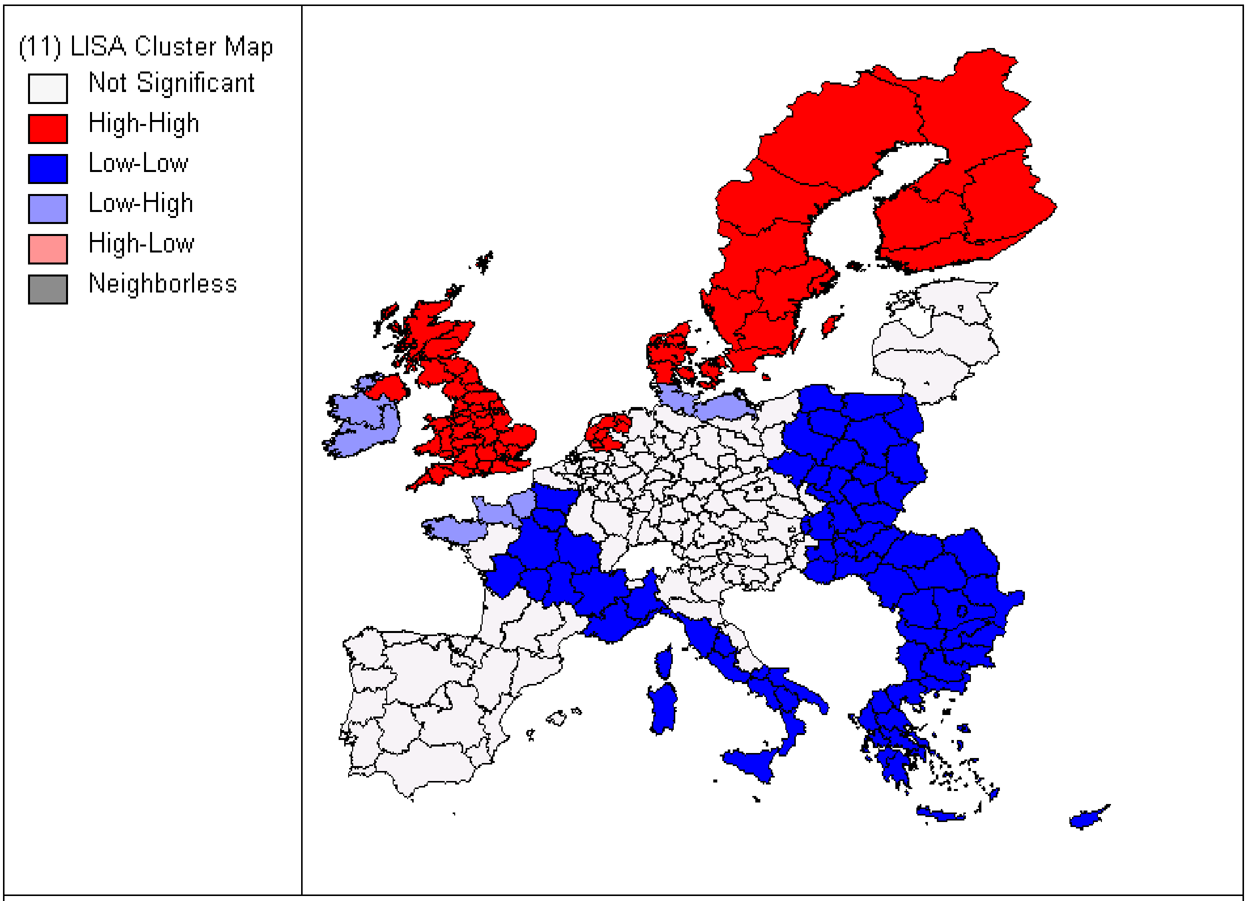


Figure 5 LISA map for life-long learning in Europe, by NUTS2 region, for 2009.

3.2 Spatial autocorrelation and Moran's I: Average farm size example

As a second example, we look at average farm size, using data for the year 2007. The pattern for average farm size shows a band of large farms from Denmark to the Slovakia and dark areas in central and northern France as well as Scotland, see *Figure 6*. Small farms predominate in Italy and South-Eastern Europe.

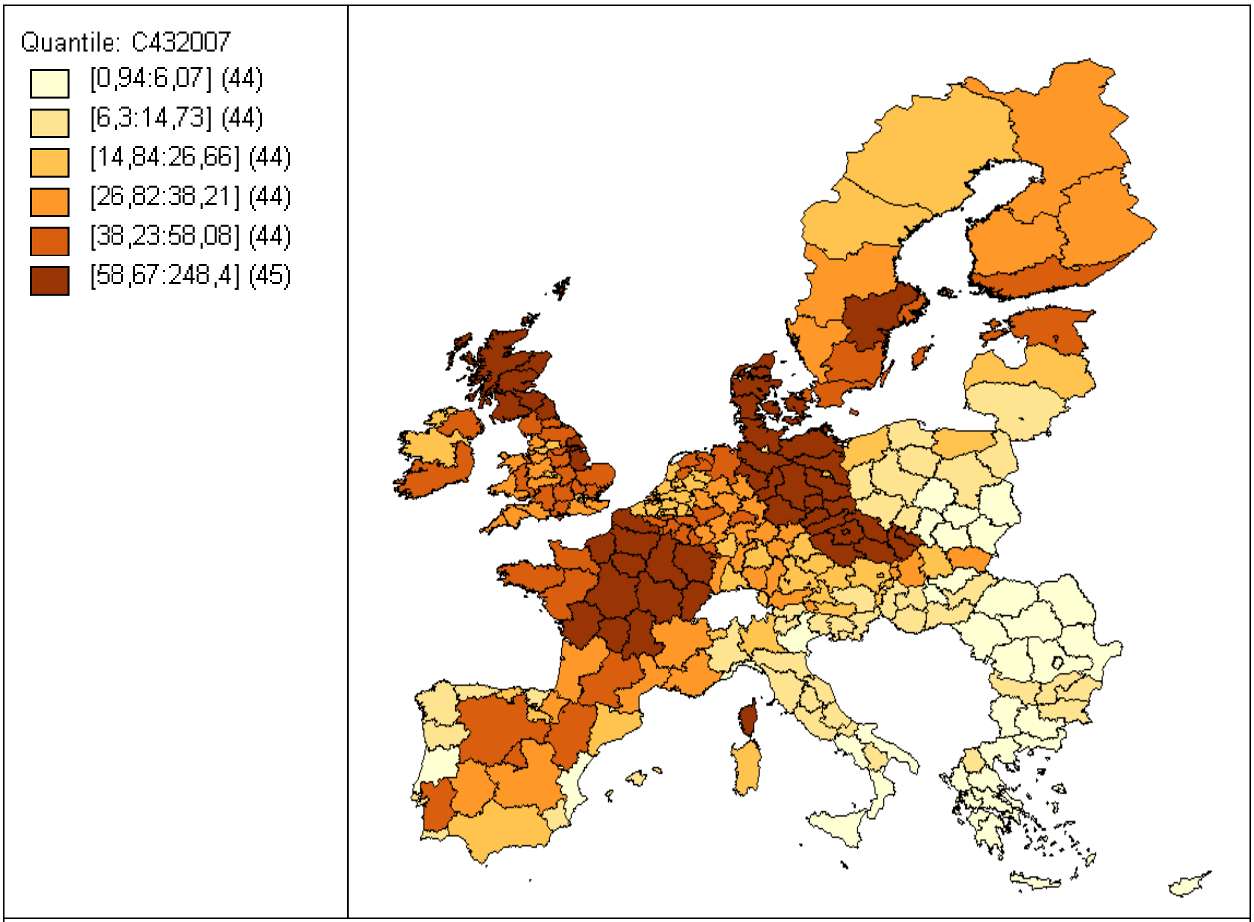


Figure 6 Average area farm size in Europe, in ha, by NUTS2 region, for 2007.

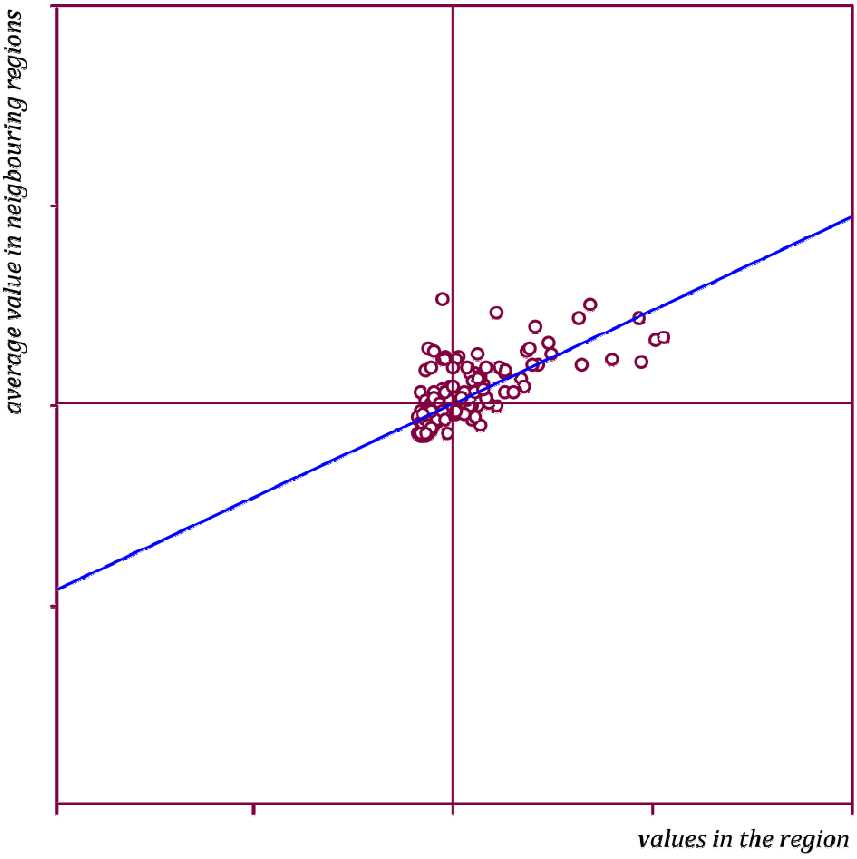


Figure 7 Scatterplot of average area farm size in Europe in a region (x-axis) and its neighbouring regions (y-axes) for 2007.

Moran’s I provides a measure of the spatial correlation between neighbours. Values range from -1 (indicating perfect dispersion) to +1 (perfect correlation), with 0 indicating a random spatial pattern. For statistical hypothesis testing, that indicates whether or not we can reject the null hypothesis. Moran's I values can be transformed to z-scores. In this case, the null hypothesis would be that there is no spatial clustering. The z-score is based on a randomization null hypothesis computation. in which values greater than 1.96 or smaller than -1.96 indicate spatial autocorrelation at the 5% significance level. For more information, see ESRI’s help page.

The graph in Figure 7 shows how observations in a region (on the horizontal axis) are related to values in surrounding regions (on the vertical axis; the axes cross at the overall average value). The slope of the blue line, basically the regression line fitted through the points, is Moran’s I, which in this case is 0.47. (This value has a meaning especially when compared to other variables, as long as the same regions and weight matrix are used.)

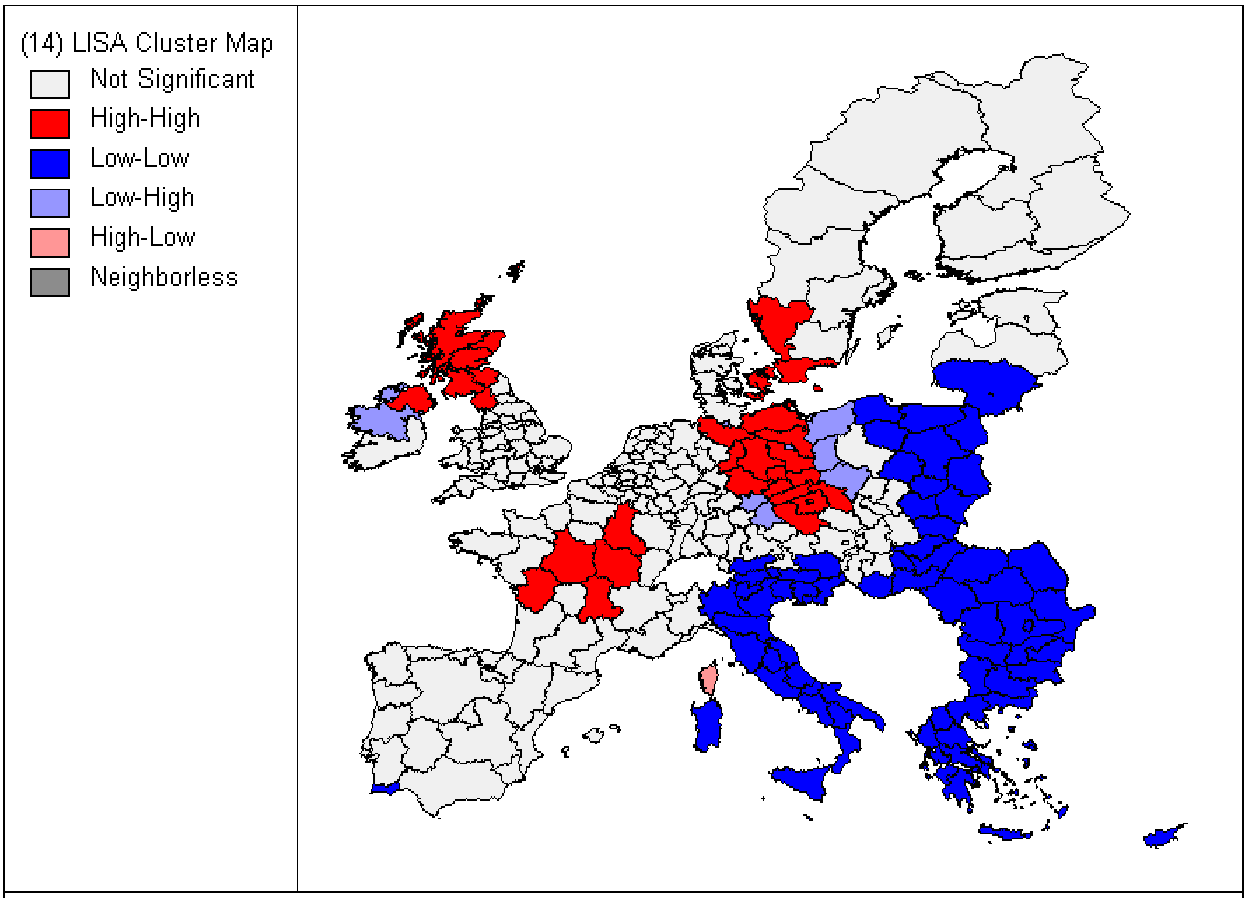


Figure 8 LISA map for average area farm size in Europe, by NUTS2 region, for 2007.

We see on the LISA map (*Figure 8*) that the clusters in Scotland and central France are indeed recognized, but Denmark is not significantly part of a cluster with eastern Germany and the Czech Republic. There is a strong Eastern European cluster of low values, which extends to Italy. However, Corsica has an exceptional high value compared to its neighbours. Likewise the north of Ireland has small holdings compared to the high values in Northern Ireland and in nearby Scotland.

4 Testing the econometric specification

4.1 Test for spatial model

Introduction

Many researchers use spatial econometrics in its simplest form, but they might not label it as such. Controlling for spatial heterogeneity using regional dummies or a distance to the nearest airport is a way of implementing spatial econometrics. Among the more advanced models, however, two main approaches are in use, covering situations where:

- either the outcome in one region is affected by the outcome in neighbouring regions (a spatial lag model);

- or the outcome in one region is affected by unknown characteristics of the neighbouring regions (a spatial error model).

4.1.1 Spatial lag model

An example of the first type would be a house price. Obviously, the price of a house depends on its age and size, the number of rooms, the presence of a garage, etc. However, the attractiveness (reflected in the prices) of nearby houses also have an impact.[1] In vector notation, we estimate

$$P = \alpha + \beta X + \rho WP + \epsilon$$

instead of

$$P = \alpha + \beta X + \epsilon$$

with *X* being a vector of house characteristics and *P* the price of a house; ρ is the coefficient estimated for the spatial lag. The most distinguishing aspect of the first formula is the vector *W*; this is the spatial weights matrix as discussed in section 3.1. Although this is a crucial element in a spatial econometric estimation, its function is fairly simple: it 'depreciates' the effects of the other observations by some distance-related characteristic. The most common characteristics are Euclidean distance (squared), travel time, and border contiguity.

We assume that the data that will be used have a panel structure (observations in space and time). As a result, we use a standard fixed effects panel data regression model in Stata for the spatial lag model.:

```
. xtreg P X Pt1_wq, fe
. estimates store FE
```

In the Stata commands X represents a list of independent variables, and Pt1_wq is the spatial lagged dependent variable. Note that the spatially lagged dependent variable is also lagged in time with one period. The construction of the spatially lagged variable is discussed in Section 4.1.4. The fixed effects are the regional specific intercepts.

[1] The example is not perfect, as all housing prices in the neighbourhood are also influenced by an unobserved "neighbourhood quality" variable.

Instead of using the fixed effect regression model, we can also use the random effects model. Then the regional specific effects are not fixed but assumed to be randomly drawn from a distribution estimated. The random effects panel data regression model in Stata is:

```
    . xtreg P X Pt1_wq, re
    . estimates store RE
```

More information can be found on Stata help on `xtreg`.

Fixed effects versus random effects

If effects are fixed, the RE estimator is inconsistent. The FE estimator (or within estimator) is less desirable because it assumes only within variation which leads to less-efficient estimation results and it is not able to estimate coefficients of time invariant regressors. With the Hausman test, the choice between fixed-effects and random effects models can be tested with a χ^2-test. The null-hypothesis is that individual or regional effects are fixed: .

```
    . hausman FE RE, sigmamore
```

Here, FE and RE are the data on the coefficients for the fixed effects and random effect regression estimations respectively. The null hypothesis is rejected if the probability of the $\chi^2(k)$ is smaller than 0.05 where *k* is the number of coefficients to be tested in the model. This means that the RE model estimation is preferred over the FE model estimation. Cameron and Trivedi (2010, p. 266 and further) present an example of the Hausman test.

4.1.2 Spatial error model

For the second case, the so-called spatial error model, we can think of productivity in a factory. If we have information on just inputs of labour and capital as well as the sector of a firm, and estimate

$$Prod = \alpha + \beta \text{Labour} + \gamma \text{Capital} + \delta \text{Sector}_{\text{dummies}} + \epsilon$$

then a map of the error terms ϵ might show a spatial pattern – most likely, clusters of high and low values together. Those unobserved effects are probably agglomeration effects, and if we cannot control for them, they will distort the estimates for β, $\gamma$and δ. We can prevent this by setting

$$\epsilon = \lambda W\epsilon + u$$

with λ as the coefficient estimated for the spatial error, and *W* again as the spatial weight matrix. *u* is the unobserved non-spatial error for every observation.

4.1.3 Testing for spatial lag or spatial error model

Tests exist to decide whether spatial econometrics are appropriate in a regression, and if so, whether a spatial lag or a spatial error model is more useful. Such tests are best executed in GeoDa, where an OLS regression command is available and will always report test results. More information is available at http://geodacenter.asu.edu/node/397#ols. Testing can also be done in Stata through `spmat`, but these unfortunately require a manually created weight matrix different from the ones used by GeoDa and the Stata module `sppack`, and investing time in producing yet another weight matrix is probably more inconvenient than using GeoDa.

GeoDa reports five statistics after OLS, of which the fifth can be ignored. The other four are two statistics for the Lag model and two for the Error model, each once regular and once robust. Anselin (2005) gives the following decision tree to decide which model is most appropriate: it boils down to looking at the regular statistics first, and using the robust alternatives only if the regular statistics are significant for both Lag and Error.

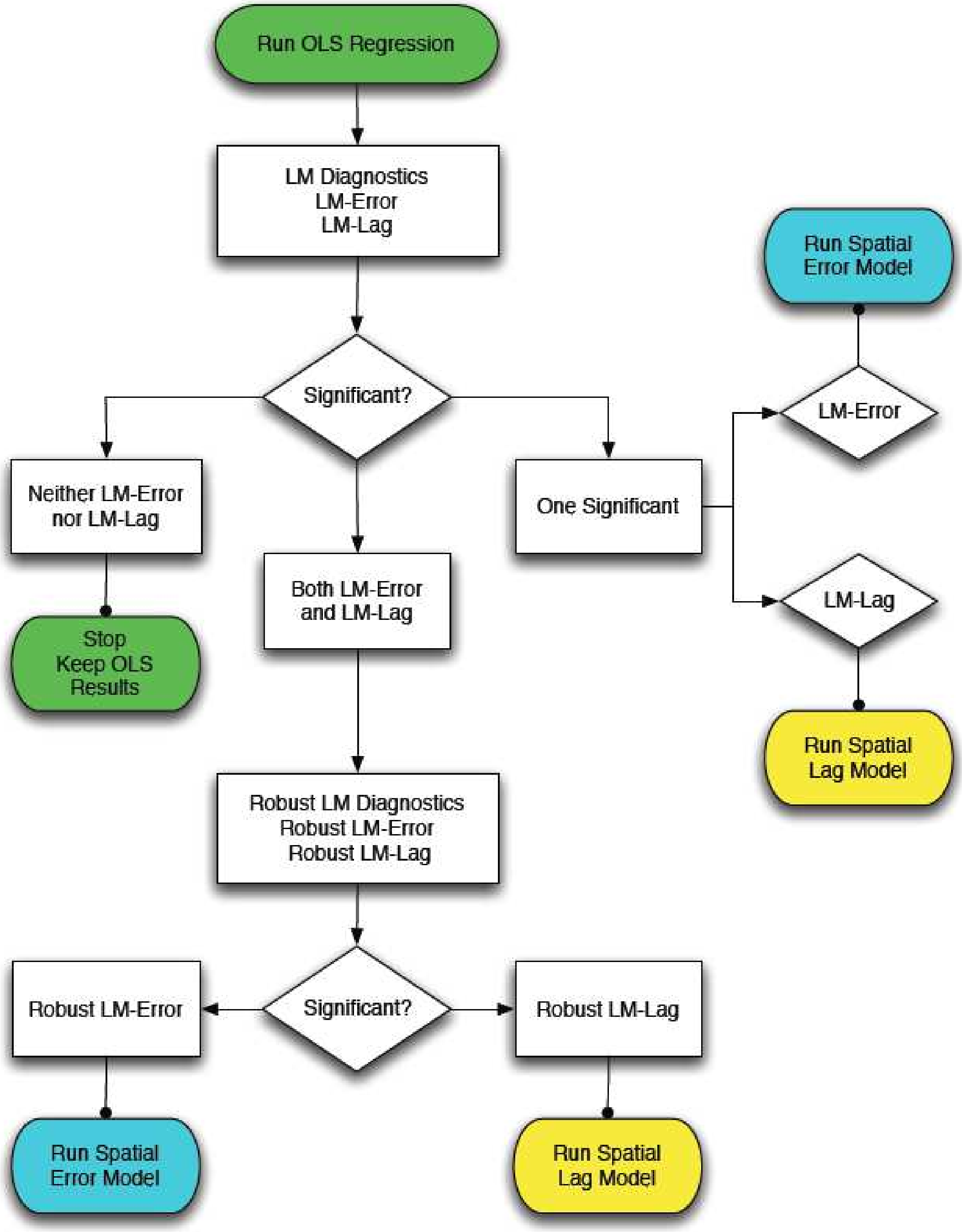


Figure 9 Decision model, from Anselin (2005).

4.1.4 Creating spatially lagged variables

To calculate a spatially lagged variable, we use the packages `shp2dta` and `spmat` for Stata. We then read a standard GISCO map into STAT using `shp2dta`. This produces two files, a database file and a coordinates file. For more information, use `help shp2dta` within Stata.

```
ssc install shp2dta
ssc install sppack

shp2dta using "NUTS_RG_10M_2003.shp", database(nuts2db)
   coordinates(nuts2coord) replace
```

We now keep only the NUTS2 regions in the database by dropping all others, and then likewise for the coordinates database.

```
use nuts2db, clear
describe
rename NUTS_ID nuts_id
drop if STAT_LEVL!=2
save "nuts2db.dta", replace

use nuts2coord, clear
merge m:1 _ID using nuts2db, keep(match) keepusing( )
drop POLY_ID-_merge
save "nuts2coord.dta", replace
```

We now make a queen contiguity matrix, which we call `nuts2q.` Subsequently, we include the regional ID's in our data file and set it to use the queen contiguity matrix `nuts2q`. Finally, we create a spatially lagged variable of `var` using said matrix, and we choose to call this variable `var_wq`. We can now use this variable just like any ordinary variable in regressions. For more information, see `help spmat` in Stata.

```
use nuts2db.dta
spmat contiguity nuts2q using nuts2coor, id(_ID)
spmat save nuts2q using nuts2q.spmat

use datafile.dta
rename region NUTS_ID
merge m:1 NUTS_ID using nuts2db, keepusing(_ID)
rename NUTS_ID region
drop if _merge==1

spmat use nuts2q using nuts2q.spmat
spmat lag var_wq nuts2q var
```

4.2 Model specification test (diagnostics)

A model specification error can occur when one or more relevant variables are omitted from the model or one or more irrelevant variables are included in the model. If relevant variables are omitted from the model, the common variance they share with included variables may be wrongly attributed to those variables, and the error term is inflated. On the other hand, if irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to them. Model specification errors can substantially affect the estimate of regression coefficients. This section presents a summary of chapter 2 of the online Stata web book:

http://www.ats.ucla.edu/stat/Stata/webbooks/reg/chapter2/statareg2.htm.

The advantage of using this source is that it uses examples of Stata commands and estimation. The online version is much more elaborated than the summary presented here.

This section will emphasize a number of specification test

- Independence of observations: Durbin –Watson test, see section 4.2.1
- Endogeneity of regressors Durbin-Wu-Hausman test, see section 4.2.2
- Omitted variables: Ramsey test, see section 4.2.3
- Multicollinearity, see section 4.2.4
- Homogeneity test, see section 4.2.5
- Normality test, see section 4.2.6

4.2.1 Test on independence of observations

Econometric estimation procedures usually assume the IID (identically and independently distributed) property of the errors which means that the errors associated with one observation are not correlated with the errors of any other observation cover several different situations. Consider the case of collecting data from students in eight different elementary schools. It is likely that the students within each school will tend to be more like one another than students from different schools, that is, their errors are not independent. Another way in which the assumption of independence can be broken is when data are collected on the same variables over time. Let us say that we collect truancy data every semester for 12 years. In this situation it is likely that the errors for observation between adjacent semesters will be more highly correlated than for observations more separated in time. This is known as autocorrelation (in time series context). When you have data that can be considered to be time-series you should use the `dwstat` command that performs a Durbin-Watson test for correlated residuals.

We do not have any time-series data, so we will use the `elemapi2` dataset and pretend that `snum` indicates the time at which the data were collected. We will also need to use the `tsset` command to let Stata know which variable is the time variable.

```
use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi2
tsset snum
       time variable:  snum, 58 to 6072, but with gaps

regress api00 enroll
```

```
( output omitted )

dwstat

Number of gaps in sample:  311
Durbin-Watson d-statistic(  2,   400) =  .2892712
```

The Durbin-Watson statistic has a range from 0 to 4 with a midpoint of 2. The observed value in our example is very small, close to zero, which is not surprising since our data are not truly time-series. A simple visual check would be to plot the residuals versus the time variable.

4.2.2 Test on endogeneity of regressors

There is an endogeneity problem if the coefficient of the variable res_hat is statistically significant. As a consequence, the regression model has to be estimated by Instrumental Variables (IV) techniques to take into account the endogeneity of regressors. More information on IV-estimators can be found at the Stata help function.

```
Regress depvar var1 var2 var3
Predict res_hat
Regress depvar res_hat var1 var2 var3
```

More information on the endogeneity test can be found on Cameron and Trivedi (2010, pp. 188).

4.2.3 Test on omitted variables: Ramsey test

The ovtest command performs a regression specification error test (RESET) for omitted variables. It also creates new variables based on the predictors and refits the model using those new variables to see if any of them would be significant.

```
ovtest

Ramsey RESET test using powers of the fitted values of api00
      Ho:  model has no omitted variables
                F(3, 393) =      4.13
                Prob > F =      0.0067
```

The ovtest command in the example above indicates that there are omitted variables.

More information on the Ramsey test can be found on Cameron and Trivedi (2010, pp. 98-100) or at the Stata help on ovtest.

4.2.4 Test on multicollinearity of explanatory variables

When there is a perfect linear relationship among the predictors, the estimates for a regression model cannot be uniquely computed. The term collinearity implies that two variables are near perfect linear combinations of one another. When more than two variables are involved it is often called multicollinearity, although the two terms are often used interchangeably.

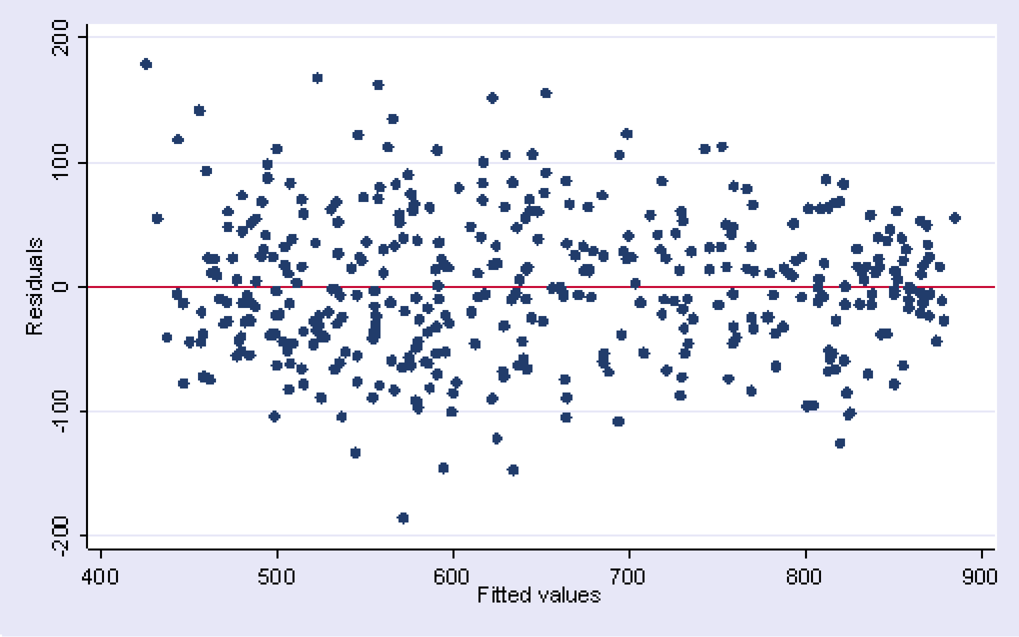
The primary concern is that as the degree of multicollinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients can get wildly inflated. In this section, we will explore some Stata commands that help to detect multicollinearity.

We can use the `vif` command after the regression to check for multicollinearity. `vif` stands for *variance inflation factor*. As a rule of thumb, a variable whose VIF values are greater than 10 may merit further investigation. Tolerance, defined as 1/VIF, is used by many researchers to check on the degree of collinearity. A tolerance value lower than 0.1 is comparable to a VIF of 10. It means that the variable could be considered as a linear combination of other independent variables. Let us first look at the regression we did from the last section, the regression model predicting the variable api00 from the variables meals, ell and emer and then issue the `vif` command."

More information on the Ramsey test can be found on Cameron and Trivedi (2010, p. 379).

4.2.5 Test on homogeneity

One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroscedastic". There are two test for heteroscedasticity:



Now let's look at a couple of commands that test for heteroscedasticity. The first test on heteroscedasticity given by `itmest` is the White's test and the second one given by `hettest` is the Breusch-Pagan test. Both test the null hypothesis that the variance of the residuals is homogenous. Therefore, if the p-value is very small, we would have to reject the hypothesis and accept the alternative hypothesis that the variance is not homogenous.

```
estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |       18.35      9     0.0313
            Skewness |       7.78      3     0.0507
            Kurtosis |       0.27      1     0.6067
---------------------+-----------------------------
               Total |       26.40     13     0.0150
---------------------------------------------------
estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
        Ho: Constant variance
        Variables: fitted values of api00
        chi2(1)      =     8.75
        Prob > chi2  =   0.0031
```

So in this case, the evidence is against the null hypothesis that the variance is homogeneous. These tests are very sensitive to model assumptions, such as the assumption of normality. Therefore it is a common practice to combine the tests with diagnostic plots to make a judgment on the severity of the heteroscedasticity and to decide if any correction is needed for heteroscedasticity. In our case, the plot above (to be added) does not show too strong an evidence. So we are not going to get into details on how to correct for heteroscedasticity even though there are methods available.

4.2.6 Test on distribution of the residuals

Normality of residuals is only required for valid hypothesis testing, that is, the normality assumption assures that the p-values for the t-tests and F-test will be valid. Normality is not required in order to obtain unbiased estimates of the regression coefficients. OLS regression merely requires that the residuals (errors) be identically and independently distributed (IID). Furthermore, there is no assumption or requirement that the predictor variables be normally distributed. If this were the case than we would not be able to use dummy coded variables in our models.

After we run a regression analysis, we can use the `predict` command to create residuals and then use commands such as `kdensity`, `qnorm` and `pnorm` to check the normality of the residuals. More information on `qnorm` and `pnorm` can be found on the help function of Stata http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

Let us use the `elemapi2` data file for an example on the `kdensity` statement Let us predict the academic performance (api00) from the percentage receiving free meals (meals), the percentage of English language learners (ell), and percentage of teachers with emergency credentials (emer).

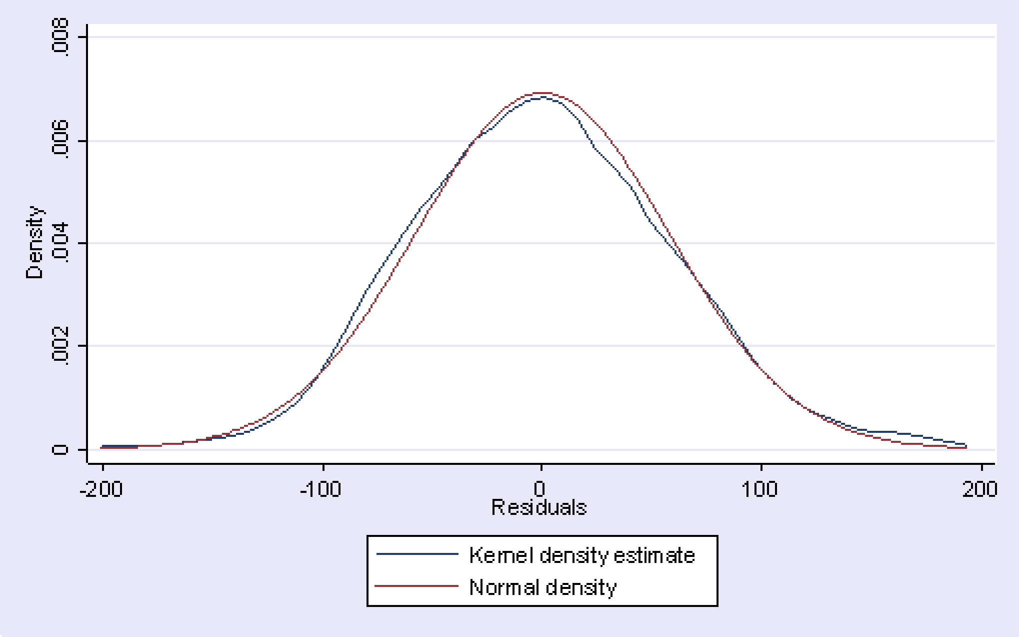
```
use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi2
regress api00 meals ell emer

  Source |       SS       df       MS          Number of obs =     400
---------+------------------------------       F(  3,   396) =  673.00
   Model |  6749782.75     3  2249927.58       Prob > F      =  0.0000
Residual |  1323889.25   396  3343.15467       R-squared     =  0.8360
---------+------------------------------       Adj R-squared =  0.8348
   Total |  8073672.00   399  20234.7669       Root MSE      =   57.82

--------------------------------------------------------------------
   api00 |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
---------+----------------------------------------------------------
   meals |  -3.159189   .1497371    -21.098   0.000      -3.453568   -2.864809
     ell |  -.9098732   .1846442     -4.928   0.000      -1.272878   -.5468678
    emer |  -1.573496    .293112     -5.368   0.000      -2.149746   -.9972456
   _cons |   886.7033    6.25976    141.651   0.000       874.3967    899.0098
------------------------------------------------------------------------------
```

We then use the `predict` command to generate residuals.

```
predict r, resid
```

Below we use the `kdensity` command to produce a kernel density plot with the `normal` option requesting that a normal density be overlaid on the plot. `kdensity` stands for kernel density estimate. It can be thought of as a histogram with narrow bins and moving average, see the Stata output below.

```
kdensity r, normal
```



There are also numerical tests for testing normality. Another test available is the Shapiro-Wilk W test for normality. The `swilk` command performs the Shapiro-Wilk W test for normality:

```
swilk r

                  Shapiro-Wilk W test for normal data
Variable |    Obs          W         V         z    Pr > z
---------+-------------------------------------------------
       r |    400    0.99641    0.989    -0.025   0.51006
```

The p-value is based on the assumption that the distribution is normal. In our example, it is very large (.51), indicating that we cannot reject that `r` is normally distributed.

(Source: section 2.2 in
http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm)

4.3 Tests on coefficients

4.3.1 Test on linear restrictions of coefficients

Testing a single coefficient

Suppose one of the variables *VAR1* in our regression has a coefficient β_1. The hypothesis we would like to test is whether β_1 is equal to 0. If this hypothesis is rejected, the coefficient β_1 of *VAR1* is significantly different from 0. Stata uses a WALD test for testing the hypothesis, see C&T406. To test H_0:β_1=0, we have

```
. * Testing a single coefficient equal to 0
. Test beta1
  ( 1) beta1 = 0
        Chi2(  1) =  70.80
      Prob .> chi2 =   0.000
```

The null-hypothesis is rejected if the probability is smaller than 0.05. As a consequence, the coefficient of variable VAR1 is significant. If the null-hypothesis is not rejected, one can consider to exclude the variable from the regression equation.

Testing multiple coefficients

Suppose we have the variables *VAR1* to VAR3 in our regression with coefficient β_1 to β_3. The hypothesis we would like to test is whether β_1 is equal to 0, and whether the sum of the coefficients of the variables VAR2 and VAR3 is equal to1. If this hypothesis is rejected, the coefficient β_1 of *VAR1* is significantly different from 0, and the sum of the coefficients of VAR2 and VAR 3 is not equal to 1. Stata uses a WALD test for testing the hypothesis, see Cameron and Trivedi (2009, p. 406). To test H_0: β_1=0 and β_2+ β_3=1, we have

```
. * Testing two hypotheses jointly
. xtreg y VAR1 VAR2 VAR3 VAR4, fe
. Test (beta1) (beta2 + beta3 = 1)
  ( 1) beta1 = 0
  ( 2) beta2 + beta3 = 1
        Chi2(  2) = 122.29
      Prob .> chi2 =   0.000
```

If the `mtest` option is added to the multiple tests command in Stata, each hypothesis is tested in isolation as well, i.e.

```
    . Test (beta1) (beta2 + beta3 = 1), mtest
```

More information on testing linear restrictions can be found in Cameron and Trivedi (2009, 403-409) or the Stata help on `test`.

4.3.2 Test on structural change

LR test on two models, one restricted model and one unrestricted model (this is not the same as imposing linear restrictions).

4.3.3 Tests on linearity in variables

This is more a procedure of trial and error than a straightforward test. In linear regression models, variables are either nominal variables or dummy variables. Nominal variables are included as linear function of the dependent variable, in other words . Alternatively, one can use quadratic or third-order polynomial of nominal variables to add to linear regression equations.

```
    . * Testing two hypotheses jointly
    . xtreg y VAR1 VAR2 VAR3 VAR4, fe
    . estimate store Regr1
    . generate VAR1sq = VAR1*VAR1
    . xtreg y VAR1 VAR1sq VAR2 VAR3 VAR4, fe
    . estimate store Regr2
    . LRtest Regr1 Regr2, force
```

To test whether higher-order polynomials of nominal functions add explanatory power to the estimation one could perform a Likelihood Ration test (Cameron and Trivedi, 2010, p. 416). In addition, one has to check two aspects. First, is the coefficient of VAR1sq significant.

Secondly, is the LR test rejected, then VAR1sq has significantly additional explanatory power for the regression estimation and has to be maintained.

One should keep in mind that VAR1sq might correlate significantly to VAR1. If this is the case, then the results of the estimation with VAR1sq might be biased.

4.4 Goodness of fit tests of the model

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. The most commonly used measure of goodness of fit is the R-squared statistic. Goodness of fit measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions, i.e. Kolmogorov–Smirnov test.

5 Concluding remarks

In this document the general econometric test to assess the impact of RDPs are presented.

In WP4 we will focus on (modelling) a selection of indicators for Rural Development Measures that differ with respect to impact and provide an overview of the relevant aspects of RDPs. We analyse which relations between Rural Development Indicators (and other data available) are affected by spatial interactions and thus have to be tested using spatial econometrics. In the forthcoming document 4.2 the database is analysed for spatial patterns. Explanatory Spatial Data Analysis (ESDA) to assess the spatial distribution of the relevant data at the relevant scale level (NUTS0-NUTS2-NUTS3). To apply ESDA the weight matrix

has to be adjusted to each relevant scale level.

Thereafter the model is specified and estimated at NUTS0 level - EU wide with focus on the variation between the member states. The difference in impact of RD Measures is explained at member state level. Then the model for the case studies - EU-depth (NUTS2 and NUTS3 level) is specified in a generic form for WP5. The necessary information is provided by the case studies. Finally we report on general methodology with recommendations for use in EU RDPs.

Acknowledgement

All maps are © EuroGeographics for the administrative boundaries

References

Anselin, L. (2005), *Exploring Spatial Data with GeoDa: A Workbook*, http://geodacenter.asu.edu/system/files/geodaworkbook.pdf

Anselin, L., J. Le Gallo and H. Jayet (2008), *Spatial Panel Econometrics*. In: L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data.* Springer Verlag, Heidelberg.

Cameron A. Colin and Pravin K. Trivedi (2009), *Microeconometrics using Stata*. Stata Press, College Station.

Getis, A. (2009), "Spatial Weight Matrices". *Geographical Analysis* 41, pp. 404–410.

Harris, R., J. Moffat and V. Kravtsova (2011). In Search of 'W'. *Spatial Economic Analysis* 6(3), 249-270.

Uthes, S., T. Kuhlman, S. Reinhard, P. Nowicki, M. J. Smit, E. van Leeuwen, A. L. Silburn, I. Zasada and A. Piorr (2011) Report on analytical framework – conceptual model, data sources, and implications for spatial econometric modeling, Müncheberg, ZALF.