

# The role of spatial scale in regional convergence: the effect of MAUP in the estimation of $\beta$ -convergence equations

Alberto Díaz Dapena<sup>1</sup> · Esteban Fernández Vázquez<sup>1</sup> ·  
Fernando Rubiera Morollón<sup>1</sup>

Received: 31 December 2014 / Accepted: 4 February 2016 / Published online: 18 February 2016  
© Springer-Verlag Berlin Heidelberg 2016

**Abstract** Empirical analysis of regional convergence is normally based on data collected at a geographical scale corresponding to states or large regions (NUTS-2 or NUTS-3 for the case of Europe). However, it could be more realistic to consider that the dynamics generating economic growth take place at a smaller spatial scale. Potential heterogeneity across local areas might be not correctly quantified if the analysis is made at an aggregated geographical scale, which produces the so-called modifiable areal unit problem (MAUP). The objective of this paper is to explore to which extent MAUP has an effect on convergence analysis, in particular in the empirical estimation of  $\beta$ -convergence equations. First, we show how aggregation of spatial data can generate a problem of bias in the OLS estimator of  $\beta$ -convergence equations from cross-sectional data, as well as inflating its variance. Second, by means of a numerical simulation, we quantify the effect of geographical aggregation on the estimates of  $\beta$ -convergence. Our experiment is based on real spatial structures of aggregated and disaggregated data for different countries, and it numerically illustrates how a modification in the spatial scale has a significant effect on this type of studies.

**JEL Classification** R11 · R12

## 1 Introduction

The study of convergence in GDP per capita, or similar variables, among territories has been one of the central issues in the literature on economic growth and regional economics. This interest is logical from the point of view of the design of economic

---

✉ Fernando Rubiera Morollón  
frubiera@uniovi.es

<sup>1</sup> Regional Economics Laboratory – REGIOlab, University of Oviedo, Oviedo, Spain

policy as well as in the arena of economic theory discussion, because it is fundamental to find empirical evidence about where and under what conditions it is possible to observe processes of economic convergence or divergence. For example, identifying patterns of divergence or very slow convergence among European territories would provide the empirical evidence to support active and expensive territorial cohesion policies, such as the European Union Cohesion Policy, which is now the most expensive policy in the EU budget. From an academic point of view, the measurement of convergence processes in different scenarios implies finding an empirical evidence for the hypothesis of decreasing returns while divergence means the rejection of this hypothesis giving evidence in line with endogenous growth models or the framework of urban and regional economics.

There are different ways of studying convergence among territories, but the so-called  $\sigma$  and, especially,  $\beta$ -convergence are the approaches most commonly applied.  $\sigma$ -Convergence is perhaps the simplest approach. It basically quantifies the dispersion of income per capita or a similar variable in different moments along time: if the standard deviation of the variable of interest decreases along time, this is considered as an indication of convergence. This kind of analysis is usually conducted as an exploratory or preliminary analysis in the study of convergence. The seminal paper of Baumol (1986) introduced the concept of  $\beta$ -convergence: the relation between the growth rate of a particular economic variable during a period of time with the initial level of that variable in a set of territories—countries or regions. The literature on the econometrics for estimating  $\beta$ -convergence has been growing in the last few years. Islam (2003) or Magrini (2004) present surveys of this literature classifying the studies of  $\beta$ -convergence into different types of approaches: (1) panel data, (2) time series, (3) club convergence and (4) spatial dependence.

In general, the literature on  $\beta$ -convergence does not pay much attention to the level of spatial disaggregation on which the data are observable. This could be partially explained by a practical reason: the lack of information at a detailed spatial scale for many economies. For instance, in the case of European countries information on value added or income is normally available only at the scale of NUTS-2 or NUTS-3 administrative regions, which are constructed as the aggregation of a range of smaller areas of different characteristics. Additionally, a more theoretical reason justified that the role of the spatial scale of the data was neglected: the neoclassical framework on which the initial models were built on did not consider this issue as important. Oppositely, alternative approaches for modelling economic growth explicitly consider the role played by variables at a local level, where agglomeration economies and other centripetal forces have an effect at a sub-regional level. As a consequence, the aggregation of spatially disaggregated data into larger regions could cast some doubts on the empirical evidence found in convergence studies based on spatially aggregated data if regions are characterized by a high degree of sub-regional heterogeneity.

This paper studies the consequences of aggregation of spatial data in convergence analysis. More specifically, we aim at quantifying the effect of neglecting small-scale processes derived from estimating  $\beta$ -convergence equations based on spatially aggregated data. Our research bases on previous studies that have already called the attention to the effect of the aggregation, like in the work by Theil (1954) for the general case

on linear regression models or, more recently, by [Arbia and Petrarca \(2011\)](#) for the case of spatially dependent data.

The paper is structured as follows. Section 2 reviews the literature on economic growth, particularly on how the spatial scale plays a role on this literature. Section 3 derives the properties of ordinary least squares (OLS) estimators of  $\beta$ -convergence equations from cross-sectional data, and Sect. 4 quantifies this effect by means of numerical simulations applied to different structures of spatial data. Finally, Sect. 5 closes the paper with some remarks and potential future research lines and possible econometric solutions for the empirical analysis.

## 2 The relevance of the spatial scale in the regional convergence analysis

The neoclassical economic growth theories are mainly based on the role of the decreasing returns in the different production factors. Solow's model ([1956](#)) concludes that, in the long run, all territories will converge to the same level of GDPpc, provided that we have taken into account the relevant factors of an economy and that there exist decreasing returns in the production factors. This model also predicts that there is a constant growth of GDPpc in the steady state, which is equal to the technological growth. The  $\beta$ -convergence translates this theoretical framework into a simple empirical equation: the relation between the growth rate of a particular economic variable during a period with the initial level of that variable in a set of territories. When another regressor is considered it is absolute  $\beta$ -convergence, whereas if other explanatory variable is included, it refers to a conditional  $\beta$ -convergence analysis. This estimation framework allows for testing if poorer areas grow faster or not than the richer ones. If the parameter  $\beta$  is estimated with a negative sign, this indicates that lower levels of income per capita produce higher growth rates, leading to a process of convergence in the long run. A positive estimate of  $\beta$  would reveal a process of divergence. Under this approach, the spatial scale is not relevant because the logic of decreasing returns operates in the same way in all spatial scales.

Alternative approaches in the literature on economic growth, however, pointed out again the relevance of the spatial scale in the empirical analysis by taking into account the presence of local processes of endogenous growth, as well as the relevance of the spatial scale and agglomeration economies. For instance, [Myrdal \(1957\)](#), [Boudeville and Montefiore \(1968\)](#) or [Dixon and Thirlwall \(1975\)](#), among others, highlight the importance of cumulative processes in rich territories due to the movements of capital and workers, which makes them even more attractive while the opposite situation happens to poor places. [Romer \(1990\)](#) developed the model of endogenous technological change, which was later extended by [Mankiw et al. \(1992\)](#) considering human capital as a relevant factor. These models argue that endogenous growth takes place mainly at the local level. Additionally, a vast body of literature also pays attention to the role of the scale—economies of scale—and agglomerations—economies of agglomerations, starting from the contributions by [Marshall \(1920\)](#). The gains derived from large-scale production and from positive externalities associated with size lead to the concentration of economic activity in central locations from where the largest possible market is accessible. Additionally, more recent literature also stresses the positive link

between productivity and the presence of a diversified, highly qualified and versatile labour pool in large cities (Duranton and Puga 2000; Glaeser 1994, 1998; or Quigley 1998). In line with all this literature, it is possible to identify central and peripheral areas within regions, which is one of the essential concepts in the New Economic Geography (NEG) models (Krugman 1991; Krugman and Venables 1995 and Fujita and Krugman 1995). According to this literature: (1) there are incentives to largely concentrate the production in central areas; and (2) the intra-regional and inter-country processes of specialization and trade reinforce the processes of concentration and, in consequence, of divergence. Under a NEG approach, cities and metropolis—local areas—are in the centre of the analysis, drawing the attention to cities as the missing link between the macroeconomic theories of growth and the spatial empirical analysis.

Summarily, approaches such as the classical regional economics or the NEG models have a more local-based perspective than their neoclassical counterparts, which pays no attention to spatial aggregation.

Besides the theoretical discussion on the appropriate spatial scale, from the perspective of the empirical estimation of  $\beta$ -convergence equations, the role played by the spatial scale on which models are estimated is equally interesting if the conclusions of the empirical analysis could partially depend on this scale. This issue is generally denominated as a modifiable areal unit problem (MAUP), and its consequences have been explored since the 1930s (see Gehlke and Biehl 1934), and later explained in detail by Openshaw and Taylor (1979) or Openshaw (1983). Basically, one of the effects of the MAUP—the so-called scale effect—refers to the aggregation bias that emerges if data are aggregated into larger units—for example, cities to regions.<sup>1</sup>

The study of the effect of data aggregation on the estimation of empirical models has a relative long tradition in economics. For instance, Theil (1954) already studied these effects for the case of linear regression models. More recently, Arbia and Petrarca (2011) explored the effects of aggregation in a scenario of special dependence in the data.<sup>2</sup> However, the estimation of  $\beta$ -convergence equations has some particularities that make the issue of data aggregation specially interesting. First, the literature generally focuses on the effects of data aggregation in linear models, while the usual functional forms applied for  $\beta$ -convergence equations are nonlinear. Moreover, the study of processes of convergence normally distinguishes between convergence between countries or between regions. While the definition of “country” is univocal, the definition of “region” is a more unclear concept—as argued previously—and several alternatives for grouping basic spatial units could be used to construct aggregated regions. This makes the empirical study of regional convergence to be at least partially conditioned by the particular configuration of regions on which the study is based.

To illustrate the role played by the spatial scale for regional convergence, a simple estimation of an absolute  $\beta$ -convergence equation has been made for the case of the European Union. Annual data on GDP per capita in Purchasing Power Standard (PPS) have been taken at the scale of NUTS-3 regions, and absolute  $\beta$ -convergence

---

<sup>1</sup> Additionally, the zoning effect refers to the shape of the spatial units and the problem that a modification of their shape can also change any empirical result. However, in this paper we just pay attention to the problem of aggregation.

<sup>2</sup> Also see Rey and Montouri (1999) and Janikas and Rey (2005).

**Table 1**  $\beta$ -Convergence equation for the EU-28 (2000–2011) for different spatial scales

	NUT-1	NUT-2	NUT-3
$\beta^a$	-0.31***	-0.26***	-0.23***
Constant	3.36***	2.85***	2.52***
$\lambda^b$ (%)	3.37	2.74	2.38
$R^2$	57.72	44.82	36.44
$N$	98	272	1305

\*\*\* Estimates significantly different from zero at 1 %

<sup>a</sup> Source: Eurostat REGIO database, ESA-1995. Parameter obtained from the estimation of a unconditional  $\beta$ -convergence equation like Eq. (1)

<sup>b</sup> The speed of convergence ( $\lambda$ ) is obtained from the following expression:  $\lambda = \frac{-\ln(1+\beta)}{T} 100$ , being  $T$  the number of years

equations were estimated for different definitions of regions, namely NUTS-1, NUTS-2 and NUTS-3. The dependent variable is the growth rate of GDP per capita between 200 and 2011 to be regressed on the (log of) GDP per capita in 2000. A summary of the results is reported in Table 1.

Results in Table 1 show how the estimate of the  $\beta$  parameter at the scale of large NUTS-1 regions is remarkably higher than if the equations were estimated at the scale of NUTS-2 or NUTS-3 regions. Paying attention not only to the estimates of  $\beta$  parameters, but to the speed of convergence, it ranges between 2.38 (NUTS-3) and 3.38 % (NUTS-1) again depending on the specific definition of region applied. As a consequence, the time required to reduce the regional differences in the EU to one half of their initial levels—the so-called half-life, would be of around 20 years if the regions are defined as NUTS-1 units but approximately 30 if they were as NUTS-3.

This type of issues on regional convergence analysis has deserved some attention in previous empirical literature. Miller and Genc (2005), for example, estimated  $\beta$ -convergence equations under several possible spatial divisions for the US aggregating data available at county level, finding only a very minor effect of the scale on their results. More recently, Resende (2011) based on data collected at several spatial scales for the case of Brazil finding that their results were heavily conditioned by the specific criterion used to form regions: by using data grouped by Brazilian states he estimated a significant and negative  $\beta$  parameter, but the conclusion was the opposite when the  $\beta$ -convergence equations were estimated at a municipal scale. Even when these studies are interesting, they are limited to specific cases and particular periods of time, which limits the possibilities of drawing any general conclusions from them. The next section studies analytically the estimation of  $\beta$ -convergence equations and the properties of a least squares estimator of equations from spatially disaggregated and aggregated data.

### 3 The effect of the aggregation on the OLS estimation of $\beta$ -convergence equations with cross-sectional data

The literature studying the empirics of estimating of  $\beta$ -convergence equations started with the cross-sectional analyses of Baumol (1986), Barro (1991), Barro and Sala i

Martin (1991) or Mankiw et al. (1992), to later accommodate estimators capable to exploit panel-data structures as proposed by Islam (1995) or Lee et al. (1997).<sup>3</sup>

While panel-data estimators are the type of estimation strategy most commonly followed by far in the context of analysing country data, in the context of regional analysis is not uncommon to base the estimation of  $\beta$ -convergence equations on cross-sectional data due to information availability (see, e.g., Azzoni 2001, for Brazil; Rodríguez-Pose and Sánchez-Reaza 2002, for Mexico; Cuadrado 2001, for Europe; or Raiser 1998, for China). This section studies the properties of a traditional ordinary least squares (OLS) estimator of  $\beta$ -convergence equations based on a cross section of data.

Let us assume an economy that is divided into different spatial units that are created according to several criteria for geographical aggregation. More specifically, suppose that the economy is divided into  $i = 1, \dots, n$  basic spatial units—municipalities or cities—that are aggregated into  $j = 1, \dots, m (m < n)$  groups—regions. In line with the ideas of New Economic Geography and endogenous growth theories, we assume that the process of income generation takes place at the basic spatial scale of  $n$  units. This section studies the effects on the conclusions of convergence analysis depending on the scale at which the outcome data are observable: directly observable at the original scale ( $n$  local places) or at the aggregated scale ( $m$  regions). If the conclusions about the coefficient depend on the level of aggregation, this will be a signal that a potential MAUP is somehow “contaminating” our analysis.

Our starting point will be the formulation developed in Arbia and Petrarca (2011) for the case of cross-sectional data in a linear regression model that are generated at a given spatial level, but then observed at a more aggregate scale. The following equation describes the model to be estimated at a disaggregated scale with  $n$  spatial units:

$$y = X\beta + u \tag{1}$$

where  $y$  is the  $(n \times 1)$  vector with the dependent variable,  $X$  is a  $(n \times K)$  matrix with the  $K$  regressors considered in the equation,  $\beta$  is the  $(K \times 1)$  vector with the parameters to be estimated and  $u$  is the typical  $(n \times 1)$  disturbance, which is assumed to distribute normally around zero with a constant variance  $\sigma^2$ . If the data of the  $n$  units are aggregated at a higher geographical scale with  $m$  locations, the new data set is defined by:

$$y^* = Gy \tag{2}$$

$$X^* = GX \tag{3}$$

$$u^* = Gu \tag{4}$$

Being  $G$  the aggregation matrix with dimensions  $(m \times n)$ , including elements like:

$$G = \begin{bmatrix} g_{11} & \dots & g_{1r_1} & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & g_{21} & \dots & g_{2r_2} & 0 & \dots & 0 \\ & \dots & & & & & & \dots & \\ 0 & \dots & 0 & \dots & & & g_{m1} & \dots & g_{mr_m} \end{bmatrix} \tag{5}$$

<sup>3</sup> For a recent review of advanced estimation strategies, see Elhorst et al. (2010).

where each row indicates that the original data are aggregated—grouped—into  $m$  different locations, being the number of original spatial units differently aggregated in each case  $(r_1, r_2, \dots, r_m)$ .

In this context, the aggregated equation is defined as:

$$y^* = X^* \beta^* + u^* \tag{6}$$

where

$$E(u^*) = E(Gu) = \mathbf{0} \tag{7}$$

$$Var(u^*) = E(u^*u^{*'}) = E(Guu'G') = GG'\sigma^2 \tag{8}$$

In their paper, [Arbia and Petrarca \(2011\)](#) deal with the specific case of perfect aggregation where the elements of this aggregation matrix  $G$  are unitary values:

$$G = \begin{bmatrix} 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 \\ & \dots & & & & & \dots \\ 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{bmatrix} \tag{9}$$

Being the number of ones in every row always equal to  $r = m/n$ . They show how the OLS estimator of  $\beta^*$  ( $\hat{\beta}^*$ ) of equation (6) is an unbiased estimator of  $\beta$  in the original equation (1), being the variance of the OLS estimator in the aggregated equation (6) bigger than the original variance of the OLS estimator in (1):

$$\begin{aligned} E(\hat{\beta}^*) &= E\left([X'^* X^*]^{-1} X'^* y^*\right) = E\left([X' G' G X]^{-1} X' G' G y\right) \\ &= E\left([X' X]^{-1} X' y\right) = E(\hat{\beta}) = \beta \end{aligned} \tag{10}$$

$$Var(\hat{\beta}^*) = GG'\sigma^2 [X' G' G X]^{-1} > Var(\hat{\beta}) \tag{11}$$

In other words, the scale effect does not represent a problem of bias, although it generates an efficiency problem.

The  $\beta$ -convergence equations, however, are not characterized by this same response to the scale effect, due to some particularities in the aggregation scheme of the dependent and the independent variables and the logarithmic form of the equation. In order to justify this claim, let us state the typical absolute  $\beta$ -convergence equations estimated for a cross section of  $n$  spatial units as:<sup>4</sup>

$$\begin{aligned} \ln\left(\frac{y_{it}}{y_{i0}}\right) &= \alpha + \beta \ln(y_{i0}) + u_{it}; \text{ or} \\ \ln(y_{it}) &= \alpha + (1 + \beta) \ln(y_{i0}) + u_i \end{aligned} \tag{12}$$

<sup>4</sup> A similar exercise could be done for conditional  $\beta$ -convergence equations just by adding more regressors to this basic equation. We have opted for working with this simple case for the sake of simplicity, but the main conclusions in terms of the effects of aggregation on its estimation, however, would hold.

where the growth in an economic indicator  $y$  as GDP or income, value added, etc., per capita between periods 0 and  $t$  in location  $i$  regressed on the logs of the initial variable per capita ( $y_{i0}$ ) on the same location. One problem with aggregated data for estimating equations like (12) is that the nonlinearities in the dependent and explanatory variables are not compatible with the equivalences between the aggregated and disaggregated equation. More specifically, the aggregate version of the absolute  $\beta$ -convergence equations equation will be:

$$\begin{aligned} \ln\left(\frac{y_{jt}^*}{y_{j0}^*}\right) &= \alpha^* + \beta^* \ln(y_{j0}^*) + u_{jt}^*; \text{ or} \\ \ln(y_{jt}^*) &= \alpha^* + (1 + \beta^*) \ln(y_{j0}^*) + u_{jt}^* \end{aligned} \tag{13}$$

Being:

$$\begin{aligned} \mathbf{y}_0^* &= \mathbf{G}\mathbf{y}_0 & (14) \\ \mathbf{u}^* &= \mathbf{G}\mathbf{u} & (15) \end{aligned}$$

Matrix  $\mathbf{G}$  represents the aggregation scheme for the initial values per capita, with a typical element  $g_{ij}$  indicating the population share of the basic spatial unit  $i$  on the aggregated location  $j$  measured in the initial period. In contrast to the type of equations aggregated as in (6), the dependent variable of the equation estimated with aggregate data is given by:

$$\ln(\mathbf{y}_t^*) = \ln(\mathbf{H}\mathbf{y}_t) \neq \mathbf{G}\mathbf{y}_t \tag{16}$$

where  $\mathbf{H}$  is the aggregation matrix where a typical element  $h_{ij}$  indicates the population share of the spatial unit  $i$  on region  $j$  measured in the final period. In general, this matrix is not necessarily equal to  $\mathbf{G}$ , given that the elements of  $\mathbf{H}$  are the population shares in the final period and the populations in each period can be different.

Note that Eq. (10) states that the expected value of the OLS estimator with aggregated data is given by  $E\left([\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{y}\right)$  and it is equal to  $\beta$ , while a different aggregation scheme would modify the form of the estimator being its expected value  $E\left([\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{G}'\mathbf{H}\mathbf{y}\right)$ . When the elements of matrix  $\mathbf{H}$  are larger than the elements of  $\mathbf{G}$ , the estimator will present a positive bias, while a negative bias will be the consequence of the elements of  $\mathbf{H}$  being smaller than those in  $\mathbf{G}$ . The comparison between these two matrices can be made in terms of the Euclidean norms of their row vectors, comparing  $\sqrt{\mathbf{h}'_j\mathbf{h}_j}$  with  $\sqrt{\mathbf{g}'_j\mathbf{g}_j}$ . These norms would account for the concentration of population shares on each region  $j$ —they can be interpreted as a Herfindahl index for the distribution of population in region  $j$ . If population in the final period is more unequally distributed than in the initial period and, in general,  $\sqrt{\mathbf{h}'_j\mathbf{h}_j} \geq \sqrt{\mathbf{g}'_j\mathbf{g}_j}$  this would lead to a positive bias in the estimation of  $\beta$ . The opposite situation will happen when the population in the final period is more evenly distributed within regions than in the initial period.



Even if the aggregation criterion reflected in  $H$  was the same as the aggregation scheme present in matrix  $G$ , an additional problem derived for the nonlinear nature of the  $\beta$ -convergence equation will be present, affecting the properties of the OLS estimation from aggregated data. Assuming a case where  $G = H$ , note that  $\ln(y_t^*) = \ln(Gy_t) \neq Gy_t$ . This problem is the same with the matrix of explanatory variables  $X^*$  (which in the case of absolute  $\beta$ -convergence equations corresponds to the log of the initial levels  $y_0^*$ ) given that  $\ln(y_0^*) = \ln(Gy_0) \neq G \ln(y_0)$ .<sup>5</sup> Specifically, we could argue that  $\ln(y_t^*) \leq H \ln(y_t)$  and  $\ln(y_0^*) \leq G \ln(y_0)$  basing on Jensen's inequality. These inequalities imply that Eqs. (10) and (11) do not hold, affecting the expected value and the variance of the OLS estimator of an aggregate equation as (6). The dependent variable  $y_t^*$  in the case of  $\beta$ -convergence equations with aggregated data is  $\ln(Hy_t)$ , being the matrix of regressors  $X^*$  given by  $\ln(GX)$ . The expected value and the variance of the OLS estimator for this aggregated equation are, respectively:

$$E(\hat{\beta}^*) = E\left([X'^* X^*]^{-1} X'^* y_t^*\right) = E\left([\ln(GX)'\ln(GX)]^{-1} \ln(GX)'\ln(Hy_t)\right) \neq E\left([X'G'GX]^{-1} X'G'Gy_t\right) \neq \beta \tag{17}$$

$$\begin{aligned} Var(\hat{\beta}^*) &= Var\left([X'^* X^*]^{-1} X'^* y_t^*\right) \\ &= Var\left([\ln(GX)'\ln(GX)]^{-1} \ln(GX)'\ln(Hy_t)\right) \geq \sigma^2 [X'G'GX]^{-1} \\ &\geq Var(\hat{\beta}) \end{aligned} \tag{18}$$

The result in (18) is equivalent to (11), indicating the augmenting effect of the aggregation on the variance of the estimator. However, Eq. (17) shows how a problem of bias emerges now as well, in contrast to the result in (10).<sup>6</sup> The scale effect in the estimation of the  $\beta$ -convergence equations leads, in summary, to estimates that can be biased and with higher variance than in the original disaggregated equations. The next section of the paper explores by means of a numerical simulation the empirical implications of this problem.

### 4 Convergence with spatially disaggregated and aggregated data: some numerical experiments

Once the effect of the aggregation level on the OLS estimator has been studied, it is important to quantify its consequences when applied to the empirical analysis of  $\beta$ -convergence. A numerical experiment is conducted in this section with this purpose in

<sup>5</sup> For the sake of clarity in the exposition, in the remaining of this section we refer to the matrix of potential regressors  $X$  included as explanatory variables in the specification of a general  $\beta$ -convergence equation. Absolute  $\beta$ -convergence equation only considers initial values  $y_0$  in matrix  $X$ .

<sup>6</sup> The positive or negative sign of the bias depends on the aggregation schemes represented on matrices  $G$  and  $H$ —because of the per capita nature of the dependent and explanatory variables—and it is not straightforward, since their elements are influenced by the population dynamics of the spatial units aggregated into larger regions. The issue of the logarithmic transformation adds more complexity to the study of the bias. Details are provided in the “Appendix”.

mind. Our experiment assumes that the data are generated at the level of  $i = 1, \dots, n$  basic spatial units by the following equation that determines the growth in the relevant variable as:

$$\begin{aligned} \dot{y}_i &= \alpha + \beta \ln(y_{i0}) + u_i; \text{ or} \\ \ln(y_i) &= \alpha + (1 + \beta) \ln(y_{i0}) + u_i \end{aligned} \quad (19)$$

being  $y_{i0}$  the value of the relevant variable at the starting period and  $y_i$  its final value. In the experiment, we have arbitrarily set the value of the intercept  $\alpha$  at 1.1, and  $u \sim N(0, 0.5)$ . The idea is to compare the OLS estimates of parameter  $\beta$ , which is the key element in the analysis of  $\beta$ -convergence, in two situations that vary on the spatial scale on which the data are observed:

1. the reference situation or benchmark, that assumes that we have data observable at the same scale at which they are generated, i.e., for the  $i = 1, \dots, n$  basic spatial units
2. a case where the data are only observable at an aggregated spatial scale into  $j = 1, \dots, m$  units. In this second scenario, we assume that we only have data on  $y_j^*$  and  $y_{j0}^*$  and from them we estimate the parameters of the equation:

$$\ln(y_j^*) = \alpha + (1 + \beta) \ln(y_{j0}^*) + u_j^* \quad (20)$$

In order to have a numerical experiment as realistic as possible, we have taken as reference for simulating possible structures of aggregation of spatial data the real sub-regional and regional divisions in three different countries: namely the USA, Germany and Chile. These three countries are taken as examples of developed economies, each of them presenting a particular configuration in their regional divisions. For example, the basic spatial units for the case of Chile are the *comunas* ( $n = 100$ ) that form the total of  $m = 13$  administrative regions. Similarly, in Germany we can find the basic spatial units defined by the concept of *kreise* ( $n = 393$ ) that are aggregated into  $m = 14$  *länders*. Finally, the USA is divided into  $n = 3088$  counties that are aggregated forming the  $m = 50$  states.

In order to provide with sensible values to the growth equation depicted in (20), we have taken real data for the initial value of the variable of interest. In the cases of the USA and Chile, we have defined  $y_{i0}$  as the income per capita, while in the case of Germany—due to data availability at the desired spatial scale—it is defined as GDP per capita. The time span on which we estimate (20) is also different for each country and conditioned by data limitations: for the USA there is a series of income at county level from 1969 to 2011 published by the Bureau of Economic Analysis; in Chile we have data on income for the *comunas* between 1996 and 2006 available in the *Casen Survey* of the Ministry of Planning; and for Germany the *Destatis Statistisches Bundesamt* contains estimates of GDP for the *kreise* between 2000 and 2011. Additionally, data on population are required to have indicators of income or GDP per capita. We have opted for using real data on population as well. Note that data of population in the initial and the final periods are required in order to aggregate spatially the per capita values of the variable of interest. The values per capita in the initial and final periods—the

**Table 2** Summary statistics for the different scenarios

	Germany (2000–2011) 393 <i>Kreise</i> , 14 <i>Länders</i>	USA (1969–2011) 3088 Counties, 50 states	Chile (1996–2006) 100 <i>Comunas</i> , 13 Regions
$x_0$	23,406 (GDP p.c., € 2000)	3024 (Income p.c., \$ 1969)	348,419 (Income p.c., CLP 1996)
Initial population (means across $n$ units)	191,189	64,904	89,950
Final population (means across $n$ units)	189,926	100,606	107,197

explanatory and dependent variable in (20), respectively—are aggregated by weighting the values in levels at the scale of basic spatial units by their population shares on these periods. Summary statistics of all these variables are given in Table 2.

All these pieces of information have been used for the data generating process described in Eq. (19). The key element on this equation is the parameter  $\beta$ , whose value determines if we have a process of convergence—if negative—or divergence—if positive. In the experiment, different scenarios have been considered depending on the value of parameter  $\beta$ , setting its values ranging between  $-0.3$  and  $0.3$ . For each value of the parameter and for each country, we have simulated 5,000 trials and we have estimated the parameter by applying OLS in scenarios (1) and (2).

Table 3 summarizes the results obtained on each case, reporting the true value of the parameter together with the average OLS estimate, the empirical variability of the estimates—standard deviation—and a measure of deviation—mean squared error—between the true values and the OLS estimates.

Additionally, Fig. 1 visually illustrates the results of the simulations reported in Table 3. In these plots the x axis represents the true value of the  $\beta$  parameter considered in Eq. (20). For each value of  $\beta$ , the mean estimate obtained in the 5000 trials using disaggregated or aggregated data is represented in the y axis. If the results were not biased, we would expect a  $45^\circ$  line crossing the origin of the two axes with the true values and the estimates. 95% confidence bandwidths are also plotted, based on the normal distribution of the estimates.

As expected, the empirical variability of the OLS estimates are substantially lower when estimated from the  $n$  basic data points than in the case of the  $m$  aggregated spatial units, since the sample size are smaller when working with aggregate data. Not surprisingly, these differences are more remarkable for the case of the USA when compared to the other two countries in the experiment, given that the ratio  $r = m/n$  is much smaller for the USA. The loss of efficiency derived from estimating equation (20) with  $m$  aggregated regions instead of estimating (19) with  $n$  spatial units is not entirely produced, however, by this inflation of the variance. One substantial part can be attributed to the bias as stated in Eq. (17). The estimates based on aggregated data

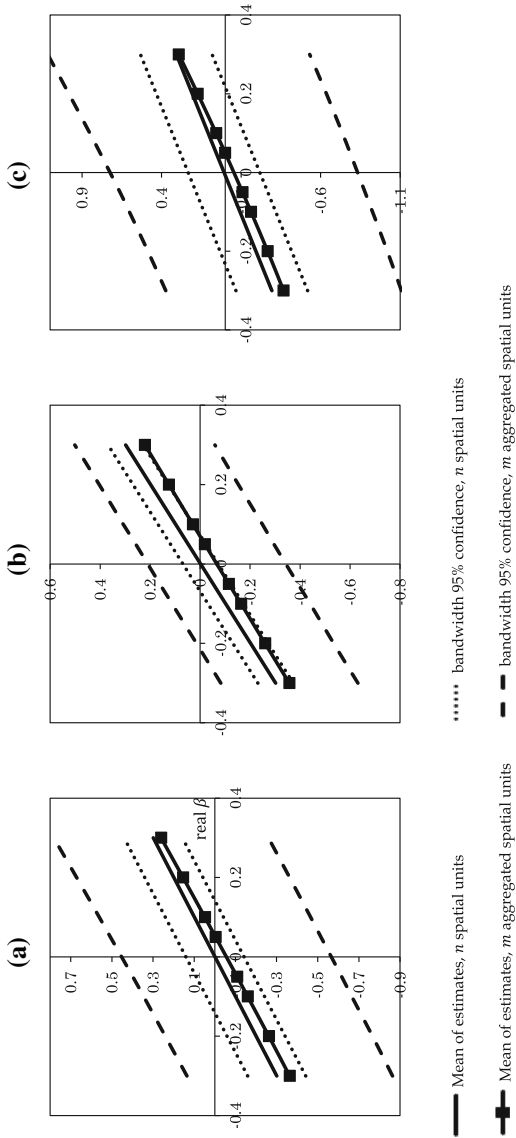
**Table 3** Results of an OLS estimation with different spatial configurations (5000 trials)

True $\beta$	OLS estimates					
	Germany (2000–2011)		USA (1969–2001)		Chile (1996–2006)	
	$n = 393$	$m = 14$	$n = 3088$	$m = 50$	$n = 100$	$m = 13$
-0.30	-0.298	-0.360	-0.301	-0.351	-0.299	-0.398
	(0.073)	(0.256)	(0.037)	(0.142)	(0.114)	(0.379)
	[0.005]	[0.069]	[0.001]	[0.023]	[0.013]	[0.153]
-0.20	-0.202	-0.258	-0.201	-0.254	-0.199	-0.298
	(0.073)	(0.258)	(0.037)	(0.143)	(0.114)	(0.384)
	[0.005]	[0.070]	[0.001]	[0.023]	[0.013]	[0.157]
-0.10	-0.098	-0.155	-0.099	-0.157	-0.099	-0.195
	(0.073)	(0.259)	(0.037)	(0.143)	(0.114)	(0.389)
	[0.005]	[0.070]	[0.001]	[0.024]	[0.013]	[0.160]
-0.05	-0.049	-0.103	-0.051	-0.109	-0.051	-0.142
	(0.073)	(0.260)	(0.037)	(0.144)	(0.114)	(0.392)
	[0.005]	[0.071]	[0.001]	[0.024]	[0.013]	[0.162]
0.050	0.051	0.001	0.051	-0.012	0.051	-0.033
	(0.073)	(0.262)	(0.036)	(0.144)	(0.114)	(0.399)
	[0.005]	[0.071]	[0.001]	[0.025]	[0.013]	[0.166]
0.1	0.100	0.054	0.099	0.037	0.102	0.023
	(0.073)	(0.263)	(0.037)	(0.145)	(0.114)	(0.402)
	[0.005]	[0.071]	[0.001]	[0.025]	[0.013]	[0.168]
0.2	0.201	0.160	0.199	0.132	0.201	0.138
	(0.073)	(0.265)	(0.037)	(0.145)	(0.114)	(0.410)
	[0.005]	[0.072]	[0.001]	[0.026]	[0.013]	[0.172]
0.3	0.302	0.267	0.301	0.225	0.301	0.257
	(0.073)	(0.268)	(0.036)	(0.144)	(0.114)	(0.418)
	[0.005]	[0.073]	[0.001]	[0.026]	[0.013]	[0.176]

Average estimates are reported for each true value of parameter  $\beta$ . Empirical standard deviations are shown in parentheses. Mean squared errors between true values and estimates are shown in brackets

present a negative bias underestimating the true value of the  $\beta$  parameter. The negative bias is partially a consequence of populations generally more uniformly distributed within each type of aggregated region (US states, German *land*ers or Chilean *regiones*) in the final period (2011 for the USA and Germany and 2006 for Chile) than in the initial one (1969 for the USA, 2000 for Germany and 1996 for Chile).

Although the simulations have been made for countries with different characteristics and spatial configurations, the results seem to be robust. As expected, the mean of the OLS estimates with  $n$  data points are practically equal to the true coefficient. In contrast, for each value of the true parameter, the regression based on aggregate regions tends on average to estimates smaller than the real coefficient. The mean bias of the eight values set for parameter  $\beta$  in the simulation is  $-0.051$  for Germany,  $-0.061$  for the USA and  $-0.082$  for Chile. In summary, the effect produced by the aggregation



**Fig. 1** OLS estimator with local and aggregate data, 1000 replications. **a** Germany, 393 *Kreise*, 14 *Länder* (2000–2011). **b** USA, 3088 counties, 50 states (1969–2011). **c** Chile, 100 *Comunas*, 13 regions (1996–2006). *Solid line* mean of estimates,  $n$  spatial units. *Dotted line* bandwidth 95% confidence,  $n$  spatial units. *Square* mean of estimates,  $m$  aggregated spatial units. *Dashed line* bandwidth 95% confidence,  $m$  aggregated spatial units

of the spatial units in our experiments negatively biases the conclusions drawn from the OLS estimation of  $\beta$ -convergence equations.

## 5 Conclusions

The study of convergence is one of the more prolific research lines in the literature on regional economics. Conclusions derived from convergence analysis provide the support to maintain, reduce or increment expensive policies, such as the Regional Cohesion Policy in the EU. Different improvements have been proposed in the estimation techniques applied to quantify empirically the speed of convergence or divergence among territories. However, most of this empirical literature does not pay attention to how relevant could be the geographical scale in which the convergence is measured, although one of the most important differences among neoclassical theoretical equations and other alternative approaches is the spatial scale in which economic growth is studied.

The objective of this paper is to provide an evaluation of the empirical consequences on changes in the spatial scale in the most commonly used approach for convergence analysis: the estimation of equations of  $\beta$ -convergence. The characteristics of an OLS estimator applied to cross-sectional data—which is a relatively common situation in empirical studies—are derived. We found that geographical aggregation produces estimators with higher variance—part of it produced by the reduction in the sample size, but also biased if compared with the OLS estimator based on the original disaggregated spatial units.

To provide quantitative evidence about the effect of the spatial scale in  $\beta$ -convergence analysis, we conduct numerical simulations with different spatial configurations of real countries: Germany, USA and Chile. The results in the simulation confirm the loss of efficiency caused by the aggregation of spatial data, some of which is due to differences in sample size, but the negative bias generated is also significant. One important implication derived from our results is that the estimation of  $\beta$ -convergence equations based on aggregated data should take into account that an important part of the information, related with intra-regional dynamics, could be missing.

Our results, however, do not necessarily indicate that estimates of  $\beta$ -convergence equations with aggregated data are misleading or not useful: in some situations the availability of spatially disaggregated data is very limited and some type of aggregation is required. In addition, in economies where aggregate regions are characterized by low levels of intraregional heterogeneity, aggregation of spatial data could be not a real issue when dealing with convergence analysis. Our results, however, suggest that the spatial scale on which data are taken for estimating  $\beta$ -convergence equations should be carefully defined, since this specification can be partially affecting the conclusions of the analysis.

Our analysis opens the discussion about the suitability of econometric techniques that are not affected by MAUP problems. In this regard, multilevel estimation (see, among others, [Goldstein 1986, 2011](#), [Hox et al. 2010](#)), which allow for using data at

different scales is particularly interesting if we want to identify different spatial scales of convergence avoiding the potential bias derived from the data aggregation.

Finally, there are relevant issues not studied here that would require further research. For instance, this paper studied the MAUP effect on a simple OLS estimator with cross-sectional data. The proliferation of time series with regional data has made possible, however, applying estimators based on a structure of panel data. The consequences of spatial aggregation in the context of estimators applied to dynamic panels are an important issue that should be included in the research agenda on the estimation of  $\beta$ -convergence equations.

### Appendix: The bias of OLS estimation in $\beta$ -convergence equations from aggregated data

Equation (17) shows how an OLS estimation of  $\beta$ -convergence based on aggregated data can be affected by a problem of bias. This problem is caused by the differences in the aggregation matrices  $G$  and  $H$ , which, respectively, affect the values of the explanatory and dependent variables, and for the nonlinear nature of the  $\beta$ -convergence equations. We will show this basing on the basic formulation:

$$\ln(y_{it}) = \alpha + (1 + \beta) \ln(y_{i0}) + u_i \tag{21}$$

Considering vector  $y_0$ , which contains the initial values included as regressor in the  $\beta$ -convergence equation, Jensens's inequality states that  $\ln(Gy_0) \leq G \ln(y_0)$ . Note that it is possible to rewrite this inequality as:

$$\ln(Gy_0) = \hat{c}_0 G \ln(y_0) \tag{22}$$

where  $\hat{c}_0$  is a diagonal ( $m \times m$ ) matrix with a typical element  $\hat{c}_{0j}$  defined as:

$$\hat{c}_{0j} = \frac{\ln(g'_j y_{j0})}{g'_j \ln(y_{j0})} \leq 1 \tag{23}$$

In (23),  $g'_j$  refers to the (row) vector of matrix  $G$  that aggregates the initial values of  $y_0$  that belong to the aggregated region  $j(y_{j0})$ . Similarly, concerning the aggregation of the dependent variable, we can write:

$$\ln(Hy_t) = \hat{c}_t H \ln(y_t) \tag{24}$$

where the elements of the diagonal matrix  $\hat{c}_t$  are given by the expression:

$$\hat{c}_{tj} = \frac{\ln(h'_j y_{jt})}{h'_j \ln(y_{jt})} \leq 1 \tag{25}$$

Equation (17) can be consequently rewritten as:

$$\begin{aligned} E(\hat{\beta}^*) &= E\left([\mathbf{y}_0^{*'} \mathbf{y}_0^*]^{-1} \mathbf{y}_0^{*'} \mathbf{y}_t^*\right) = E\left(\left[\ln(\mathbf{G}\mathbf{y}_0)' \ln(\mathbf{G}\mathbf{y}_0)\right]^{-1} \ln(\mathbf{G}\mathbf{y}_0)' \ln(\mathbf{H}\mathbf{y}_t)\right) \\ &= E\left(\left[\ln(\mathbf{y}_0)' \mathbf{G}' \hat{\mathbf{c}}_0' \hat{\mathbf{c}}_0 \mathbf{G} \ln(\mathbf{y}_0)\right]^{-1} \ln(\mathbf{y}_0)' \mathbf{G}' \hat{\mathbf{c}}_0' \hat{\mathbf{c}}_t \mathbf{H} \ln(\mathbf{y}_t)\right) \end{aligned} \quad (26)$$

In a situation as the described in [Arbia and Petrarca \(2011\)](#), where the equation is linear ( $\hat{c}_{tj} = \hat{c}_{0j} = 1; j = 1, \dots, m$ ) and the aggregation scheme is the simple sum of spatial units ( $\mathbf{G} = \mathbf{H}$ ) makes (26) to be equal to Eq. (10) and the OLS estimator is unbiased.  $\hat{\beta}^*$  will be biased, however, in situations that depart from that baseline. The specification of a  $\beta$ -convergence equation as depicted in (21), with nonlinear relations and different aggregation schemes in the dependent variable and the regressor makes the OLS estimation biased, depending the sign of the bias on the relationship between the matrices  $\mathbf{G}$ ,  $\mathbf{H}$ ,  $\hat{\mathbf{c}}_0$  and  $\hat{\mathbf{c}}_t$ .

## References

- Arbia G, Petrarca F (2011) Effects of MAUP on spatial econometric models. *Lett Spat Resour Sci* 4(3):173–185
- Azzoni CR (2001) Economic growth and regional income inequality in Brazil. *Ann Reg Sci* 35(1):133–152
- Barro R J (1991) Economic Growth in a Cross-Section of Countries. *Q J Econ* 106(2):407–443
- Barro RJ, Sala i Martin X et al (1991) Convergence across states and regions. *Brook Pap Econ Act* 1:107–182
- Baumol WJ (1986) Productivity growth, convergence, and welfare what the long-run data show. *Am Econ Rev* 76:1072–1085
- Boudeville JR (1968) Problems of regional economic planning, vol 3. Edinburgh University Press, Edinburgh
- Cuadrado JR (2001) Regional convergence in the European Union: from hypothesis to the actual trends. *Ann Reg Sci* 35(3):333–356
- Dixon R, Thirlwall AP (1975) A model of regional growth-rate differences on Kaldorian lines. *Oxf Econ Pap* 27(2):201–214
- Duranton G, Puga D (2000) Diversity and specialization in cities: why, where and when does it matter? *Urban Stud* 37(3):533–555
- Elhorst JP, Piras G, Arbia G (2010) Growth and convergence in a multi-regional model with space–time dynamics. *Geogr Anal* 42:338–355
- Fujita M, Krugman P (1995) When is the economy monocentric? von Thünen and Chamberlin unified. *Reg Sci Urban Econ* 25(4):505–528
- Gehlke CE, Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *J Am Stat Assoc* 29(185):169–170
- Glaeser EL (1994) Why does schooling generate economic growth? *Econ Lett* 44:333–337
- Glaeser EL (1998) Are cities dying?. *J. Econ. Perspect* 12(2):139–160
- Goldstein H, (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73(1):43–56
- Goldstein H, (2011) Multilevel statistical models, Vol. 922.. John Wiley & Sons
- Hox, JJ, Moerbeek M, van de Schoot R (2010) Multilevel analysis: Techniques and applications. *Routledge*
- Islam N (2003) What have we learnt from the convergence debate? *J Econ Surv* 17(3):309–362
- Islam N (1995) Growth empirics: a panel data approach. *Q J Econ* 110(4):1127–1170
- Janikas MV, Rey SJ (2005) Spatial clustering, inequality and income convergence. *Reg Dev* 21:45–65
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99(2):483–499
- Krugman P, Venables AJ (1995) Globalization and the inequality of nations. *Q J Econ* 110:857–880
- Lee K, Pesaran MH, Smith R (1997) Growth and convergence in a multi-country empirical stochastic Solow model. *J Appl Econom* XII:357–392



- Mankiw NG, Romer D, Weil DN (1992) A contribution to the empirics of economic growth. *Q J Econ* 107(2):407–437
- Magrini S (2004) Regional (di)convergence. *Handb Reg Urban Econ* 4:2741–2796
- Marshall A (1920) *Principles of economics*, 8th edn. Macmillan, New York
- Miller JR, Genc I (2005) Alternative regional specification and convergence of U.S. regional growth rates. *Ann Reg Sci* 39:241–252
- Myrdal G (1957) *Economic theory and under-developed regions*. Duckworth
- Openshaw S (1983) *The modifiable areal unit problem*. GeoBooks, Norwich
- Openshaw S, Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed) *Statistical applications in the spatial sciences*. Pion, London, pp 127–144
- Quigley JM (1998) Urban diversity and economic growth. *J Econ Perspect* 12(2):127–138
- Raiser M (1998) Subsidizing inequality: economic reforms, fiscal transfers and convergence across Chinese provinces. *J Dev Stud* 34(3):1–26
- Resende GM (2011) Multiple dimensions of regional economic growth: the Brazilian case, 1991–2000. *Pap Reg Sci* 90(3):629–662
- Rey S, Montouri BD (1999) US regional income convergence: a spatial econometric perspective. *Reg Stud* 33(2):143–156
- Rodríguez-Pose A, Sánchez-Reaza J (2002) The impact of trade liberalization on regional disparities in Mexico. *Growth Change* 33:72–90
- Romer P (1990) Endogenous technological change. *J Polit Econ* 98(5):71–102
- Solow R (1956) A contribution to the theory of economic growth. *Rev Econ Stat* 39:312–320
- Theil H (1954) *Linear aggregation of economic relations*. North Holland Publishing, Amsterdam