Misspecification of Space: An Illustration Using Regional Growth Convergence Regressions

JAN MUTL[†]

February 17, 2009

Abstract

I illustrate the importance of choosing the correct space in empirical applications of spatial econometric models. I consider different spatial weighting matrices in a SAR(1) model – contiguity matrix, distance based matrix and their variants adjusted for 'size' of each observation. I show formally that only the modified weighting matrices imply specifications that are robust to changes in the sample size. I demonstrate the effect of spatial misspecification by presenting simulations of a regional convergence model. The different specification of space are also estimated using European regional data. The results confirm the sensitivity of the conclusion with respect to the choice of the space.

JEL Codes: C21, C23, F43, O47

Keywords: beta growth convergence; spatial econometrics; spatial weights specification; aggregation

Wordcount: 4,800

[†]*Institute for Advanced Studies, Stumpergasse 56, A-1060 Vienna, Austria (email: mutl@ihs.ac.at)*

I. Introduction

The motivation for this paper is to point out certain type of aggregation problems in models with spatial dependence. In order to demonstrate the issue on a practical example, I revisit the question of determining whether poor regions converge to the richer regions. There is large body of literature dealing with the topic of so called β and $\sigma-$convergence, following the work of Barro and Sala-i-Martin (2004). Recently, it has attracted a renewed interest and many authors re-estimated the convergence regression using spatial specifications. See, for example, the papers in the special issue of Papers in Regional Science (Bode and Rey, 2006 give an overview). Many applications use European NUTS regional data[1] on economic and social variables and then assume that the relevant space is based on a spatial weighting matrix that is the row normalized contiguity matrix for the regions in the sample. This implies that the relevant distance of two regions is determined only by the number of regions one has to cross to get from one to the other. Given the diversity of the NUTS regions in terms of geographical and population size, this can be misleading and many authors in this literature admit that such specification is chosen in the absence of better data. Other papers (for example Le Gallo and Cern, 2003) use distance based weighting matrices. Overall, there is no clear consensus as to what is the appropriate specification of the economic space. The lack of consensus might not be problematic if the empirical results are robust with respect to the choice of the spatial weights. Note that both contiguity and distance based weights yield spatial weighting matrices that are in some sense 'similar' and it might be difficult to statistically distinguish these as alternatives.

In this paper, I illustrate the importance of choosing the correct space in the empirical application and document the surprisingly large sensitivity of the empirical results to the choice of the spatial weights. This gives some hope for development of formal statistical tests with reasonable power properties that would be able to pose different spatial specifications as alternatives. Note that the existing tests (such as Kelejian and Prucha,

[1] NUTS stands for 'Nomenclature des unités territoriales statistiques', see http://europa.eu.int/comm/eurostat/ramon/nuts for definitions.

2001), provide a test of a particular spatial weights matrix against the alternative of no spatial autocorrelation.

I consider several different specification of the spatial weighting matrix in a convergence regression. The first is the usual contiguity matrix and its row normalized variant. The second spatial weights specification improves on the first by calculating the population weighted centers for each region and then constructing the weights as a function of the distances among the centers of the regions. The last set of weighting matrices reflects the suggestions made in this paper and takes the contiguity and distance based weights and adjusts them by the relative sizes of the appropriate regions.

The adjustment to the weight matrices is motivated by the fact that many applications use regional data that is available for widely heterogeneous (in terms of size) regions. Note that regional data available in Europe follows the NUTS classification. At the NUTS 1 level, the largest region in terms of population is Nordrhein-Westfalen in Germany with 18 million inhabitants, while the smallest region of Åland in Finland has only 26 thousand inhabitants.[2] At a finer level of classification (NUTS 2), the size varies in a similar fashion.[3] The adjustment for the size of a region should allow the researcher to better fit the data when using such heterogenous samples. The modification of the spatial weights also allows to consider samples that combine data at different levels of aggregation (i.e. combining NUTS 1 level data with NUTS 2 level data where available). In the rest of the paper I provide a theoretical justification for the modification and, in particular, I show that my preferred specification is robust to changes in the sample when a particular region is replaced by its sub-regions. I also examine the differences that are likely to arise in practical applications.

In the next section I look at what happens to the different specifications of space when the sample size increases and show that adjusting the spatial weights for the size of the appropriate regions is the only possibility to retain an invariant specification. I take this

[2]Based on Eurostat data for regional population as of January 2002, series d2jan from the directory 'General and regional statistics/Regions/Population and area', accessible at http://epp.eurostat.cec.eu.int.

[3]The largest NUTS 2 region is Île de France with 11 million of inhabitants. The smallest region is again Åland in Finland with only 26 thousand inhabitants.

idea to the European data in Section 3, where I estimate a $\beta-$convergence regression with spatial lag of the dependent variable using different definitions of the spatial weights. The estimated coefficients are then used as guidance for setting the parameter space for the Monte Carlo experiments in Section 4. I provide conclusions and suggestions for future research in Section 5.

II. Theoretical Considerations

When one combines heterogenous locations in one sample, one has to pay extra attention to the way the spatial weights are specified. To fix ideas, let us consider SAR(1) model (in Anselin's terminology, Anselin (1998)):

$$y_{i,n} = \mathbf{x}_{i,n}\boldsymbol{\beta} + \rho \sum_{j=1}^{n} w_{ij,n} y_{i,n} + u_{i,n}, \tag{1}$$

where $y_{i,n}$ is the dependent variable at location i for the sample size n, $\mathbf{x}_{i,n}$ is a vector of explanatory variables, $u_{i,n}$ is the disturbance term and $\boldsymbol{\beta}$ and ρ are (vector and scalar) parameters. The spatial weights w_{ij} are assumed to be nonstochastic and known to the researcher. I now consider several specifications for the spatial weights common in the applied literature and judge whether these are sensible or not when the sample contains units with different levels of aggregation and/or units that are heterogeneous in size. I have the following thought experiment in mind. I write down the model (1) for two different sample sizes n_1 and n_2, with $n_1 < n_2$. The larger sample is obtained by splitting a particular location into several new ones, due to say better data availability. Such increase can occur, for example, when the European regional data on NUTS 1 level are used and a particular region is now replaced by its NUTS 2 level components. However, this thought experiment also captures situations where it is reasonable to assume that the true data generating process operates at a finer level of disaggregation while in the observed sample some of the regions are aggregated together.

I now compare the implication of the increased sample size on the observations that remain unchanged in the sample. I propose that for sensible specifications, the change in

the sampling design should not affect the locations that are retained unchanged in the sample. I show that several of the specifications do not have this property and suggest modifications that alleviate this problem.

Observe that I consistently specify the components of our model in (1), including the spatial weights $w_{ij,n}$, to be triangular arrays and depend on the sample size. Let us now consider the new sample to be created from the old sample by splitting a fixed region. Without loss of generality I assume that this region has an index n, and it is split into two new regions, say n and $n+1$. Consider the equation (1) written for an arbitrary region $i < n$:

$$y_{i,n} = \mathbf{x}_{i,n}\boldsymbol{\beta} + \rho \left[\left(\sum_{j=1}^{n-1} w_{ij,n} y_{i,n} \right) + w_{in,n} y_{n,n} \right] + u_{i,n}, \tag{2}$$

for the sample size n, and

$$y_{i,n+1} = \mathbf{x}_{i,n+1}\boldsymbol{\beta} + \rho \left[\left(\sum_{j=1}^{n-1} w_{ij,n+1} y_{i,n+1} \right) + w_{in,n+1} y_{n,n+1} + w_{i,n+1,n+1} y_{n+1,n+1} \right] + u_{i,n+1}, \tag{3}$$

for the sample size $n+1$. I assume that the locations $1, ..., n-1$ are not affected by the change in the sampling design and hence we have $\mathbf{x}_{i,n+1} = \mathbf{x}_{i,n}$, $w_{ij,n} = w_{ij,n+1}$ and $u_{i,n} = u_{i,n+1}$ for $i, j < n$.[4] Therefore,

$$y_{i,n+1} - y_{i,n} = \rho \left(w_{in,n+1} y_{n,n+1} + w_{i,n+1,n+1} y_{n+1,n+1} - w_{in,n} y_{n,n} \right). \tag{4}$$

If the specification is to be invariant to the increase in the sample size, the above expression must be equal or close to zero. Note that in a practical application, the observed value of y_i of course does not change with the sample size change assumed in the argument. The difference between the two specifications would be absorbed in the disturbances. The interpretation adopted here is that the underlying disturbance does not change with the sample size change and hence the implied difference among the specifications $(y_{i,n+1} - y_{i,n})$ can be interpreted as the additional error due to the aggregation problem. I consider next whether two popular specifications for the spatial

[4] This implicitly assumes that the spatial weights are not normalized. Row normalization complicates the albegra but does not change the results in this paper.

weights common in the literature - contiguity matrix and distance based weights - fulfill this condition.

Contiguity Based Weights

In this case, we have $w_{ij} = 0$ when the two regions (observations) are not neighbors. As a result, when the $i-th$ region was not neighboring the region n (in the sample of size n), all the weights $w_{in,n}$, $w_{in,n+1}$, and $w_{i,n+1,n+1}$ are equal to zero and we have trivially no difference between $y_{i,n}$ and $y_{i,n+1}$. However, when the region i is a neighbor of the region n, we have $w_{in,n} = 1$.[5] Suppose additionally that in the larger sample only the region n is a neighbor of i and hence $w_{in,n+1} = 1$ and $w_{i,n+1,n+1} = 0$. As a result, we have

$$y_{i,n+1} - y_{i,n} = \rho \left(y_{n,n+1} - y_{n,n} \right). \tag{5}$$

In many applications, the dependent variable is measured in levels and hence for a subregion we have that $y_{n,n+1} < y_{n,n}$, and, consequently $y_{i,n+1} - y_{i,n} \neq 0$. Thus the contiguity weights are not well-suited for variables in levels. If the dependent variable is a rate of change or size standardized variable (such as GDP per capita or productivity), one can postulate that

$$y_{n,n} = \frac{a_1}{a_1 + a_2} y_{n,n+1} + \frac{a_2}{a_1 + a_2} y_{n+1,n+1}, \tag{6}$$

where a_i are appropriate measures of the contribution of the two subregions to the dependent variable in region n (in sample size n). However, this specification leads to similar conclusion, namely that

$$\begin{aligned} y_{i,n+1} - y_{i,n} &= \rho \left(y_{n,n+1} - y_{n,n} \right) \\ &= \rho \left(y_{n,n+1} - \frac{a_1}{a_1 + a_2} y_{n,n+1} - \frac{a_2}{a_1 + a_2} y_{n+1,n+1} \right) \\ &= \rho \frac{a_2}{a_1 + a_2} \left(y_{n,n+1} - y_{n+1,n+1} \right) \neq 0. \end{aligned} \tag{7}$$

[5] Again, assuming that the spatial weighting matrix is not row normalized - under row normalization this would be an inverse of the number of neighbors of the region i. However, this does not change the conclusion made in this section.

On the other hand, contiguity weights matrix can be appropriate in spatial lags in the disturbances where one can reasonably postulate that the disturbances for the two subregions can be expected to be the same.

Distance Based Weights

Let us now consider distance based spatial weights. Here the w_{ij} depends negatively on the distance of observations (regions) i and j. One has to choose the specific functional form and I, as an example, consider the same functional form as has been used in the literature examining extent of intra- vs. international price variation[6] as well as in the literature on $\beta-$convergence.[7] In these papers, the effect of geographic distance is captured by using the following specification:

$$w_{ij} = -\log\left(\frac{d_{ij}}{d_{\max}}\right) = \log\left(d_{\max}\right) - \log\left(d_{ij}\right), \tag{8}$$

where d_{ij} is the distance of regions i and j, and $d_{\max}$ is the largest distance in the sample. Consider now the same increase in the sample size, i.e. region n is split into two new regions indexed by n and $n+1$. I pick an arbitrary region i and assume that its distances to the new subregions are given by

$$\begin{aligned} d_{in,n+1} &= d_{in,n} \cdot (1-\varepsilon_i), \\ d_{i,n+1,n+1} &= d_{in,n} \cdot (1+\varepsilon_i), \end{aligned} \tag{9}$$

where the last subscript indicates the sample size, i.e. $d_{in,n}$ is the distance of regions i and n in the sample size n; as opposed to the distance of regions i and the new subregions n and $n+1$ in sample size $n+1$, which is denoted by $d_{in,n+1}$ and $d_{i,n+1,n+1}$ respectively.

[6] See e.g. Engel and Rogers (1996) or Beck and Weber (2001) and many others.

[7] See e.g. Le Gallo and Cern (2003), Badinger et al. (2004), Egger and Pfaffermayr (2006) and many others.

Using this notation, the spatial weights are given by

$$\begin{aligned} w_{in,n} &= \log\left(d_{\max,n}\right) - \log\left(d_{in,n}\right), \\ w_{in,n+1} &= \log\left(d_{\max,n+1}\right) - \log\left(d_{in,n}\right) - \log\left(1-\varepsilon_i\right), \\ w_{i,n+1,n+1} &= \log\left(d_{\max,n+1}\right) - \log\left(d_{in,n}\right) - \log\left(1+\varepsilon_i\right). \end{aligned} \tag{10}$$

To simplify the derivations, I assume that the largest distance remains the same in the two samples and $d_{\max,n} = d_{\max,n+1}$. As before, I allow the dependent variable to be measured in levels and/or rates of change or ratios. Hence, I assume that the dependent variable in region n is a weighted average of the dependent variables in its subregions as in (6). The difference in the dependent variable y_i in the two sample sizes is then

$$\begin{aligned} (1/\rho)\left(y_{i,n+1} - y_{i,n}\right) &= w_{in,n+1}y_{n,n+1} + w_{i,n+1,n+1}y_{n+1,n+1} - w_{in,n}y_{n,n} \\ &= \left[w_{in,n} - \log\left(1-\varepsilon_i\right)\right]y_{n,n+1} \\ &\quad + \left[w_{in,n} - \log\left(1+\varepsilon_i\right)\right]y_{n+1,n+1} \\ &\quad - w_{in,n}\left(\frac{a_1}{a_1+a_2}y_{n,n+1} + \frac{a_2}{a_1+a_2}y_{n+1,n+1}\right) \\ &= w_{in,n}\left(\frac{a_2}{a_1+a_2}y_{n,n+1} + \frac{a_1}{a_1+a_2}y_{n+1,n+1}\right) \\ &\quad - y_{n,n+1}\log\left(1-\varepsilon_i\right) - y_{n+1,n+1}\log\left(1+\varepsilon_i\right) \\ &= w_{in,n}y_{n,n}\left(\frac{a_2}{a_1+a_2}\frac{a_1+a_2}{a_1} + \frac{a_1}{a_1+a_2}\frac{a_1+a_2}{a_2}\right) \\ &\quad - y_{n,n+1}\log\left(1-\varepsilon_i\right) - y_{n+1,n+1}\log\left(1+\varepsilon_i\right) \\ &= w_{in,n}y_{n,n}\left(\frac{a_2}{a_1} + \frac{a_1}{a_2}\right) \\ &\quad - y_{n,n+1}\log\left(1-\varepsilon_i\right) - y_{n+1,n+1}\log\left(1+\varepsilon_i\right). \end{aligned} \tag{11}$$

For small ε_i we can approximate the natural logarithm by a linear function:

$$-\log\left(1-\varepsilon_i\right) \approx \log\left(1+\varepsilon_i\right) \tag{12}$$

and thus

$$\begin{aligned} & -y_{n,n+1}\log\left(1-\varepsilon_i\right)-y_{n+1,n+1}\log\left(1+\varepsilon_i\right) \\ \approx\ & \log\left(1+\varepsilon_i\right)\cdot\left(y_{n,n+1}-y_{n+1,n+1}\right), \end{aligned} \tag{13}$$

which could potentially be equal to zero, provided that the region n (in sample n) is split into equal components (in sample $n+1$). Nevertheless, even under such scenario, we would still be left with

$$y_{i,n+1}-y_{i,n}\approx\rho w_{in,n}y_{n,n}\left(\frac{a_2}{a_1}+\frac{a_1}{a_2}\right)\neq 0. \tag{14}$$

Therefore, I conclude that distance based weights as considered in the literature are not well suited for applications with heterogeneous units and the weights are applied to a variable that is measured as a rate of change or as a ratio to size.

When the dependent variable is measured in levels, the distance based weights will only be appropriate when the increases in the sample size are due to splits into equal size sub-regions. To see this, consider the dependent variable to be such that

$$y_{n,n}=y_{n,n+1}+y_{n+1,n+1}, \tag{15}$$

and derive the analogy to (11) as

$$\begin{aligned} \left(1/\rho\right)\left(y_{i,n+1}-y_{i,n}\right) &= w_{in,n+1}y_{n,n+1}+w_{i,n+1,n+1}y_{n+1,n+1}-w_{in,n}y_{n,n} \\ &= \left[w_{in,n}-\log\left(1-\varepsilon_i\right)\right]y_{n,n+1} \\ &\quad +\left[w_{in,n}-\log\left(1+\varepsilon_i\right)\right]y_{n+1,n+1} \\ &\quad -w_{in,n}\left(y_{n,n+1}+y_{n+1,n+1}\right) \\ &= -y_{n,n+1}\log\left(1-\varepsilon_i\right)-y_{n+1,n+1}\log\left(1+\varepsilon_i\right) \\ &\approx \log\left(1+\varepsilon_i\right)\cdot\left(y_{n,n+1}-y_{n+1,n+1}\right). \end{aligned} \tag{16}$$

Thus there will be no difference between the specification for the two sample sizes only

when $y_{n,n+1} - y_{n+1,n+1} = 0$ (when $y_{i.n}$ is measured in levels). This in turn might not be satisfied in many applications.

Sample Size Consistent Weights

I now suggest appropriate modification to the spatial weights that make the empirical specification consistent under changing sample size. Let us first consider the case of distance based weights.

Suppose that the weights in (8) are amended by a factor that reflect the size of each region. I show here that this modification alleviates the problem of changing the sample size. I assume that the dependent variable in a region is a weighted average of the dependent variables in its sub-regions, and that these weights are known, as in (6). Hence each location i has with it associated size which I denote $a_{i,n}$ where the second subscript indicates the sample size. I assume that the size of the regions is additive, i.e. $a_{n,n} = a_{n,n+1} + a_{n+1,n+1}$. Observe that we can then rewrite (6) as

$$y_{n,n} = \frac{a_{n,n+1}}{a_{n,n}} y_{n,n+1} + \frac{a_{n+1,n+1}}{a_{n,n}} y_{n+1,n+1}. \tag{17}$$

I take the distance based weights $w_{ij,n}$ as defined in (8) and define the proposed modified weights as

$$w^*_{ij,n} = w_{ij,n} \cdot \left(\frac{a_{j,n}}{\sum_{k=1}^{n} a_{k,n}} \right). \tag{18}$$

Observe that these weight are consistent with different sample sizes. In particular, analogical to (11) we now have

$$\begin{aligned} (1/\rho)\left(y_{i,n+1} - y_{i,n}\right) &= w^*_{in,n+1} y_{n,n+1} + w^*_{i,n+1,n+1} y_{n+1,n+1} - w^*_{in,n} y_{n,n} \\ &= \left[w_{in,n} - \log\left(1 - \varepsilon_i\right)\right] \left(\frac{a_{n,n+1}}{\sum_{k=1}^{n+1} a_{k,n+1}} \right) y_{n,n+1} \\ &\quad + \left[w_{in,n} - \log\left(1 + \varepsilon_i\right)\right] \left(\frac{a_{n+1,n+1}}{\sum_{k=1}^{n+1} a_{k,n+1}} \right) y_{n+1,n+1} \\ &\quad - w_{in,n} \left(\frac{a_{n,n}}{\sum_{k=1}^{n} a_{k,n}} \right) y_{n,n}. \end{aligned} \tag{19}$$

I assume that the size of the regions retained in both sample is the same, i.e. $a_{i,n} = a_{i,n+1}$, for $i < n$. Thus $\sum_{k=1}^{n+1} a_{k,n+1} = \sum_{k=1}^{n} a_{k,n}$ and we have

$$\begin{aligned}(y_{i,n+1} - y_{i,n}) \sum_{k=1}^{n} a_{k,n}/\rho &= w_{in,n}\underbrace{(a_{n,n+1}y_{n,n+1} + a_{n+1,n+1}y_{n+1,n+1} - a_{j,n}y_{n,n})}_{=0} \\ &\quad -a_{n,n+1}y_{n,n+1}\log(1-\varepsilon_i) - a_{n+1,n+1}y_{n+1,n+1}\log(1+\varepsilon_i).\end{aligned} \tag{20}$$

As above, for small ε_i I approximate the natural logarithm by a linear function ($-\log(1-\varepsilon_i) \approx \log(1+\varepsilon_i)$), and thus

$$\begin{aligned}(y_{i,n+1} - y_{i,n}) \sum_{k=1}^{n} a_{k,n}/\rho &\approx \log(1+\varepsilon_i) \cdot (a_{n,n+1}y_{n,n+1} - a_{n+1,n+1}y_{n+1,n+1}) \\ &= \log(1+\varepsilon_i) \cdot \\ &\quad \left(a_{n,n+1}\frac{a_{n,n+1}+a_{n+1,n+1}}{a_{n,n+1}}y_{n,n} - a_{n+1,n+1}\frac{a_{n,n+1}+a_{n+1,n+1}}{a_{n+1,n+1}}y_{n,n}\right) \\ &= \log(1+\varepsilon_i) \cdot y_{n,n}\left(a_{n,n+1} + a_{n+1,n+1} - a_{n,n+1} + a_{n+1,n+1}\right) \\ &= 0.\end{aligned} \tag{21}$$

This demonstrate that the proposed adjusted distance based weights lead to specifications that are robust to sample size increases of the kind considered in this paper.

Next, consider the case of contiguity weights. The same adjustment makes the specification robust to sample size changes in this case as well. To see this consider the same change in the sample size, i.e. splitting the $n-th$ region into two new sub-regions. Let us again consider a region i such that it is a neighbor of the first sub-region but does not neighbor the second sub-region. In terms of our notation we have $w_{in,n} = w_{i,n,n+1} = 1$ and $w_{i,n+1,n+1} = 0$. Consider a set of transformed weights $w^*_{ij,n}$ as in (18). As above we have the difference between the specification for the two sample sizes for location i given

by

$$
\begin{aligned}
(1/\rho)\left(y_{i,n+1}-y_{i,n}\right) &= \frac{a_{n,n+1}}{\sum_{k=1}^{n+1}a_{k,n+1}}y_{n,n+1}-\frac{a_{n,n}}{\sum_{k=1}^{n}a_{k,n}}y_{n,n} \qquad (22)\\
&= \frac{a_{n,n+1}}{\sum_{k=1}^{n+1}a_{k,n+1}}\frac{a_{n,n+1}+a_{n+1,n+1}}{a_{n,n+1}}y_{n,n}-\frac{a_{n,n}}{\sum_{k=1}^{n}a_{k,n}}y_{n,n}\\
&= \frac{a_{n,n+1}+a_{n+1,n+1}}{\sum_{k=1}^{n+1}a_{k,n+1}}y_{n,n}-\frac{a_{n,n}}{\sum_{k=1}^{n}a_{k,n}}y_{n,n}\\
&= \frac{a_{n,n}}{\sum_{k=1}^{n+1}a_{k,n+1}}y_{n,n}-\frac{a_{n,n}}{\sum_{k=1}^{n}a_{k,n}}y_{n,n}=0.
\end{aligned}
$$

Hence, the proposed adjusted contiguity weights also lead to specification that are robust to sample size increases of the kind considered in this paper.

III. Estimation Results

I now take the different specifications of spacial weights to the data and compare the different estimates these will yield. I consider the 72 NUTS 1 and 206 NUTS 2 European regions[8] and construct the spatial weights based on their actual locations. The contiguity matrix is constructed based on the maps provided by Eurostat on their website.[9] In particular I set $w_{ij}^{c}=1$ whenever the two regions share a common border. The distances of the regions are calculated as distances of the population weighted centers for each region. I use the data from World Gazetteer.[10] In particular, the data includes longitude and latitude coordinates and population of cities in the world. For each municipality it also provides the country and containing region that were matched (after some linguistic issues are cleared) to NUTS 2 regions.

In particular, the data includes 203,957 municipalities that I was able to match to a NUTS 2 region in my sample. In terms of population, these municipalities represent 82 percent of the EU total for these regions. The population of these municipalities varies between 1 (Rochefourchat, Rhône-Alpes, France) and 3,398,362 (Berlin, Germany)

[8] I only consider the continental regions in order to be able to construct the contiguity matrix in a sensible way. Thus I exclude UK and Ireland and the overseas regions of France, Portugal and Spain.

[9] See http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html

[10] Available freely at htpp://www.world-gazeteer.com. I use the sample contained on the website in November 2005.

inhabitants. The median population size is about 870 inhabitants. In general, for all countries except for Germany, I was able to match the data entries to their NUTS 2 regions. For Germany, I used the list of local administrative units from Eurostat (LAU) and matched their names to the Gazetteer dataset. I found that in some cases there was a multiple match, or no match, and I dropped these locations. In terms of population, the dropped locations represent between 6.2 percent of the regions population (Baden Würtenberg) and 0.0 percent (Sachsen Anhalt). To check the extent of the coverage of our data, I calculate the population totals for NUTS 1 regions based on the matched locations and compare it to the population reported by Eurostat. The ratio of my data relative to the population totals from Eurostat varies between 83 and 105 percent.

I select the locations in a particular NUTS region and calculate their average latitude and longitude, weighting each location with its population size. As a result, I obtain the coordinates of the centers of each NUTS region, denoted by LAT_i and $LONG_i$. The geographic distance of regions i and j is the calculated using the formula

$$d_{ij} = \pi r \arccos\left[\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos\left(\lambda_j - \lambda_i\right)\right], \tag{23}$$

where $r = 6731.1$ is the Earth's radius, $\phi_i = \pi \frac{LAT_i}{180}$ is the latitude coordinate of the region i in degrees radius, and $\lambda_i = \pi \frac{LONG_i}{180}$ is the longitude coordinate of the region i in degrees radius.

I estimate the following regressions:

$$g_i = \alpha + \boldsymbol{\beta} y_{i0} + \rho \sum_{j=1}^{n} w_{ij} g_j + \varepsilon_i, \tag{24}$$

where g_i is the average growth rate of GDP per capita for the $i-$th region for the period 1995 to 2002 (calculated from the GDP per head at market prices; series e2gdp95 provided by Eurostat in the directory 'Regional Statistics/Regions/Economic Accounts/Gross Domestic Product Indicators', accessible at http:/epp-eurostat.cec.eu.int), y_{i0} is the natural logarithm of the initial GDP per capita of region i, w_{ij} are the spatial weights, ε_i is a disturbance term and α, β and ρ are parameters.

I estimate the above equation with the different specifications for w_{ij} and in particular consider:

(i) contiguity matrix where

$$w_{ij}^{c} = 1, \tag{25}$$

when regions i and j are neighbors and $w_{ij}^{c} = 0$ otherwise,

(ii) row normalized contiguity matrix with

$$w_{ij}^{cn} = \frac{w_{ij}^{c}}{\sum_{k=1}^{n} w_{ik}^{c}}, \tag{26}$$

(iii) population normalized contiguity matrix with

$$w_{ij}^{c*} = \frac{w_{ij}^{c} p_j / p}{\sum_{k=1}^{n} w_{ik}^{c} p_k / p}, \tag{27}$$

where p_k is the population of region k and p is the total population of the regions in the sample,

(iv) distance based weights with

$$w_{ij}^{d} = \log\left(d_{\max}\right) - \log\left(d_{ij}\right), \tag{28}$$

where d_{ij} is the distance of regions i and j and $d_{\max}$ is the largest distance in the sample,

(v) row normalized distance weights with

$$w_{ij}^{dn} = \frac{w_{ij}^{d}}{\sum_{k=1}^{n} w_{ik}^{d}}, \tag{29}$$

(vi) population normalized distance weights

$$w_{ij}^{d*} = \frac{w_{ij}^{d} p_j / p}{\sum_{k=1}^{n} w_{ik}^{d} p_k / p}. \tag{30}$$

The empirical equation is estimated with an instrumental variable (IV) procedure where I use higher order (up to the third order) spatial lags of the exogenous variable y_{i0} as instruments for the spatial lag of the dependent variable. For detailed description and large sample results of such procedure, see Kelejian and Prucha (1998). As a robustness check, I also estimate the model without the spatial lag of the dependent variable by ordinary least squares (OLS). Table 1 summarizes the results. In order to facilitate an easier comparison across different models, I report a coefficient $\rho^* = \rho \cdot \lambda_{\max}$ where $\lambda_{\max}$ is the largest (in absolute value) eigenvalue of the appropriate spatial weighting matrix. Observe that for row normalized weights matrices $\lambda_{\max} = 1$.

TABLE 1

Estimates of $\beta-$convergence regression, NUTS 1 regions

Parameter	α		β		ρ^*		σ^2
Model							
OLS	2.67	(17.51)	−0.25	(−15.22)	*n.a.*	*n.a.*	0.12
$\mathbf{W}^c$	2.99	(13.23)	−0.27	(−12.63)	−0.22	(−1.91)	0.12
$\mathbf{W}^{cn}$	−39.83	(−7.02)	3.56	(6.70)	19.41	(8.54)	2.09
$\mathbf{W}^{c*}$	0.90	(2.77)	−0.09	(−2.87)	0.81	(6.24)	0.12
$\mathbf{W}^d$	2.81	(18.35)	−0.24	(−15.70)	−0.56	(−2.86)	0.11
$\mathbf{W}^{dn}$	2.21	(5.89)	−0.23	(−10.37)	0.81	(1.33)	0.12
$\mathbf{W}^{d*}$	2.32	(5.87)	−0.23	(−10.32)	0.68	(0.93)	0.12

Notes: T-statistics are in brackets.

TABLE 2

Estimates of $\beta-$convergence regression, NUTS 2 regions

Parameter	α		β		ρ^*		σ^2
Model							
OLS	2.43	(23.92)	−0.22	(−20.63)	*n.a.*	*n.a.*	0.12
$\mathbf{W}^d$	2.62	(24.94)	−0.22	(−21.22)	−0.68	(−4.78)	0.12
$\mathbf{W}^{dn}$	2.16	(7.14)	−0.21	(−14.09)	0.54	(0.95)	0.12
$\mathbf{W}^{d*}$	2.15	(7.08)	−0.21	(−13.69)	0.55	(1.02)	0.12

Notes: T-statistics are in brackets.

Observe that due to scaling by the largest eigenvalue of the weights matrix, only estimates of ρ^* that are less then one in absolute value imply spatially stable model. At the same time, I expect the positive spatial autocorrelation to be present. In light of this, when using the contiguity based weights, only the weights scaled by population size ($\mathbf{W}^{c*}$) yield estimates of expected magnitude and sign. The distance based weights produce more robust results, with both the row normalized and population normalized weights ($\mathbf{W}^{dn}$ and $\mathbf{W}^{d*}$) leading to sensible estimates.

As one would expect based on the theoretical considerations in the previous section, the population size adjusted weights lead to similar estimates in both sample sizes. This is not true for the other specifications of space.

IV. Monte Carlo Experiments

To get an idea of the importance of the consideration in this paper, I conduct a simulation study. I want to make the Monte Carlo designs as close as possible to those in practical applications. Thus I take the same 72 NUTS 1 European regions considered in the previous section and construct the same weighting matrices. As an alternative with a larger sample size, I consider the sample of 206 NUTS 2 regions and calculate their weights in the same fashion. I also take logs of the actual level of the economic activity for the regions in 1995 as the initial values of the data generating process. I collect the initial values in an $n \times 1$ vector $\mathbf{y}_{0,n}$ (where as above $n \in \{72, 206\}$ is the number of regions). In all experiments I generate the data according to

$$\mathbf{g}_n = (\mathbf{I}_n - \rho \mathbf{W}_n)^{-1} \left(\alpha + \boldsymbol{\beta} \mathbf{y}_{0,n} + \boldsymbol{\varepsilon}_n\right), \tag{31}$$

where $\mathbf{g}_n$ is the $n \times 1$ vector of (simulated) growth rates, α, β and ρ are parameters, and the $n \times 1$ vector of disturbances is generated as

$$\boldsymbol{\varepsilon}_n \sim N\left(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_n\right), \tag{32}$$

where σ^2 is a parameter.

Motivated by the empirical results in the previous section, I consider the parameter space to be combinations of β and ρ from the following ranges: $\beta \in \{.00, -.10, -.15, -.20, -.25, -.30\}$, $\rho \in \{.0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$. In all experiments I set $\alpha = 2$ and $\sigma^2 = .12$. The spatial weighting matrix in the data generating process is chosen to be the row normalized region size adjusted distance weights matrix, i.e. with elements

$$w_{ij}^* = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}, \tag{33}$$

where

$$w_{ij} = \left[\log\left(d_{\max}\right) - \log\left(d_{ij}\right)\right] \frac{p_j}{\sum_{k=1}^{n} p_k},$$

with p_k being the population of the region k.

Thus, I have 60 parameter combinations. For each parameter constellation, I use the same random numbers to generate the vector of disturbances $\boldsymbol{\varepsilon}_n$ one thousand times. Hence, I obtain 1,000 replications of the vector $\mathbf{g}_n$ for each of the 70 parameter designs. For a given replication of the data, I estimate the model under the different specifications of the spatial weighting matrix and calculate the biases (mean of the estimator over the 1,000 replications minus the true values) and root mean square errors of each of the estimators. The results for the two sets of experiments, corresponding to NUTS 1 and NUTS 2 samples, are reported in Tables A1-A14 in the appendix. Tables 3 and 4 below reports biases and root mean square errors for the different estimators averaged over the different parameter designs. Note that the designs where $\beta = 0$ are omitted in the calculation of the averages as under such specification the IV estimation procedure does not work. This is due to the fact that the IV estimators rely on instruments that are spatial lags of the exogenous variables. These are only valid when $\beta \neq 0$. When $\beta = 0$, the IV procedure should work badly and this is confirmed in the results.

TABLE 3

Average biases and RMSEs, NUTS 1 regions

Parameter	β		ρ^*		σ^2	
Model	Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	−0.007	0.012	*na*	*na*	−0.036	0.008
$\mathbf{W}^c$	−0.007	0.017	−0.451	0.301	−0.035	0.010
$\mathbf{W}^{cn}$	0.150	5.548	0.595	60.947	0.067	2.940
$\mathbf{W}^{c*}$	0.110	4.286	2.138	232.188	0.174	4.551
$\mathbf{W}^d$	−0.006	0.013	−0.450	0.286	−0.036	0.009
$\mathbf{W}^{dn}$	0.000	0.016	−0.102	0.942	−0.036	0.009
$\mathbf{W}^{d*}$	0.001	0.017	0.061	1.877	−0.035	0.012

TABLE 4

Average biases and RMSEs, NUTS 2 regions

Parameter	β		ρ^*		σ^2	
Model	Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	−0.010	0.008	*na*	*na*	−0.035	0.006
$\mathbf{W}^d$	−0.009	0.008	−0.450	0.285	−0.035	0.006
$\mathbf{W}^{dn}$	0.000	0.010	0.006	0.587	−0.035	0.006
$\mathbf{W}^{d*}$	0.001	0.011	0.049	0.587	−0.035	0.006

Observe that the $\beta-$convergence parameter is only modestly underestimated by procedures that incorrectly specify the space. However, the spatial autoregressive parameter is not estimated as significant by all procedures except those based on $\mathbf{W}^{dn}$ and $\mathbf{W}^{d*}$. Notice that the extent of convergence of the regions implied by the model with spatial dependence, depends on both $-\beta$ and ρ (see, for example, Egger and Pfaffermayr, 2006). Hence the estimators based on incorrectly specified space are significantly underestimating the amount of convergence in the data. Finally, I also note that the performance of the row normalized contiguity matrix $\mathbf{W}^{cn}$ can be very poor for some parameters.

These conclusion seem to be in line with the estimation results obtained using the actual European data. The spatial autocorrelation parameter was estimated to be even negative by under some specifications while the estimates of β were robust across the different specifications.

V. Conclusions and Suggestions For Future Research

The main message I hope to have illustrated is that one should pay careful attention to the specification of the relevant economic space in empirical applications using spatial econometric methods. I have examined what happens when sample size changes through redefinition of (some of) the spatial units and I have found that only spatial weights that are adjusted for the size of each observation (region) are sensible and can be expected to be robust to such sample size changes. I then look at the performance of the different specifications of space on the European data as well as on artificial data in the context of estimating $\beta-$convergence regression. I find that only normalized (row or population) distance weights provide estimates that do not change dramatically when the sample size is increased from 72 to 206. These conclusions are underpinned by the Monte Carlo experiments that use the same design as the empirical equations, in particular the same space as well as exogenous variables.

In the future research, it seems to be necessary to conduct more extensive Monte Carlo experiments and check the performance of the different estimators under different

data generating processes (using different definitions of space as the true space that generates the data). In terms of empirical research, the estimations contained in this paper are only to be taken as illustrative and it would be of interest to estimate a more comprehensive conditional $\beta-$convergence regression where other appropriate exogenous variables are added to the regression. Finally, given the importance of the correct space one would like to consider not only geographical space in the regression but also space based on economic distance among the regions. Such economic distance should be based on factors such as the strength of economic links among the regions (size of the trade or investment flows, size of investment exposure, etc.) as well as cultural links (common language, common legal tradition, etc.). At the moment the lack of data on the regional level makes such specification to be infeasible. Given that some data exists on national level, an intermediate step would be to consider the national border effect where the 'size' of the border is related to such variables.

References

[1] Anselin, L. (1998). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Boston, MA.

[2] Badinger, H, Müller, W.G. and Tondl, G. (2004). 'Regional Convergence in the European Union, 1985-1999: A Spatial Dynamic Panel Analysis,' *Regional Studies*, Vol. **38**, pp. 241-253.

[3] Barro, R. and Sala-i-Martin, X. (2004). *Economic Growth*. 2nd edition, MIT, Cambridge, MA.

[4] Beck, G.W. and Weber, A.A. (2001). 'How wide are european borders? on the integration effects of monetary unions,' *CFS Working Paper* No. 2001/07.

[5] Bode, E. and Rey, J.S. (2006). 'The spatial dimension of economic growth and convergence,' *Papers in Regional Science*, Vol. **85**, pp. 171-176.

[6] Egger, P. and Pfaffermayr, M. (2006). 'Spatial convergence,' *Papers in Regional Science*, Vol. **85**, pp. 199-215.

[7] Engel, C. and Rogers, J.H. (1996). 'How wide is the border?' *American Economic Review,* Vol. **86**, pp. 1112–1125.

[8] Kelejian, H. and Prucha, I.P. (1998). 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances,' *Journal of Real Estate Finance and Economics*, Vol. **17**, pp. 99-121.

[9] Kelejian, H. and Prucha, I.P. (2001). 'On the Asymptotic Distribution of the Moran I Test Statistic with Applications,' *Journal of Econometrics*, Vol. **104**, pp. 219-257.

[10] Le Gallo, J. and Cern, E. (2003). 'Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995,' *Papers in Regional Science*, Vol. **82**, pp. 175-201.