Lett Spat Resour Sci (2011) 4:173–185
DOI 10.1007/s12076-011-0065-9

ORIGINAL PAPER

Effects of MAUP on spatial econometric models

Giuseppe Arbia · Francesca Petrarca

Received: 15 February 2011 / Accepted: 14 August 2011 / Published online: 29 September 2011

Abstract The modifiable areal unit problem refers to the modifications of any statistical analysis when changing the scale of observation (e.g. from region to countries) or the aggregation criterion (e.g. different partitions of one country at a given scale). In a previous work (Arbia in Spatial data configuration in the statistical analysis of regional economics and related problems, Kluwer, Dordrecht 1989), we analyzed the effects of the modifiable areal unit problem on statistical analysis without referring to any specific random field. The quoted work concentrated on the modification of the first and second order properties of a random field. In this paper we reanalyse the same problem by looking specifically at the effects of MAUP on linear spatial econometric models looking in particular at the SARAR (1,1) model.

Keywords Aggregation effect · Loss in efficiency · MAUP · SARAR models · Scale effect · Spatial econometric

JEL Classification C21 · C31 · C5

A previous version of this paper was presented at the North American Meetings of the Regional Science Association International (RSAI) sponsored by the North American Regional Science Association (NARSC), Denver, November 10–13, 2010. We wish to thank all participants for the comments received in that occasion.

G. Arbia (✉)
Department of Statistical, Technological and Environment Sciences, University “G. D’Annunzio” of Chieti-Pescara, Viale Pindaro 42, Pescara, Italy
e-mail: arbia@unich.it

F. Petrarca
Department of Economy “Federico Caffè”, University of “Roma Tre”, Via Silvio d’Amico 77, Rome, Italy

1 Introduction

The MAUP (Modifiable Areal Unit Problem) is a particular manifestation of the more general Modifiable Unit Problem discussed by Yule and Kendall (1950) and has a long tradition in Statistics and Geography. It manifests itself in two ways: the scale problem dealing with the indeterminacy of any statistical measure to changes in the level of aggregation of data, and the aggregation problem having to do with the indeterminacy of any statistical measures to changes in the aggregation criterion at a given spatial scale. A full understanding of the effects of MAUP is relevant in many statistical issues: to assist the choice of the scale of analysis and of the spatial partition to be considered in geographical studies, to avoid ecological fallacies, to infer individual (or disaggregated level) relationships from spatially aggregated data, to interpret correctly the parameters' estimates, to identify the worst cases scenarios when changing the level of aggregation, to suggest how to approximate the value of parameters when we are interested at a fine level of aggregation (e.g. in EU the NUTS-3 level), but we only avail data at a coarser level of aggregation (e.g. in the NUTS-2 level), to suggest grouping criteria that preserve some properties of the estimators when choosing between alternative zoning systems at a given scale, in spatial interpolation and small area estimation to name only few. The effects of MAUP on statistical measures pioneered by Gehlke and Biehl (1934), Yule and Kendall (1950) and Robinson (1950), have been studied at length by Openshaw and Taylor (1979) and Arbia (1989) amongst the others. In particular the effects on standard econometric models are well known dating back to the early contributions of Prais and Aitchinson (1954), Theil (1954), Zellner (1962), Cramer (1964), Haitovsky (1973), Orcutt et al. (1968) and more recent contributions are found in Barker and Pesaran (1989), Okabe and Tagashira (1996), Tagashira and Okabe (2002), Griffith et al. (2003). The main results found in the literature are that the GLS estimators of regression's parameters are BLUE with a sampling variance greater than that obtained using GLS on the original data. Such underestimation of the variance of the estimators leads to biases in the hypothesis testing on the parameters leading to wrong inferential conclusions. The null hypothesis is rejected more frequently than it should and we accept as good models that should be discarded. The loss in efficiency due to aggregation depends on the grouping criterion and it is minimized when groups are made so as to maximize the within group variability with respect to the between group variability. Furthermore the coefficient of determination R^2 is emphasized and generally micro-data can better discriminate between alternative specifications of the models.

In a bivariate setting Arbia (1989) derived the formal relationship between the simple Pearson's correlation coefficient at the original process level (say level-1) and the correlation at the aggregate process level (say level-2) when data are spatial correlated. Conclusions were based on two particular specifications of the variance-covariance matrix of X and Y, namely: (i) the uniformly covariant variance-covariance matrix where all individual units are equally correlated introduced by Smith (1980) also termed the equicorrelated hypothesis by Shabenberger and Gotaway (2000) and the uniform weights assumption by Kelejian (2008), and (ii) the locally covariant variance-covariance matrix where each individual unit is correlated with only its first order neighbor through a common parameter.

In this paper we explicitly study the effects on the precision of the parameters' estimates in spatial econometric models. We present here some results of our approach and we consider, in particular, the effects on the SARAR linear models. We will present the theory and some examples based on artificial data. Results are organized in the following way. In Sect. 2 we will consider the effects of MAUP on spatial linear models while Sect. 3 contains some concluding remarks and some indications for future research.

2 Effect of MAUP on the regression coefficients estimators efficiency

Let us consider the effects of aggregation on linear spatial econometric models and let us refer to the general formulation of a SARAR (1,1) (model Kelejian and Prucha (1998)):

$$\begin{aligned} \underset{n\cdot 1}{y} &= \lambda \underset{n\cdot n}{W} \underset{n\cdot 1}{y} + \underset{n\cdot k}{X} \underset{k\cdot 1}{\beta} + \underset{n\cdot 1}{u} \qquad |\lambda| < 1 \\ \underset{n\cdot 1}{u} &= \rho \underset{n\cdot n}{W} \underset{n\cdot 1}{u} + \underset{n\cdot 1}{\varepsilon} \qquad\qquad\quad |\rho| < 1 \end{aligned} \tag{1}$$

where X is an n by k matrix of known constants, $\underset{n\cdot 1}{\varepsilon} \sim N.I.D(0, \sigma_\varepsilon^2 I)$ and W is the row-standardized n by n matrix defined according to the rook's case definition. The first equation considers the spatially lagged variable of the dependent variable as one of the regressors and may also contain spatially lagged variables of some of the exogenous variables. The second equation considers a spatial model for the stochastic disturbances. In the previous equation if $\lambda = 0$ we have the Spatial Error Model, if $\rho = 0$ the Spatial Lag Model, Arbia (2006). Finally if $\lambda = \rho = 0$ we have the classical linear econometric model with no spatial effects. We will review the effects of MAUP on such model considering different combinations of the spatial parameters λ and ρ.

2.1 Effects on the benchmarking classical econometric model

Just to produce some benchmarking results, let us start considering the classical linear econometric model with no spatial effects ($\lambda = \rho = 0$)

$$\underset{n\cdot 1}{y} = \underset{n\cdot k}{X} \underset{k\cdot 1}{\beta} + \underset{n\cdot 1}{\varepsilon} \tag{2}$$

with $\underset{n\cdot 1}{\varepsilon} \sim N.I.D(0, \sigma_\varepsilon^2 I)$ and let us define the aggregated data as:

$$\begin{aligned} \underset{m\cdot k}{X^*} &= \underset{m\cdot n}{G} \underset{n\cdot k}{X} \\ \underset{m\cdot 1}{y^*} &= \underset{m\cdot n}{G} \underset{n\cdot 1}{y} \\ \underset{m\cdot 1}{\varepsilon^*} &= \underset{m\cdot n}{G} \underset{n\cdot 1}{\varepsilon} \end{aligned}$$

with $m < n$ and where the aggregation m by n matrix G is taken in the case of *perfect aggregation*, discussed by Theil (1954), in which each level-2 unit contains

the same number of level-1 units. In this way we neglect the aggregated problem and we concentrate only on the scale problem.
G is defined as:

$$\underset{m\cdot n}{G} = \begin{bmatrix} \overbrace{1 \;\dots\; 1}^{r\ times} & 0 \;\dots\;\dots\;\dots & & 0 \\ 0 \;\dots\; 0 & 1 \;\dots\; 1 & 0 & \dots\; 0 \\ \dots & \dots & & \dots \\ 0 \;\dots\; 0 & 0 \;\dots\; 0 & 1 & \dots\; 1 \end{bmatrix}$$

and the integer r is given by $r = \frac{n}{m}$.
The aggregated model is:

$$\underset{m\cdot 1}{y^*} = \underset{m\cdot k}{X^*}\,\underset{k\cdot 1}{\beta^*} + \underset{m\cdot 1}{\varepsilon^*} \tag{3}$$

$$\underset{m\cdot 1}{E(\varepsilon^*)} = 0$$

$$\underset{m\cdot m}{E(\varepsilon^* \varepsilon^{*T})} = \sigma_\varepsilon^2 \underset{m\cdot m}{(GG^T)} = r\sigma_\varepsilon^2 \underset{m\cdot m}{I} = \sigma_{\varepsilon^*}^2 \underset{m\cdot m}{I}$$

where the regression coefficients $\underset{k\cdot 1}{\beta^*}$ are the parameters of the aggregated model, $GG^T = r \underset{m\cdot m}{I}$ and $\sigma_{\varepsilon^*}^2 = r\sigma_\varepsilon^2$.

We call $\underset{k\cdot 1}{\hat{\beta}^*_{OLS}}$ the OLS estimators of the parameters $\underset{k\cdot 1}{\beta^*}$ in (3):

$$\underset{k\cdot 1}{E(\hat{\beta}^*)} = \left(X^{*T} X^*\right)^{-1} X^{*T} y^* = \left(X^T G^T G X\right)^{-1} X^T G^T G y \equiv f(r; \sigma_\varepsilon^2) \tag{4}$$

Furthermore we have that the variance of the OLS estimators are given by

$$\underset{k\cdot k}{\Sigma_{\hat{\beta}^*}} = Var(\hat{\beta}^*) = \sigma_{\varepsilon^*} \left(X^{*T} X^*\right)^{-1} = \sigma_\varepsilon^2 \left(X^T G^T G X\right)^{-1} \equiv f(r, \sigma_\varepsilon^2) \tag{5}$$

where $\Sigma_{\hat{\beta}^*}$ is a k by k variance-covariance matrix and the OLS estimator is not fully efficient, the loss in *efficiency* depending on the aggregation criterion (r) and on the variance of the random component.

Here below we show some results based on the above expressions. A regular square lattice grid 64-by-64 ($n = 4096$) is progressively aggregated by constituting groups of 4 neighboring units into a 32-by-32 ($n = 1024$), 16-by-16 ($n = 256$), 8-by-8 ($n = 64$) and 4-by-4 ($n = 16$). Correspondingly, the number of units in each groups are $r = 4,\ 16,\ 64,\ 256,\ 1024$. We consider only one independent variable ($k = 1$) and the values 1 and 0.5 for the true β. The results are displayed in Fig. 1: the efficiency loss is monotonic and there is no systematic effect due to aggregation depending on different values of β.

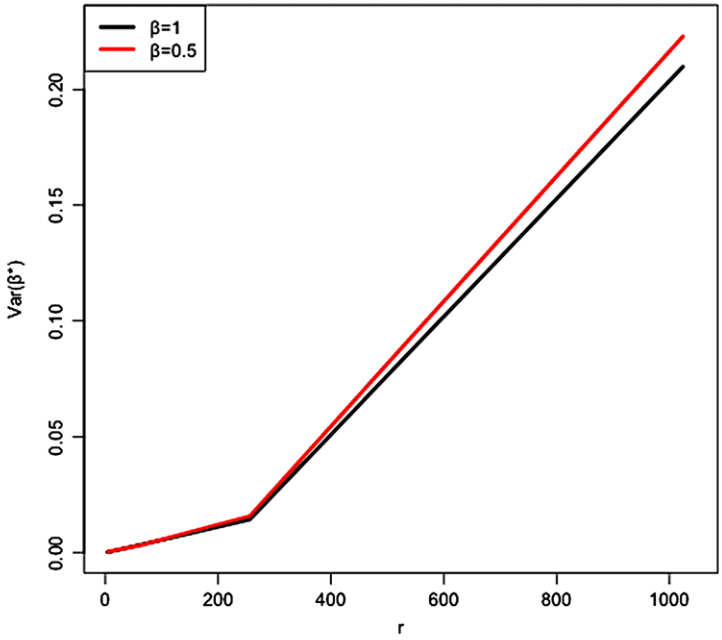

Fig. 1 Efficiency loss of the estimator of the regression coefficient at different scale levels

2.2 Effects on the Spatial Lag model

Let us now consider an alternative model where a spatially lagged dependent variable is included in the list of independent variables (Spatial Lag model. See Anselin 1988; Arbia 2006):

$$\underset{n\cdot 1}{y} = \underset{n\cdot k}{X}\ \underset{k\cdot 1}{\beta} + \lambda \underset{n\cdot n}{W}\ \underset{n\cdot 1}{y} + \underset{n\cdot 1}{\varepsilon} \tag{6}$$

where $\underset{n\cdot 1}{\varepsilon} \sim N.I.D(0, \sigma_\varepsilon^2 I)$ and W is the row-standardized n by n matrix defined according to the rook's case definition. The model can be recast:

$$\underset{n\cdot 1}{y} = \left(I - \underset{n\cdot n}{\lambda}\ W\right)^{-1} \underset{n\cdot k}{X}\ \underset{k\cdot 1}{\beta} + \underset{n\cdot 1}{u} \tag{7}$$

where $\underset{n\cdot 1}{u} = (I - \underset{n\cdot n}{\lambda} W)^{-1} \underset{n\cdot 1}{\varepsilon}$ hence

$$E(\underset{n\cdot 1}{y}) = (I - \lambda\ W)^{-1} X\beta$$

$$\underset{n\cdot n}{E(uu^T)} = \sigma_\varepsilon^2 (I - \lambda W)^{-1} (I - \lambda W)^{-T}$$

see Arbia (2006).

We define the aggregated data as:

$$\underset{m\cdot k}{X^*} = \underset{m\cdot n}{G}\ \underset{n\cdot k}{X}; \quad \underset{m\cdot 1}{y^*} = \underset{m\cdot n}{G}\ \underset{n\cdot 1}{y}; \quad \underset{m\cdot 1}{\varepsilon^*} = \underset{m\cdot n}{G}\ \underset{n\cdot 1}{\varepsilon},$$

with $m < n$, and the aggregated model is:

$$\underset{m\cdot 1}{y^*} = \underset{m\cdot k}{X^*}\ \underset{k\cdot 1}{\beta^*} + \lambda^*\ \underset{m\cdot m}{W^*}\ \underset{m\cdot 1}{y^*} + \underset{m\cdot 1}{\varepsilon^*} \tag{8}$$

where $\underset{m\cdot m}{W^*}$ is exogenously given and represent the weight matrix at the aggregated level. We can rewrite (7) as:

$$\underset{m\cdot 1}{y^*} = (I - \underset{m\cdot m}{\lambda^*\ W^*})^{-1} \underset{m\cdot k}{X^*}\ \underset{k\cdot 1}{\beta^*} + \underset{m\cdot 1}{u^*} \tag{9}$$

where $\underset{m\cdot 1}{u^*} = (I - \underset{m\cdot m}{\lambda^*\ W^*})^{-1} \underset{m\cdot 1}{\varepsilon^*}$.

Hence

$$\begin{aligned}
\underset{m\cdot m}{E(u^* u^{*T})} &= E\left[\left(I - \lambda^* W^*\right)^{-1} \varepsilon^* \varepsilon^{*T} \left(I - \lambda^* W^*\right)^{-T}\right] \\
&= E\left[(I - \lambda W)^{-1} G\varepsilon\varepsilon^T G^T (I - \lambda W)^{-T}\right] \\
&= \sigma_\varepsilon^2 \left[(I - \lambda W)^{-1} GG^T (I - \lambda W)^{-T}\right] \\
&= r\sigma_\varepsilon^2 \left[(I - \lambda W)^{-1} (I - \lambda W)^{-T}\right].
\end{aligned}$$

The model in (8) contains a spatial lag in the dependent variable. This term is a endogenous and models of this kind are usually estimated via Maximum Likelihood or Instrumental Variables (IV method) when λ is unknown. On the contrary if is known the GLS estimator coincides with the Maximum Likelihood. In the remainder of this paper we are going to consider the spatial parameter λ under experimental control and we are going to monitor how the precision of the estimator changes for different values of it. As a consequence considering the GLS estimator is not a limitation in the present analysis. The GLS estimators of $\underset{k\cdot 1}{\beta^*}$, $\underset{k\cdot 1}{\hat{\beta}^*_{GLS}}$ are given by:

$$\underset{k\cdot 1}{\hat{\beta}^*_{GLS}} = \left(\underset{k\cdot m}{X^{*T}}\ \underset{m\cdot m}{\Omega^{-1}}\ \underset{m\cdot k}{X^*}\right)^{-1} \underset{k\cdot m}{X^{*T}}\ \underset{m\cdot m}{\Omega^{-1}}\ \underset{m\cdot 1}{y^*} \tag{10}$$

where

$$\underset{m\cdot m}{\Omega} = r\left[\left(I - \underset{m\cdot m}{\lambda^*\ W^*}\right)^{-1} \left(I - \underset{m\cdot m}{\lambda^*\ W^*}\right)^{-T}\right].$$

Hence

$$\hat{\beta}^*_{GLS} = \left\{X^T G^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right) GX\right\}^{-1}$$

$$\times X^T G^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right) Gy \equiv f(r, \lambda). \tag{11}$$

In this case the variance of the GLS estimators are given by:

$$\begin{aligned}\underset{k \cdot k}{\Sigma_{\hat{\beta}*}} &= Var(\hat{\beta}^*_{GLS}) = \sigma^2_\varepsilon \left(X^{*T} \Omega^{-1} X^*\right)^{-1} \\ &= r\sigma^2_\varepsilon \left[X^T G^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right) GX\right]^{-1} \equiv f(r, \sigma^2_\varepsilon, \lambda) \end{aligned} \tag{12}$$

where $\Sigma_{\hat{\beta}*}$ is a k by k variance-covariance matrix whose element are a function of the aggregation criterion, of the stochastic component variance and of the spatial parameter λ. In order to measure the loss of efficiency produced by the aggregation it is useful to consider the ratio between the variance of the estimators (that are the diagonal elements of $\Sigma_{\hat{\beta}*}$) at the more disaggregated level and the variance after aggregation. Thus, for each level of aggregation we can define:

$$\left[EL(\hat{\beta}^*_{GLS})\right]_i = \frac{[X^T G^T (I - \lambda^* W^*)^T (I - \lambda^* W^*) GX]_{ii}}{[rX^T[(I - \lambda W)^T (I - \lambda W)]X]_{ii}} \equiv f(r, \lambda)$$
$$\text{for } i = 1, \ldots, k. \tag{13}$$

Furthermore, to measure how the presence of a spatial lag affects the loss in efficiency we can look, at any given level of aggregation, at the ratio between the variance of the estimators in the classical model and the variance in the spatial lag model, thus define a relative efficiency loss (REL) which is given by:

$$\left[REL(\hat{\beta}^*_{GLS})\right]_i = \frac{[X^T G^T (I - \lambda^* W^*)^T (I - \lambda^* W^*) GX]_{ii}}{[X^T G^T GX]_{ii}} \equiv f(r, \lambda)$$
$$\text{for } i = 1, \ldots, k. \tag{14}$$

Results on the loss in efficiency are reported in Fig. 2. In the simulations we considered only one independent variable ($k = 1$ with $\beta = 1$) and 6 values of the parameter λ (± 0.7; ± 0.24; ± 0.1).

Values of $REL < 1$ refer to cases where the REL is exacerbated by the presence of spatial effects; whereas values of $REL > 1$ refer to cases value the REL is moderated by the presence of λ.

Figure 2 shows that the loss in efficiency is more dramatic for high values of λ. However, when $\lambda = 0.7$ the loss tends to diminish increasing the level of aggregation (r). Conversely when $\lambda = -0.7$ the loss is dramatic and independent of the level of aggregation. In the remaining cases the REL increases with the level of aggregation with the only exception of $\lambda = 0.24$ when aggregation instead reduces the loss. In all other cases a positive λ has a moderating effect on the efficiency loss and this effect becomes stronger and stronger increasing the aggregation level.

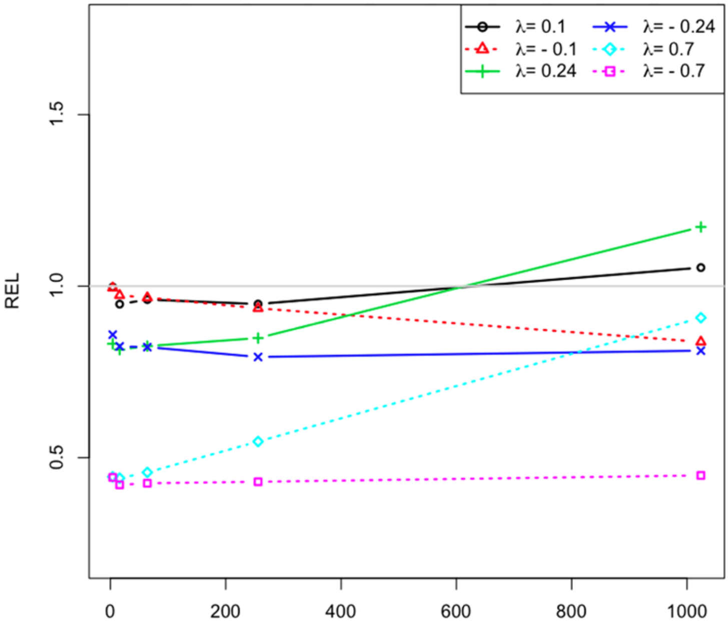

Fig. 2 Relative efficiency loss of the estimator of the regression coefficient at different scale levels in a Spatial Lag Model. Graph is based on 10000 simulations. In all simulations $\beta = 1$

2.3 Effects on the SARAR (1,1) model

Let us finally consider the full SARAR model that includes not only a lagged value of y, but also a spatial model for the residuals, given by:

$$\underset{n\cdot 1}{y} = \lambda \underset{n\cdot n}{W} \underset{n\cdot 1}{y} + \underset{n\cdot k}{X} \underset{k\cdot 1}{\beta} + \underset{n\cdot 1}{u} \qquad |\lambda| < 1 \tag{15}$$

$$\underset{n\cdot 1}{u} = \rho \underset{n\cdot n}{W} \underset{n\cdot 1}{u} + \underset{n\cdot 1}{\varepsilon} \qquad |\rho| < 1 \tag{16}$$

where $\underset{n\cdot 1}{\varepsilon} \sim N.I.D(0, \sigma_{\varepsilon}^2 I)$ and W is the row-standardized n by n matrix defined according to the rook's case definition.

Equation (15) considers the spatially lagged variable of the dependent variable y as one of the regressors and may also contain spatially lagged variables of some of the exogenous variables. Equation (16) considers a spatial model for the stochastic disturbances. This model was termed SARAR (1,1) (Kelejian and Prucha 1998) and encompasses the spatial error and the spatial lag model. The model can be rewritten:

$$\underset{n\cdot 1}{y} = (I - \lambda \underset{n\cdot n}{W})^{-1} \underset{n\cdot k}{X} \underset{k\cdot 1}{\beta} + \underset{n\cdot 1}{\eta} \tag{17}$$

where $\underset{n\cdot 1}{\eta} = (I - \lambda \underset{n\cdot n}{W})^{-1} (I - \rho \underset{n\cdot n}{W})^{-1} \underset{n\cdot 1}{\varepsilon}$.

As a consequence:

$$E(y) = (I - \lambda\ W)^{-1} X \beta, \quad E(\eta) = 0$$

$$E(\eta\eta^T) = \sigma_\varepsilon^2 \left[(I - \lambda\ W)(I - \rho\ W)\right]^{-1} \left[(I - \lambda\ W)(I - \rho\ W)\right]^{-T}$$

We can now define the aggregated data as:

$$\underset{m\cdot k}{X^*} = \underset{m\cdot n}{G}\ \underset{n\cdot k}{X}; \quad \underset{m\cdot 1}{y^*} = \underset{m\cdot n}{G}\ \underset{n\cdot 1}{y}; \quad \underset{m\cdot 1}{\varepsilon^*} = \underset{m\cdot n}{G}\ \underset{n\cdot 1}{\varepsilon},$$

with $m < n$, and the aggregated model as:

$$\begin{aligned} \underset{m\cdot 1}{y^*} &= \lambda^* \underset{m\cdot m}{W^*}\ \underset{m\cdot 1}{y^*} + \underset{m\cdot k}{X^*}\ \underset{k\cdot 1}{\beta^*} + \underset{m\cdot 1}{u^*} \qquad |\lambda| < 1 \\ \underset{m\cdot 1}{u^*} &= \rho^* \underset{m\cdot m}{W^*}\ \underset{m\cdot 1}{u^*} + \underset{m\cdot 1}{\varepsilon^*} \qquad\qquad\quad |\rho| < 1 \end{aligned} \tag{18}$$

where W^* is exogenously given and represents the weight matrix at the aggregated level.

We can rewrite (18) as:

$$\underset{m\cdot 1}{y^*} = (I - \underset{m\cdot m}{\lambda^*\ W^*})^{-1} \underset{m\cdot k}{X^*}\ \underset{k\cdot 1}{\beta^*} + \underset{m\cdot 1}{\eta^*} \tag{19}$$

where $\underset{m\cdot 1}{\eta^*} = (I - \underset{m\cdot m}{\lambda^*\ W^*})^{-1}(I - \underset{m\cdot m}{\rho^*\ W^*})^{-1} \underset{m\cdot 1}{\varepsilon^*}$.

Hence

$$\begin{aligned} &E(\eta^* \eta^{*T}) \\ &\quad = E\left[\left(I - \lambda^* W^*\right)^{-1} \left(I - \rho^* W^*\right)^{-1} G\varepsilon\varepsilon^T G^T \left(I - \lambda^* W^*\right)^{-T} \left(I - \rho^* W^*\right)^{-T}\right] \\ &\quad = r\sigma_\varepsilon^2 \left(I - \lambda^* W^*\right)^{-1} \left(I - \rho^* W^*\right)^{-1} \left(I - \lambda^* W^*\right)^{-T} \left(I - \rho^* W^*\right)^{-T}. \end{aligned}$$

The GLS estimators of β^*, say $\hat{\beta}^*$, now are transformed into:

$$\hat{\beta}^*_{GLS} = \left(X^{*T} \Omega^{-1} X^*\right)^{-1} X^{*T} \Omega^{-1} y^* \tag{20}$$

with

$$\underset{m\cdot m}{\Omega} = r \left(I - \lambda^* W^*\right)^{-1} \left(I - \rho^* W^*\right)^{-1} \left(I - \rho^* W^*\right)^{-T} \left(I - \lambda^* W^*\right)^{-T}.$$

Hence

$$\begin{aligned} \underset{k\cdot 1}{\hat{\beta}^*_{GLS}} &= \left[X^T G^T \left(I - \rho^* W^*\right)^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right) \left(I - \rho^* W^*\right) GX\right]^{-1} \\ &\qquad \times X^T G^T \left(I - \rho^* W^*\right)^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right) \left(I - \rho^* W^*\right) Gy \\ &\equiv f(r, \rho^*, \lambda^*). \end{aligned} \tag{21}$$

In this case the variance of the GLS estimator is given by:

$$\underset{k\cdot k}{\Sigma_{\beta^*_{GLS}}} = Var\left(\hat{\beta}^*_{GLS}\right) = \sigma_\varepsilon^2 \left(X^{*T} \Omega^{-1} X^{*T}\right)^{-1}$$

$$
\begin{aligned}
&= r\sigma_\varepsilon^2 \left(X^T G^T \left(I - \rho^* W^*\right)^T \left(I - \lambda^* W^*\right)^T \left(I - \lambda^* W^*\right)\left(I - \rho^* W^*\right) G X\right)^{-1} \\
&\equiv f(r, \sigma_\varepsilon^2, \lambda^* \rho^*) \qquad (22)
\end{aligned}
$$

which is now a function of the level of aggregation r, of the stochastic component variance and of the SARAR spatial parameters ρ and λ. Similarly to what we did with the Spatial Lag Model, we can measure the loss of efficiency produced by the aggregation on SARAR models. In this case this is equal to:

$$
\left[EL(\hat{\beta}^*_{GLS})\right]_i = \frac{[(X^T G^T (I - \lambda^* W^*)^T (I - \rho^* W^*)^T (I - \lambda^* W^*)(I - \rho^* W^*) G X)]_{ii}}{[[r(X^T (I - \lambda W)^T (I - \rho W)^T (I - \lambda W)(I - \rho W))X]]_{ii}} \qquad (23)
$$

for $i = 1, \ldots, k$ and for each level of aggregation.

Furthermore, in order to assess how the presence of spatial effects modifies the loss in efficiency, we can look, for any given level of aggregation, at the Relative Efficiency Loss (REL), that is at the ratio between the variance of the estimators in the classical model and the variance in the SARAR model:

$$
\begin{aligned}
&\left[REL(\hat{\beta}^*_{GLS})\right]_i \\
&\quad = \frac{[(X^T G^T (I - \lambda^* W^*)^T (I - \rho^* W^*)^T (I - \lambda^* W^*)(I - \rho^* W^*) G X)]_{ii}}{[X^T G^T G X]_{ii}} \qquad (24)
\end{aligned}
$$

for $i = 1, \ldots, k$ and for each level of aggregation.

Figures 3 and 4 report the REL for various combinations of the parameters λ and ρ. Here we consider 6 values of the parameters combining the values of ρ (± 0.7; ± 0.24; ± 0.1) with the values of λ (± 0.7; ± 0.24; ± 0.1).

Figures 3 and 4 show that, at low levels of aggregation, the loss is exacerbated ($REL < 1$) with respect to the case of no spatial effect for moderated levels of λ when ρ is also moderated. Conversely when $\lambda = \pm 0.7$ the loss is always exacerbated apart from the case in which $\lambda = \rho = 0.7$. On the contrary the loss is always mitigated ($REL > 1$) at high levels of aggregation when $\rho = 0.7$ and when $\rho > 0$ and $|\lambda| < 0.7$. The worst case is when $\lambda = \rho = -0.7$. Finally when $\lambda = -0.7$ we observe a perfect ranking of the REL with respect to the value of ρ that ranges from the lowest to the highest values.

3 Summary and conclusion

In this paper we presented a general framework to analyze the effects of MAUP on spatial econometric models. In particular dealing with linear spatial econometric models of the SARAR (1,1) class we concentrated on the loss in efficiency of the parameters' estimators due to aggregation showing how the presence of spatial effects affects the classical results. In our analysis we considered the spatial parameters as given and studied how they influence the efficiency of the estimators. For these reasons we limited ourselves to the GLS estimators that, in this experimental

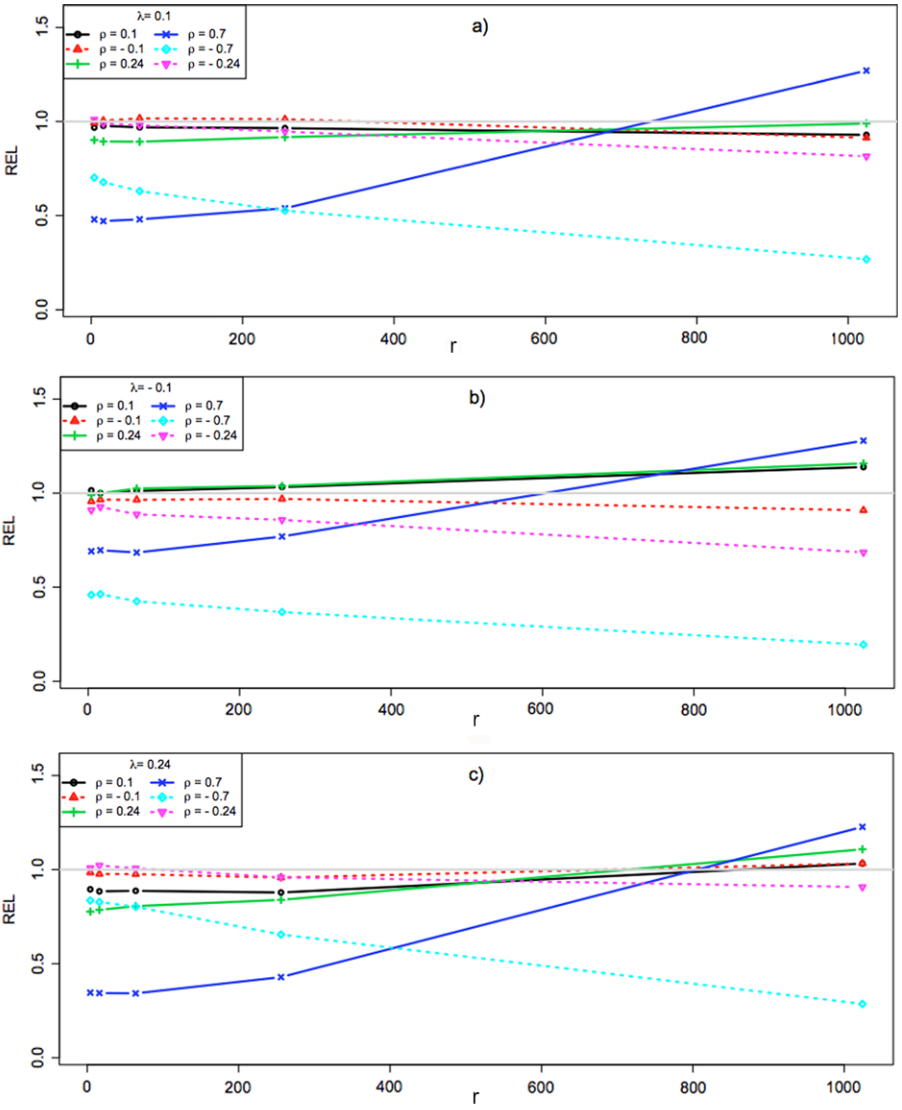

Fig. 3 Relative Efficiency Loss at various scales for various combinations of the SARAR(1,1) parameters, with $\beta = 1$. Graphs are based on 10000 simulations

situation for any given values of the spatial parameters coincide with the Maximum Likelihood estimators that are more commonly considered in the literature to tractile the endogenously of the spatially lagged dependent variable.

The efficiency loss connatural to aggregation is, generally speaking, mitigated by the presence of a positive spatial correlation parameter and conversely exacerbated by the presence of a negative spatial correlation parameter. This effect is observed both with respect to the spatial dependence in the dependent variable and in the error component even if the pattern of interaction between the two parameter is rather

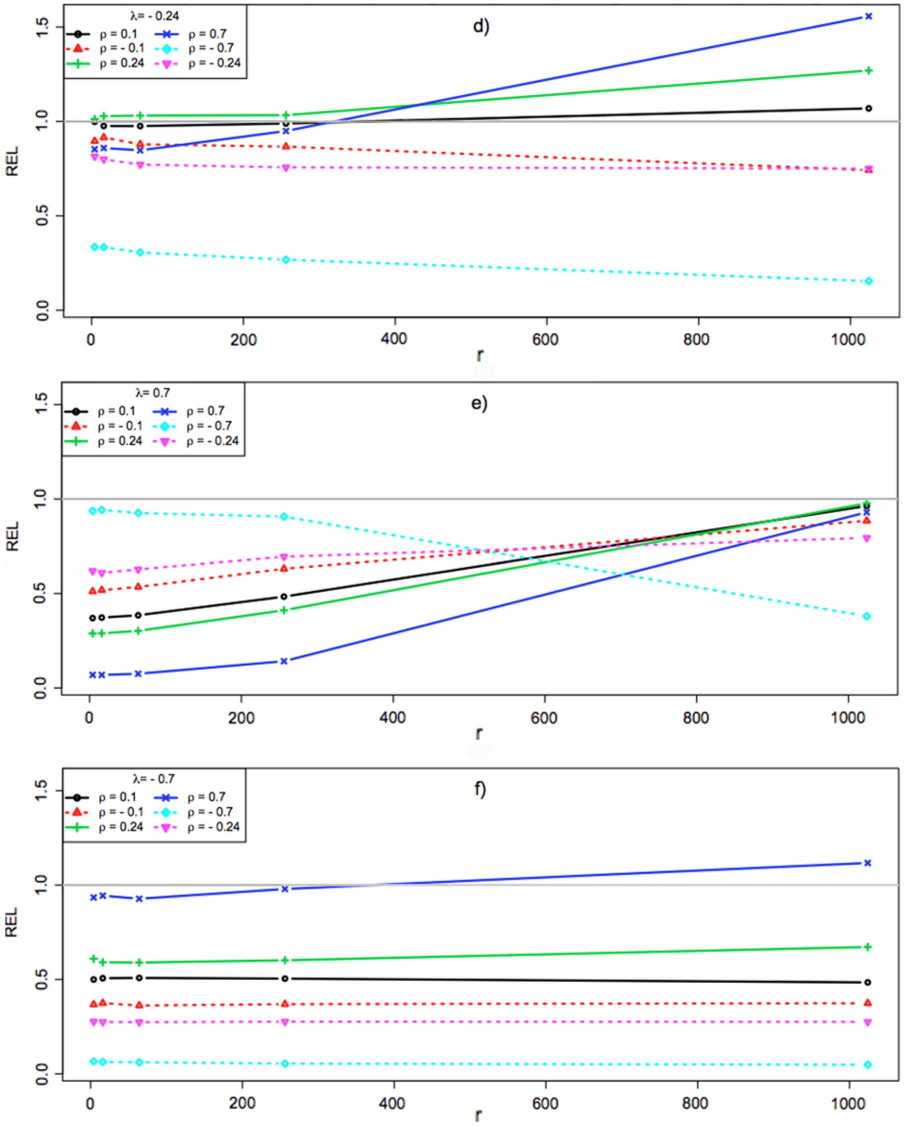

Fig. 4 Relative Efficiency Loss at various scales for various combinations of the SARAR(1,1) parameters, with $\beta = 1$. Graphs are based on 10000 simulations

complicated. This result is coherent with the theoretical expectation. Positive spatial correlation implies that we aggregate between similar values thus preserving variability. Conversely negative spatial correlation implies aggregation between very different values thus implying a more dramatic enhancement of variability. The results are also coherent with the previous findings of Arbia (1989). The results presented here could help to study into details the influence of spatial effects on aggregation of econometric models. This in turn could lead to important results. First of all it could help in identifying worst cases scenarios when changing the level of aggregation. Secondly it could help to infer the likely value of the models' parameters at a finer level of aggregation having in hand data only at a coarser level of aggregation e.g. exploiting a Bayesian strategy using our results as the likelihood and imposing reasonable priors. In the future we aim at extending the results obtained here to the case of non-perfect aggregation considering the aggregation (or zoning) problem which occurs at any given spatial scale. Furthermore we aim at considering the effects of MAUP on non-linear econometric models like those usually considered in gravity modelling of spatial interaction flows.

References

Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic, Dordrecht (1988)

Arbia, G.: Spatial Data Configuration in the Statistical Analysis of Regional Economics and Related Problems. Kluwer, Dordrecht (1989)

Arbia, G.: Spatial Econometrics - Statistical Foundations and Applications to Regional Convergence. Springer, Berlin (2006)

Barker, T., Pesaran, M.H. (eds.): Disaggregation in Econometric Modeling. Routledge, London (1989)

Cramer, J.S.: Efficient grouping, regression and correlation in Engel curve analysis. J. Am. Stat. Assoc. **59**, 233–250 (1964)

Gehlke, C.E., Biehl, K.: Certain effects of grouping upon the size of the correlation coefficient in census tract material. J. Am. Stat. Assoc., Suppl. **29**, 169–170 (1934)

Griffith, D.A., Wong, D.W.S., Whitfied, T.: Exploring relationships between the global and regional measures of spatial autocorrelation. J. Reg. Sci. **43**(4) (2003)

Haitovsky, J.: Regression Estimation from Grouped Observations. Griffin's Statistical Courses and Monographs, vol. 33. Griffin, London (1973)

Kelejian, H., Prucha, I.: A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J. Real Estate Finance Econ. **17**(1), 99–121 (1998)

Kelejian, H.: Spatial models in econometrics. Lecture notes (2008)

Okabe, A., Tagashira, N.: Spatial aggregation bias in a regression model containing a distance variable. J. Geogr. Syst. **2**, 83–202 (1996)

Openshaw, S., Taylor, P.J.: A million or so of correlated coefficients. Three experiments on the modifiable areal unit problem. In: Wrigley, N., Bennet, R.J. (eds.) Statistical Applications in the Spatial Sciences, pp. 127–144. Pion, London (1979)

Orcutt, et al.: Data aggregation and information loss. Am. Econ. Rev. 773–787 (1968)

Prais, S., Aitchinson, J.: The grouping of observations in regression analysis. Rev. Inst. Int. Stat. , **1**, 1–22 (1954)

Robinson, W.S.: Tecological correlations and the behavior of individuals. Am. Sociol. Rev. **15**, 351–357 (1950)

Shabenberger, O., Gotaway, C.A.: Statistical Methods for Spatial Data Analysis. Chapman and Hall, London (2000)

Smith, T.: A central limit theorem for spatial samples. Geogr. Anal. **12**, 299–324 (1980)

Tagashira, N., Okabe, A.: The modifiable areal unit problem in a regression model whose independent variable is a distance from a predetermined point. Geogr. Anal. **34**(1) (2002)

Theil, H.: Linear Aggregation in Economic Relations. North-Holland, Amsterdam (1954)

Yule, U., Kendall, M.S.: An Introduction to the Theory of Statistics. Griffin, London (1950)

Zellner, A.: An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. J. Am. Stat. Assoc. **57**, 348–368 (1962)