Quality & Quantity
https://doi.org/10.1007/s11135-023-01818-1

Top–down disaggregation of life expectancy up to municipal areas, using linear self-regressive spatial models

Vincenzo Basile[1] · **Stefano Cervellera**[2] · **Carlo Cusatelli**[3] · **Massimiliano Giacalone**[4]

Accepted: 10 December 2023
© The Author(s) 2024

Abstract
The paper aims to analyse a statistical procedure for the definition of territorial indicators, such as the biometric function of life expectancy of citizens of a territory e^0, applying a methodology of Top–Down spatial disaggregation, using census data from 2011 in Italy. Through spatial regressions with the methodology proposed in 1971 by Chow and Lin with the use of ISTAT elaborations of annual mortality tables, which are structured from the national level to the regional level, up to the smallest details of the main level, as a dependent variable and predictors a number k of census variables plus accidents in regression models, life expectancy can also be defined at municipal levels (not elaborated by ISTAT) and even at sub-municipal levels (Census Area). The structure of the 2011 census was characterized by 152 variables, collected with CAPI and universal CAWI survey on all the survey units, throughout the national territory, divided into administrative areas of competence and 402,677 more granular areas in census sections. This structure represents a very relevant and useful information asset for applying a spatial disaggregation of indicators, based on three fundamental assumptions:

1. Structural similarity, whereby the aggregate model and the disaggregate model are structurally similar, i.e., the relationships between the variables are valid both at the aggregate (Top) and at the disaggregate (Down) level, with the consequence that the regression parameters are the same in the two models.
2. Error similarity: for spatially correlated errors they present the same structure at both aggregate (Top) and disaggregate (Down) levels, significantly testing for zero spatial correlation;
3. Reliable indicators: the variables in the regression models are efficient predictors at both aggregates (Top) and disaggregate (Down) levels, estimable from model efficiency tests.

As we see in the following, compared to others, the best predictors of the census and income variables show us a good interaction in terms of active regressors on the estimation variable.

Keywords Spatial analysis · Spatial autoregressive models · Projection pursuit · Disaggregation of territorial indicators · Census variables · Census areas

Extended author information available on the last page of the article

Published online: 27 January 2024

1 Introduction

In Italy, the Census has been carried out every 10 years[1] since 1871 (Unification of Italy), creating a unique and exceptional information patrimony, at the level of microdata such as the family unit and the house, surveyed in a universal and direct way. Until 2011 the Census was carried out with the classic method,[2] while in 2018 Istat started the permanent one in Italy, with the assistance of the administrative archives, losing its universal nature, passing to continuous sample surveys (Cervellera et al. 2021). For this reason, we wanted to use in our article the immense patrimony of the Territorial Bases of the Istat 2011 census, as a geographic and spatial data warehouse of the entire open-source[3] census survey, since Istat has not yet updated the bases themselves to the data of the new permanent censuses carried out so far. Eurostat bases the territorial administrative classification in NUTS (Nomenclature of Territorial Units for Statistics)[4] according to the European legislation governing official EU statistics; three territorial levels are regulated: NUT1 (State), NUT2 (Region) and NUT3 (Sub-region—Province in Italy). Therefore, the National Institutes of the European Statistical System, including the Italian Istat, are obliged to define the statistics at a minimum level of NUT3. Life expectancy is a very important indicator (e.g., for the quality of health and life) that Istat, like many other European institutes, provides on a NUT3 provincial basis (Cervellera and Cusatelli 2022). Knowing these data at the sub-provincial level, by the municipality and sub-municipal areas (e.g., ACE—census areas) would be very important for citizens and public institutions (Gennaro et al. 2022), and small-area disaggregation analyses are usually performed in the USA, by the Census Bureau, down to very granular sub-municipal levels (5000–7000 inhabitants).[5] In the next section, we present the data to be processed and the methodology that will allow, in Sect. 3, linear modelling of the spatial dependence, explain the results in Sect. 4, and conclude this article with the proposal of some future developments.

2 Materials and methods

The spatial and geographical structure of the Territorial Bases starts from the smallest and main granule (Down), which are the Census Sections (about 403,000), to the Census Areas (CEA), the Sub-Municipal Areas (SMA: municipalities, districts, etc.), the Localities and the areas and administrative limits of Municipalities, Provinces, and Regions (Top).

The 152 variables detected have high information on 5 Levels,[6] in addition to the area descriptors, which are homogeneous in Buildings, Families, Foreigners, Households and Population. Many of these variables will be excellent predictors of biometric functions, particularly life expectancy at birth e°.

[1] Except in some years for various causes such as the war, different census lags (1881–1901, 1936–1951).
[2] Universal, periodic, direct, and simultaneous.
[3] www.istat.it/it/archivio/104317 Territorial Bases.
[4] https://ec.europa.eu/eurostat/web/nuts/background
[5] The United States Small-Area Life Expectancy Project (USALEEP).
[6] Five homogeneous levels, plus an area writing level.

Type	Count
Descriptive areas	12
Building structure (B)	31
Family structure (FS)	9
Foreigners (FO)	15
House Structure (H)	9
Population structure (P)	76
Total	152

At the territorial level of spatial granularity, however, the structure is as follows:

	Census year		
Roofing area	1991	2001	2011
Regions	20	20	20
Provinces	95	103	110
Municipalities	8100	8101	8092
Sub Municipal Areas[11]	70,742	60,482	60,447
Census sections	323,616	382,534	402,677
Census population	56,778,031	56,995,744	59,433,744

In municipalities of Bari, Bologna, Brescia, Cagliari, Catania, Ferrara, Firenze, Foggia, Genova, Livorno, Messina, Milano, Modena, Monza, Napoli, Padova, Palermo, Parma, Perugia, Pescara, Prato, Ravenna, Reggio Calabria, Reggio Emilia, Rimini, Roma, Salerno, Sassari, Siracusa, Taranto, Torino, Trieste, Venezia, and Verona.

The data of the biometric functions and life expectancy at birth are taken from the ISTAT elaborations of the mortality tables, up to the maximum provincial level (Down) of 2011, homogeneously pre-testing the model and the structure. To derive indicators with a Top–Down methodological approach of the highest territorial units (Top) to the bases (Down), using their information to structure an autoregressive spatial correlation of spatial dependence of the territorially adjacent base nodes. The derivation of the disaggregated indicator makes use of the method proposed by Chow and Lin (1971): it is a technique designed and used for temporal disaggregation also known as temporal distribution. Temporal disaggregation is the process of deriving data from low-frequency (Top, e.g., years) to high-frequency data (Down, e.g., quarters and months). Since the results of the Chow-Lin method depend on the information on a different variable, the method can be considered as an indirect approach, while the dynamic time autoregressive dependence, can derive spatial and areal dependence, on cross-sectional and panel data, in an autoregressive form on linking nodes between disaggregation areas, correlated with spatial weights matrices (W matrix) and linking matrices between levels and indicators (C matrix), by using Cran's R package of Spatial Dependence[7] with the Spatial Autoregressive Regression (SAR) model. Polasek and Sellner (2008) have definitively adapted this method and it has been effectively extended to cross-sectional data based on a spatial autoregressive model, for panel data and for spatial flow models. An implicit assumption of the Chow-Lin approach is the *summability* of disaggregated variables to aggregate variables, a property

[7] Package spdep, spatialreg.

that holds for so-called intensive variables. This paper shows how to extend the Chow–Lin spatial approach for cross-sectional data to non-extended or intensive variables, such as growth rates. A widely used and congenial method in applied econometrics, consisting of introducing a spatial autocorrelation term into a classical multivariate regression model (Giacalone 2021).

To derive the indicators at a territorially disaggregated level from the corresponding aggregate level indicators, three basic assumptions are defined that the model must respect:

1. Structural similarity: the aggregate and disaggregate models are structurally similar, which implies that the relationships between the variables considered are the same at the aggregate and disaggregate level, with the consequence that the regression parameters are the same in the two models;
2. Error similarity: spatially correlated errors have the same structure at both aggregate and disaggregate levels. This is equivalent to saying that the spatial correlations are not significantly different at the two levels;
3. Reliable indicators: the variables used as regressors have good predictive power at both aggregate and disaggregate levels, i.e., the goodness-of-fit measures of the regression are significantly different from zero.

The model proposed by Chow and Lin is built as a temporal disaggregation model[8] of time series components, so it is modified and adapted to cross-sectional and spot analyses, already conducted in the econometric field by Bollino and Polinori (2007), by Mazziotta and Vidoli (2009a, b).

The model is characterized, on the one hand, by a functional relationship between synthetic indicators at the provincial level and a series of explicative variables observable at the disaggregated level (and, obviously, also at the aggregated level), a relationship which has been subjected to a first verification in previous work, again with reference to the infrastructural endowment, and, on the other hand, by a methodology of inference of the unknown parameters at higher (Top), Provincial and Regional, levels. The model assumes that at the disaggregated level the simple econometric relation of a linear model of the following type is valid:

$$y_{Down} = \beta_{Down} X_{Down} + \epsilon_{Down}$$

y_{Down} is the vector of indicators at the disaggregated level, X_{Down} is the matrix (n*k) of observable predictor variables at the disaggregated level. The dimensions of X_{Down} are the number of levels of Down areas equal to p, if they are provinces with respect to the Top Region, m if they are municipalities with respect to its Top Province and l if they are Census Areas, while the number of explanatory variables as predictors of the model always remains k.

For structural similarity, the indicators shall be:

$$I_a = \frac{\sum_{i=1}^{n} I_d}{n} \rightarrow \left\{ I_{Reg} = \frac{\sum_{i=1}^{p} I_{prov}}{p}; I_{Prov} = \frac{\sum_{i=1}^{m} I_{Mun}}{m}; I_{Mun} = \frac{\sum_{i=1}^{l} I_{CEA}}{l} \right\}$$

[8] On time series analysis.

For each spatial disaggregation of spatial indicators, it is also assumed that C is a matrix of dimension (n_{Down}*N_{Top}), where n is the number of Italian provinces, capable of transforming the disaggregated observations into those of a higher level (denoted by N), whatever the aggregation operator used. In particular, if the sum operator is adopted, regional estimates are obtained by summing the corresponding provincial levels ($y_a = \sum y_d$) and the generic element $C_{i,j}$ will be constructed as: $C_{i,j} = 1$, if province i ∈ region j 0, otherwise

$$C_{Reg,Prov} = \begin{bmatrix} \frac{1}{p_{Reg}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{p_{Reg}} \end{bmatrix} C_{Prov,Mun} = \begin{bmatrix} \frac{1}{m_{Prov}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{m_{Prov}} \end{bmatrix} C_{Mun,CEA} = \begin{bmatrix} \frac{1}{l_{SMA}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{l_{SMA}} \end{bmatrix}$$

and the regional estimates will be obtained by averaging the provincial estimates ($y_a = E(y_d)$).

Under the following aggregate constraints $y_a = Cy_d$, $X_a = CX_d$ and $\varepsilon_a = C\varepsilon_d$.

$$Prov \Rightarrow Reg : \left\{ y_{Reg} = y_{Prov} C_{Reg,Prov}, X_{Reg} = X_{Prov} C_{Reg,Prov} \wedge \varepsilon_{Reg} = C_{Reg,Prov} \varepsilon_{Prov} \right\}$$

$$Mun \Rightarrow Prov : \left\{ y_{Prov} = y_{Mun} C_{Prov,Mun}, X_{Prov} = X_{Mun} C_{Prov,Mun} \wedge \varepsilon_{Prov} = C_{Prov,Mun} \varepsilon_{Mun} \right\}$$

$$SMA \Rightarrow Prov : \left\{ y_{Mun} = y_{SMA} C_{Mun,SMA}, X_{Mun} = X_{SMA} C_{Mun,SMA} \wedge \varepsilon_{Mun} = C_{Mun,SMA} \varepsilon_{SMA} \right\}$$

In the past literature, Polasek and Sellner (2008) presented an advance, or rather an interesting generalization of the model, consisting of the introduction of a spatial autocorrelation term in a classical multivariate regression model. From an application point of view, this means that the level of the dependent variable Y in each area depends not only on the independent variables considered but also on the level of the same variable Y in the surrounding areas.

In fact, if one assumes the existence of spatial correlation effects not only in the levels of competitiveness between provinces but also and especially within very similar provinces, one can hypothesize (Anselin 1988) that, given a matrix of spatial weights WN and a spatial lag parameter ρ ∈ [0,1], a "mixed autoregressive and spatial regressive relationship" is verifiable at the disaggregated level. This model is called SER, Spatial Estimated Regression:

$$y_a = \rho_d W_N + \beta_d X_d + \varepsilon_d \text{ with } \varepsilon_d \, N\left[0, \rho_d^2 I_N\right]$$

by development in series $(I_N - \rho_d W_N)^{-1}$

$$E\left(y_d | X_d\right) = (1 - \rho_d W_N + \rho_d^2 W_d^2 + \cdots) X_d \beta_d$$

With $R_N = (1 - \rho_d W_N)$.

$$y_a = R^{-1} \beta_d X_d + R^{-1} \varepsilon_d \text{ with } \varepsilon_d \, N\left[0, \sigma_d^2 (R_N^T R_N)^{-1}\right]$$

$$\Sigma_d = \sigma_d^2 (R_N^T R_N)^{-1}$$

$$y_a = \rho_d W_N y_a + \beta_d C X_d + \varepsilon_d \text{ with } \varepsilon_d \, N\left[0, \Sigma_a I_N\right]$$

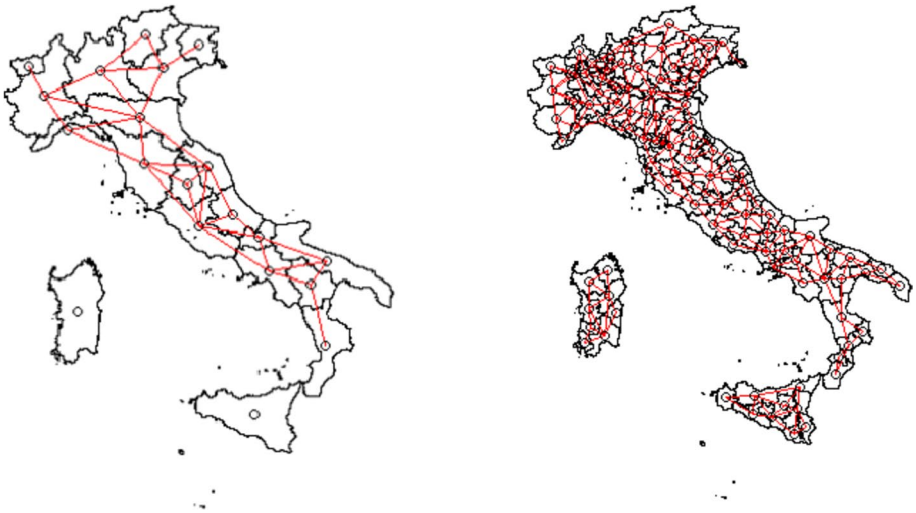

Fig. 1 Regional and provincial nodes

we get $\hat{\rho}_q$ and $\hat{\sigma}_a^2$ consistent with the assumptions of structural similarity.

$$\rho_a = \hat{\rho}_q, \beta_d = \hat{\beta}_a \text{ and } \sigma_d^2 = \hat{\sigma}_a^2.$$

Regarding the estimation of β_a according to the classical Chow-Lin approach, we obtain.

$$\hat{\beta}_a = \left(X_a^T \left(C\hat{\Sigma}_d C^T\right)^{-1} X_a\right)^{-1} X_a^T (C\hat{\Sigma}_a C^T)^{-1} y_a$$

and we can finally estimate the disaggregated dependent variable y at the administrative Down level, first of the Municipality, and then of the SCA with the 2011 Census and territorial bases.

$$\hat{y}_d = R_N^{-1} \hat{\beta}_a X_d + \hat{\Sigma}_d \left(C\hat{\Sigma}_d C^T\right)^{-1} (y_a - C\hat{R}_N^{-1} C^T \hat{\beta}_a X_a).$$

Istat calculates mortality tables, biometric functions, and life expectancy, at birth (e°) and at all ages, or age classes, each year, by provincial aggregates, online[9] from 1974 to 2021. Life expectancy is a particularly important indicator of an area's quality of life. The lack of granularity of indication that Istat limits to only the provincial aggregate (i.e., ISTAT 2006) limits its power, as it would be a most useful tool in welfare and public health policies if made available to municipal and sub-municipal administrative levels. Life expectancy has a highly variable structure both territorially and by sex discrimination. Women have a higher e° than males, on average by almost 4 years (in Italy), due to essentially biological and then social differences.

Since the census variables (as of 2011) are generally structured into sum data, Males and Females, and Males-only data (whereby the Females figure is derived by difference),

[9] https://demo.istat.it/

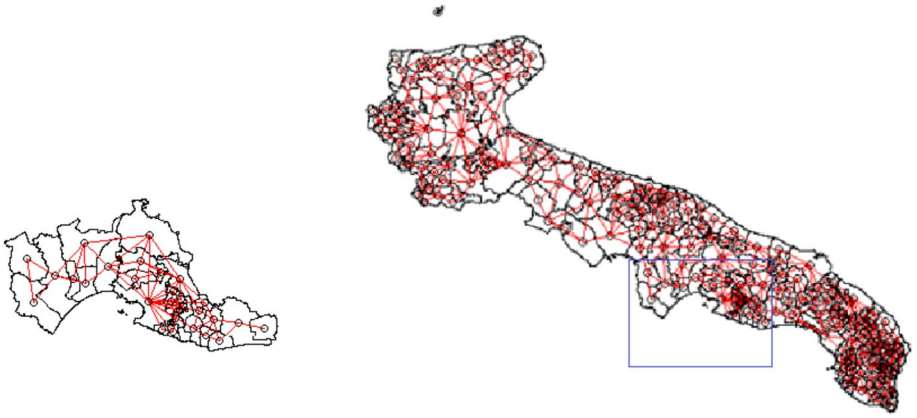

Fig. 2 Municipal nodes in Province of Taranto and Apulia Region

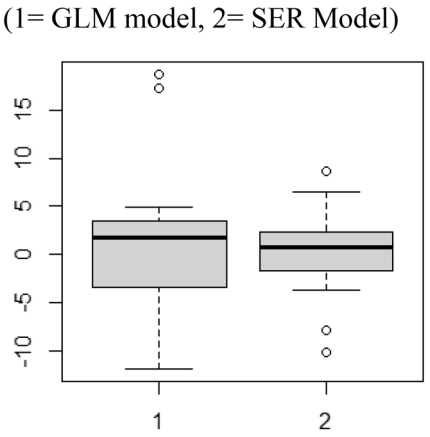

Fig. 3 Boxplot of regional residuals (1 = GLM model, 2 = SER Model)

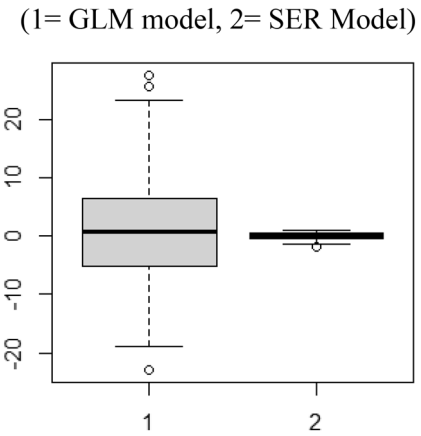

Fig. 4 Boxplot of provincial residuals (1 = GLM model, 2 = SER Model)

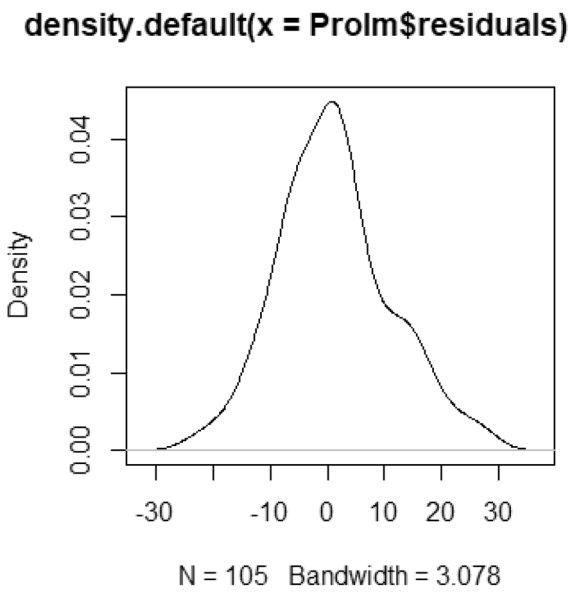

Fig. 5 GLM Model residual distribution

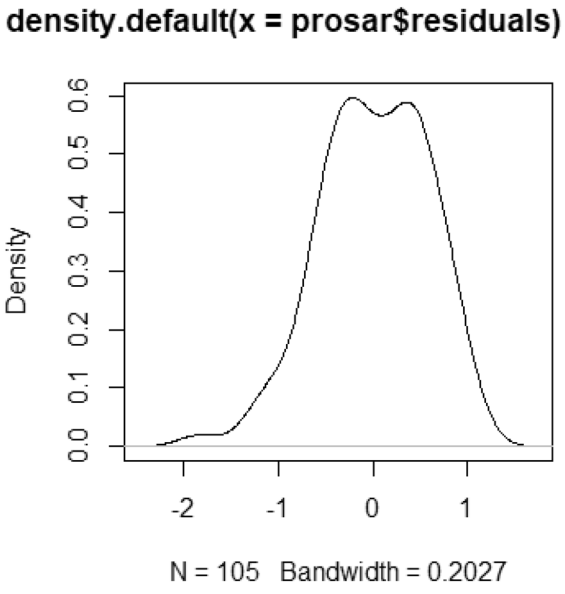

Fig. 6 SER Model residual distribution

it was considered to evaluate the auto regression spatial model only for Males on e°, thus skimming the variables referring to that sex (Figs. 1, 2, 3, 4, 5, 6).

The analysis data used were the Istat territorial bases,[10] where there are all the geographical reference shape files, based on the administrative units of work, from the Top

[10] www.istat.it/it/archivio/104317

of the Regions up to the Municipalities and the SMAs, for the largest municipalities. The granular basis of all data was the census section, for Istat data, which through linked operations on mixed queries[11] produced the reference shape files for analysis in R Cran. The e° of the reference administrative levels, down to the province, were also linked.

For the choice of the best explanatory variables as predictors of the dependent variable (e°), a series of simple, concatenated linear regressions of life expectancy at birth in 2011 was run, starting first with all 152 explanatory variables, gradually eliminating those less representative, where the coefficient was zero or very close to zero, since the model with a very high number of variables does not allow, in R, the definition of the S.E, T test and *P* value. Regression models were estimated on the provincial data. At the first step of the 152 variables, those referring to Males were first selected and then reduced for those with coefficient significantly equal to zero, with a 95% confidence level.

$$\hat{e}^{o}_{Mun} = \hat{\beta}_1 P_1 + \hat{\beta}_2 P_2 + \cdots + \hat{\beta}_{31} E_{31} + \varepsilon_{Prov}$$

The remaining 67, 2 of code H, 2 of code P, 2 of code FS and 2 of code FO, in the next step were reduced to 20, more significant: 16 of code P and 4 of code FO, up to 99.9% confidence level.

Using an autoregressive spatial model (SAR) is useful both for analyzing the spatial relationships between observations of different areal levels, and for scaling knowledge to lower levels, with extrapolation of sub-area indicators. The SAR model considers observations at a specific location as influenced by observations in the spatial vicinity. This allows you to capture the spatial dynamics within the data.

The relationship between the spatial model and the temporal model, in terms of frequency conversion, allows greater usability of implementation and application of the SAR model, with the synchronization of spatial and temporal data, so that each spatial observation is associated with a moment in time specific, allowing the definition of time series from spatial distribution, however within the limits of the qualitative validity of the data and in the face of not high and homogeneous levels of variability between geographical units.

With a semi-parametric P-Spline model, such as space–time ANOVA, for spatial data, one can include a uniform space–time trend, a spatial lag of the dependent and independent variables, a time lag of the dependent variable and its lag spatial and an autoregressive noise of the time series. Specifically, we consider a spatio-temporal ANOVA model, disaggregating the trend into spatial and temporal main effects, as well as second and third-order interactions between them.

Having assessed the goodness of fit of the SAR model to the data, we use goodness-of-fit measures and statistical tests to determine whether the model can satisfactorily explain the spatial and temporal variations in the data.

Interpret SAR model coefficients to understand how spatial relationships influence temporal dynamics. For example, suppose you have a positive value in a SAR coefficient. In that case, this suggests that observations in the spatial vicinity have a positive effect on the value of the variable over time.

[11] Spatial on-shape file and CSV tables of all census variables on the census section, using the Population and Housing Census zip file and aggregating the 20 regional tables in a single query.

3 A spatial dependence linear modelling

Spatial autocorrelation measures the degree to which a phenomenon of interest is related to itself in space (Ayuga-Téllez et al. 2011). In other words, similar values appear close to each other, or clusters, in space (positive spatial autocorrelation) or close values are dissimilar (negative spatial autocorrelation). Zero spatial autocorrelation indicates that the spatial pattern is random (Drago and Hoxhalli 2020). We can express the existence of spatial autocorrelation with the following moment condition:

$$\mathrm{Cov}\left(\mathrm{y_i}, \mathrm{y_j}\right) \neq 0 \text{ for } \mathrm{i} \neq \mathrm{j}$$

with $\mathrm{y_i}$ and $\mathrm{y_j}$ being observations of a spatially localized random variate at position (i,j), one should either estimate N, from the N covariances of the N observations themselves, or perform heavy iterative computational methods. Alternatively, applying spatial econometric analysis methods, theory is extensively elaborated by Anselin and Bera (1998) and Arbia (2014) and the practical aspect is an updated version of Anselin (2003). We introduce some restrictions in defining for each data point a relevant "neighborhood set", which in spatial econometrics is operationalized through the matrix of spatial weights. The matrix usually denoted by **W** of size $\mathrm{N} \times \mathrm{N}$ is positive and symmetric denoting in the first of each observation those places that belong to its surroundings set as non-zero elements (Anselin and Bera 1998), Arbia (2014), with characterization:

$$W_{i,j} \begin{cases} 1 \text{ if } j \in N(i) \\ 0 \end{cases}$$

N(i) is the set of (spatial) neighbours with position j, with diagonal values equal to 0. The criteria of spatial specification of proximity are various, but the main ones[12] are essentially two: the criterion "Rook", where two units are neighbours if they share a side, and the criterion "Queen", where, instead, the two units are neighbours if they share a side or an edge.

The "queen" model, compared to the "tower" model, allows the establishment of more links between adjacent areal nodes, especially in territorial situations where the geometries are very variable in terms of areas, dimensions, and shapes. This is very noticeable at the lower levels, such as for the provinces and municipalities, which are our main analysis objectives. The “queen” model guarantees better levels of quality of determination and validity of the model and results.

In addition to the neighbourhood location criterion, there is the distance evaluation criterion, within ranges defined by $j \in N(i) \text{se} d_{i,j} < d_{max}$ determined a priori. In our study we use the *queen* criterion of the `R` packages spdep, spatialreg, rgdal, maptools, leaflet and RcolorBrewer, to get the weights matrix we use two functions, `poly2nb and nb2listw,` the first one that builds a list of neighbours, if the `queen=TRUE` option is specified it will be built using the queen criterion, the second one for calculating w for the spatial weights. In this article, the method used of node proximities was only the "Queen" one, at the Regional, Provincial, Municipal and SMA levels.

Processing was done with R codes in R Studio version 1.4.1717: the principle package is *spdep* (spatial dependence: weighting schemes, statistics and model), a collection of functions to create spatial weights matrix objects from polygon contiguities, from point patterns by distance and tessellations, for summarizing these objects, and for permitting their use in spatial data analysis, including regional aggregation by minimum spanning

[12] Used in R and Payton packages.

tree; a collection of tests for spatial autocorrelation, including global Moran's Test; package *spData* to import and use territorial bases, census files and ESRI files; *sp* for generic spatial data analysis. The script of the software is in appendix.

4 Results

The application of SER at the regional level, shows only a live improvement of lower variability within the first and third quartiles, but this is since the number of regions is only 20, and the distance nodes created are few, there are 2 regions (Sardinia and Sicily) isolated, so only 62% of the nodes are non-zero. Much more significant is the analysis of nodes and weights at the provincial level, which instead generates a relevant number of nodes, with a good matrix of W (provincial) weights, with as many as 490 valid, non-zero nodes, generating 95% non-zero W_p weights. The difference in the nodes and weights generated, brings much more efficiency in the provincial level, with a very strong attenuation of the variability and covariability of the predictors, which is strongly attenuated in the SER spatial autoregressive model set up. This poses a strong quality of the produced estimates, which at the provincial level can be evaluated in the estimation error, compared to the actual value of e° that at the provincial level is detected.

The elaboration on all Regions (20) determined neighbour's nodes of 62 number by nonzero, percentage nonzero weights of 15.5, an average number of links of 3.1 and 2 regions with no links: the Islands of Sicily and Sardinia. The elaboration on all Provinces (110) determined neighbour's nodes of 490 number by nonzero, percentage nonzero weights of 4.05, an average number of links of 4.45 and 0 provinces with no links. The elaboration on all Municipalities (8092) determined neighbour's nodes of 47,638 number by nonzero, percentage nonzero weights of 0.07, an average number of links of 5.88 and 14 Municipalities with no links (a little municipal island). The graphical representation of the total nodes on municipalities is very dense, so we prefer to report an example on a single region (Apulia, with 258 nodes of municipalities, 1364 links and 1 municipality with no links (Tremiti Island). The application of the Chow-Lin type Top Down territorial disaggregation method, to the life expectancy indicator at birth and e° detected by Istat up to the administrative level of the provinces, from the life tables, allows to make an assessment of the robustness and efficiency, through the comparison between the variable elaborated by Istat, up to the provincial level, with the estimate of the indicator and the residuals of both linear regression (with lm package of R) and of autocorrelated spatial regression (lagsarm of R).

Among the best predictors of the census and income variables, 8 were selected which demonstrated, compared to others, a good interaction in terms of active regression on the estimation variable, with no know term. Figure 7 shows the spatial structure of life expectancy at birth on a municipal basis: it assumes a very linear and continuous trend, with a very detailed level of information; the low variability and the Moran Test show a strong robustness of the disaggregated indicator.

The SAR model used was developed on 2011 census data and variables, with integration of tax data on disposable income, and defined with a two-step analysis.

All census variables were preliminarily tested to determine the main components based on the p-value and the model determination index. The census variables that presented the best relationships were: P2, P18, P32, P44, P54, P64 and P66:

The proxy on economic conditions, detected by the Remx income available at various levels, has shown great reliability. The model has no intercept:

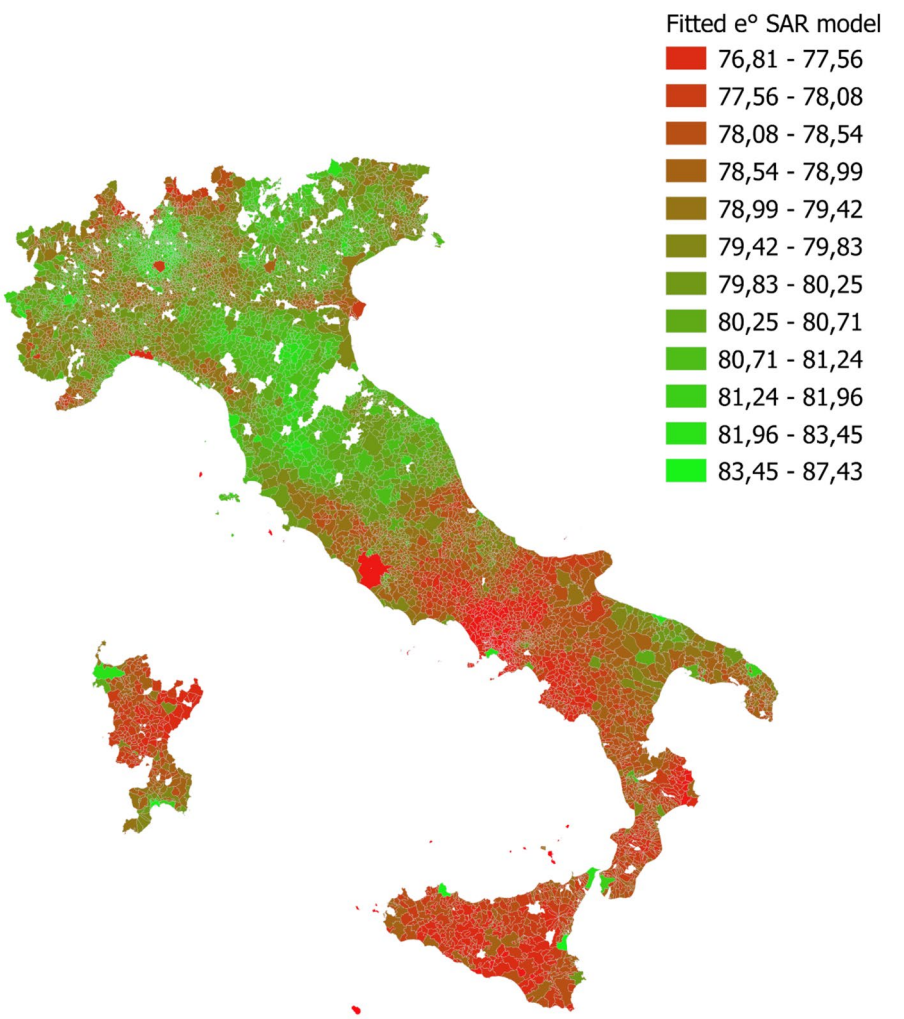

Fig. 7 Estimated e° with linear SAR model, in municipalities

$$\hat{e}^{o}_{Mun} = \hat{\beta}_1 disp_{income} + \hat{\beta}_2 P_2 + \hat{\beta}_3 P_{18} + \hat{\beta}_3 P_{18} + \hat{\beta}_4 P_{32} + \hat{\beta}_4 P_{44} + \hat{\beta}_5 P_{54} + \hat{\beta}_6 P_{64} + \hat{\beta}_7 P_{66} + \epsilon_{Prov}$$

Coefficients:
Estimate Std. Error t value Pr(>|t|).
Remp 5.917e − 03 1.591e − 04 37.196 < 2e − 16***
P18 3.730e − 03 9.600e − 04 3.885 0.000188***
P2 − 1.179e − 03 3.984e − 04 − 2.959 0.003889**
P32 − 1.286e − 03 2.408e − 03 − 0.534 0.594462.
P44 6.720e − 03 2.255e − 03 2.981 0.003644**
P54 5.885e − 04 2.988e − 04 1.969 0.051792.
P64 1.157e − 03 3.684e − 04 3.141 0.002242**
P66 7.105e − 04 6.773e − 04 1.049 0.296770.

5 Managerial implications

The managerial implications of this paper are that a top–down disaggregation method using linear self-regressive spatial models can be used to estimate life expectancy at the municipal level using census data from 2011 in Italy. This method utilizes census variables and accident data as predictors and is based on the assumptions of structural similarity, error similarity, and reliable indicators. The use of this method allows for a more detailed understanding of life expectancy at the local level and can inform policy decisions related to public health and well-being in specific areas. Additionally, the use of census data and the ability to estimate life expectancy at sub-municipal levels may also provide valuable information for businesses and organizations looking to make decisions about investments or operations in specific regions. Moreover, life expectancy at sub-municipal levels can have a variety of impacts on businesses and organizations considering environmental energy efficiency (Li et al. 2022) and excessive consumption (Yang et al. 2022). For example, areas with higher life expectancies may have more older residents who are more likely to be at home during the day and therefore more likely to use energy, while areas with lower life expectancies may have a higher proportion of working-age individuals who are away from home during the day and therefore less likely to use energy. Additionally, areas with lower life expectancies may have lower economic resources, making it more difficult for residents to invest in energy-efficient technologies or make other changes to reduce energy consumption. These factors could influence decisions made by businesses and organizations regarding energy efficiency and consumption.

6 Limitations and further research

There are several limitations to research applying a methodology of "Top–Down spatial disaggregation, using census data from 2011 in Italy." *Data availability*: The study relies on census data from 2011, which may not be representative of the current population or may not include all relevant information. Additionally, the use of ISTAT elaborations of annual mortality tables may not include all necessary data for the analysis. Further research could use more recent data and explore other data sources to improve the robustness of the results. *Assumptions*: The study relies on several assumptions, such as structural similarity and error similarity, that may not always hold true in reality. Further research could explore alternative methods or models that do not rely on these assumptions. *Spatial correlation*: The study assumes that errors are spatially correlated and that this correlation structure is the same at both aggregate and disaggregate levels. Further research could use more advanced models to capture spatial correlation, such as spatial econometric models. *Model limitations*: The study uses a linear self-regressive spatial model, which may not be appropriate for all cases and could lead to oversimplification of the data. A limitation of the model is that it assigns equal weight, and always, to all adjacent nodes, on the "queen" and "rook", therefore the same $W_{i,j}$. By introducing a discriminant on the distance between nodes on $W_{i,j}$, benefits could be created in terms of the validity of the SAR model. Given that all relationships between nodes generate, however, interactions and dependencies both between the nodes and in the border sub-nodal units, the weight of each node cannot be eliminated or significantly reduced. Considering the barycentric distances of the nodes $d_{i,j} > 0$:

$$W_{i,j}\begin{cases} \frac{2-\frac{\max d_{i,j}}{d_{i,j}}}{2} & \text{if} j \in N(i) \\ 0 \end{cases}$$

Future research could consider using alternative models such as non-linear models. *Scale of analysis*: The study focuses on municipal levels, which may not be granular enough to capture all the relevant variations in life expectancy. Further research could consider sub-municipal level (Census Area) where the variations are also present. It is important to note that these limitations should be considered when interpreting the results and conclusions of the study. Future research could aim to address these limitations by using more recent data, incorporating more variables, and using more sophisticated models. Moreover, the next step would be useful to analyse country-specific contributions to the increase of the best-practice life expectancy (Nigri et al. 2022). For example, could be useful in additional multi-country clustering-based forecasting of healthy life expectancy (HLE). In fact, according to Levantesi et al. (2023), the HLE is an indicator that measures the number of years individuals at a given age are expected to live free of disease or disability.

7 Conclusions

The determination of granular spatial data, particularly public health data such as life expectancy, is essential for both decision makers and public institutions, and for researchers. Istat, like many other institutes in Europe, has all the information to directly provide data on life expectancy on a municipal basis, without the need to estimate any breakdown, but unfortunately it does not make them public. While, at the processing level, it could use this method to spatially disaggregate data from the municipal level to the sub-municipal level, with much more precision than we could do in this application, as we have moved from the provincial level (NUT3) to the municipal level. Integrating life expectancy data in small areas into the official public statistics of the National Statistical Institutes would result in much deeper and more precise levels of health knowledge. In the present paper, we address an innovative topic and application framework for measuring, by the definition of territorial indicators such as the biometric function of citizens life expectancy, the performance at a highly disaggregated level. Moreover, we apply a methodology of Top–Down spatial disaggregation, offering a perspective that has been pursued to a limited extent. Considering a decentralized country such as Italy as a case study and micro-territorial information, a composite indicator of citizens life expectancy at the municipal and sub-municipal level was first proposed. Then, using a spatially disaggregated methodology, derived from the well-known Chow–Lin techniques, a municipality-level indicator was estimated to identify the citizens life expectancy. This method avoids artificial assumptions, and thus provides objective results that successfully realize self-regressive linear models in the spatial context (Tang et al. 2021). The potentially very high number of explanatory variables in spatial regression needs to be addressed by some dimension reduction method (Xia and An 1999). At present time, projection pursuit (PP) appears to be the most studied dimension reduction method. Some review on the topic might be found in Sun (2006), Jee (2009) and Loperfido (2018, 2019). To the best of our knowledge, however, there are no papers applying PP to spatial data. We are currently investigating an extension of the works of Galeano et al. (2006) and Loperfido (2020), who applied PP to multivariate time series.

Appendix: Measurement Model

Percentage nonzero weights: 4.049587
Average number of links: 4.454545

```
# estimates e° in Province areas, whit linear regression model, without constant term
summary(prosar)

Call: lagsarlm (formula = Males ~ 0 + Rempp + P18 + P2 + P32 + P44 +
   P54 + P64 + P66, data = ProM, listw = WqProv, zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-1.850294 -0.351893  0.032442  0.441068  1.121787

Type: lag
Coefficients: (asymptotic standard errors)
       Estimate  Std. Error z value  Pr(>|z|)
Rempp  7.9046e-05  2.8244e-05  2.7987 0.0051311
P18    1.1400e-04  5.1827e-05  2.1997 0.0278291
P2    -4.6252e-05  1.8613e-05 -2.4849 0.0129581
P32   -4.7723e-05  9.9489e-05 -0.4797 0.6314550
P44    1.8092e-04  9.5241e-05  1.8995 0.0574926
P54   -3.4001e-06  8.1705e-06 -0.4161 0.6773014
P64    5.2059e-05  1.5683e-05  3.3194 0.0009022
P66    5.2138e-05  3.7408e-05  1.3938 0.1633928

Rho: 0.98482, LR test value: 545.05, p-value: < 2.22e-16
Asymptotic standard error: 0.0043815
   z-value: 224.76, p-value: < 2.22e-16
Wald statistic: 50519, p-value: < 2.22e-16

Log likelihood: -120.3915 for lag model
ML residual variance (sigma squared): 0.32347, (sigma: 0.56874)
Number of observations: 105
Number of parameters estimated: 10.
AIC: 260.78, (AIC for lm: 803.84)
LM test for residual autocorrelation
test value: 6.8492, p-value: 0.0088681


Call: lagsarlm (formula = Males ~ 0 + Remc + P18 + P2 + P32 + P44 +
   P54 + P64 + P66, data = ComM, listw = WqCom, zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-8.582508 -0.497057  0.001314  0.748648 76.614372

Type: lag
Regions with no neighbours included:
 724 1814 4689 4948 5462 5607 6797 7377 7382 7423 7471 7558
Coefficients: (asymptotic standard errors)
       Estimate  Std. Error z value  Pr(>|z|)
Remc  2.9793e-04  1.5489e-05 19.2348 < 2.2e-16
P18  -2.7258e-04  5.8290e-04 -0.4676 0.6400463
P2    1.9541e-04  1.8181e-04  1.0747 0.2824877
P32  -1.1298e-03  1.1102e-03 -1.0176 0.3088496
P44  -6.3657e-04  9.6475e-04 -0.6598 0.5093665
```

P54 9.1153e-05 7.0009e-05 1.3020 0.1929091
P64 -3.0804e-04 1.7563e-04 -1.7539 0.0794398
P66 1.7646e-03 4.7855e-04 3.6875 0.0002265

Rho: 0.95563, LR test value: 26570, p-value: < 2.22e-16
Asymptotic standard error: 0.002313
z-value: 413.15, p-value: < 2.22e-16
Wald statistic: 170690, p-value: < 2.22e-16

Log likelihood: -21293.56 for lag model
ML residual variance (sigma squared): 9.2905, (sigma: 3.048)
Number of observations: 7962
Number of parameters estimated: 10
AIC: 42607, (AIC for lm: 69175)
LM test for residual autocorrelation
test value: 80.695, p-value: < 2.22e-16

Call:
lm(formula = Males ~ 0 + Remp + P18 + P2 + P32 + P44 + P54 + P64 + P66 + data = Proc)

Residuals:
Min 1Q Median 3Q Max
-22.9665 -5.0705 0.8681 6.5524 27.6547

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*Remp 5.917e-03 1.591e-04 37.196 < 2e-16 ****
*P18 3.730e-03 9.600e-04 3.885 0.000188 ****
*P2 -1.179e-03 3.984e-04 -2.959 0.003889 ***
P32 -1.286e-03 2.408e-03 -0.534 0.594462
*P44 6.720e-03 2.255e-03 2.981 0.003644 ***
P54 5.885e-04 2.988e-04 1.969 0.051792 .
*P64 1.157e-03 3.684e-04 3.141 0.002242 ***
P66 7.105e-04 6.773e-04 1.049 0.296770

*Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1*

Residual standard error: 10.28 on 96 degrees of freedom
(5 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared: 0.9848, Adjusted R-squared: 0.9833
F-statistic: 689.6 on 9 and 96 DF, p-value: < 2.2e-16

Call: lm (formula = Total_e° ~ 0 + P1 + P10 + P14 + P32 + P35 + P40 + P44 + P46 + P54 + P60 + P61 + P62 + P64 + P65 + P66 + P9 + FO2 + FO4 + FO6 + FO7, data = Pw)

Residuals:

Min	1Q	Median	3Q	Max
-62.295	-6.557	13.528	34.093	66.994

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
P1	-0.013157	0.066423	-0.198	0.843547	
P10	0.003042	0.002203	1.381	0.171586	
P14	0.004733	0.076983	0.061	0.951149	
P32	-0.041234	0.015080	-2.734	0.007885	**
P35	-0.010564	0.008683	-1.217	0.227751	
P40	-0.015029	0.010259	-1.465	0.147361	
P44	-0.028674	0.009900	-2.896	0.005013	**
P46	0.014211	0.066285	0.214	0.830852	

P54	-0.005461	0.001501	-3.637	0.000518	***
P60	-0.035880	0.016522	-2.172	0.033223	*
P61	0.035226	0.016410	2.147	0.035238	*
P62	0.031564	0.017882	1.765	0.081835	.
P64	0.075714	0.034889	2.170	0.033342	*
P65	-0.073013	0.033984	-2.148	0.035090	*
P66	-0.064839	0.037173	-1.744	0.085443	.
P9	0.006033	0.003605	1.674	0.098599	.
FO2	0.070818	0.020903	3.388	0.001152	**
FO4	0.001899	0.006667	0.285	0.776608	
FO6	-0.072932	0.023764	-3.069	0.003041	**
FO7	-0.081428	0.021317	-3.820	0.000283	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 1

Residual standard error: 34.94 on 71 degrees of freedom
(19 observations deleted due to missingness)
Multiple R-squared: 0.8588, Adjusted R-squared: 0.819.
F-statistic: 21.58 on 20 and 71 DF, p-value: < 2.2e-16

Variable code	Type
P1	Resident population - total
P9	Resident population - single
P10	Resident population - married or de facto separated
P14	Resident population - age <5 years
P32	Resident population - age 10 - 14 years
P35	Resident population - age 25 - 29 years
P40	Resident population - age 50 - 54 years
P44	Resident population - age 70 - 74 years
P46	Resident population - of 6 years and over
P61	Resident population - aged 15 and over employed (FL)
P62	Resident population - total of 15 years and more unemployed looking for new employment
P64	Resident population - age 15 and over from the labour force
P65	Resident population - age 15 and over employed (FL)
P66	Resident population - age 15 and over unemployed looking for new employment
FO2	Foreigners and stateless persons residing in Italy
FO4	Foreigners and stateless persons residing in Italy - age 30 - 54 years
FO6	Foreigners and stateless persons residing in Italy - males - ages 0 - 29 years
FO7	Foreigners and stateless persons residing in Italy - males - age 30 - 54 years

The software's script

required packages

```
requiere(maptools)
require(spdep)
require(leaflet)
library(spData)
library(spdep)
library(sp)
```

 # import Data Census by ISTAT's Territorial bases and Spatial Vectors ESRI

```
dsn <- system.file("vectors", package = "rgdal")[1]
cities <- readOGR(dsn=dsn, layer="Com2011_WGS84")
provinces <- readOGR(dsn=dsn, layer="Prov2011_WGS84")
regions <- readOGR(dsn=dsn, layer="Reg2011_WGS84")
taranto<- readOGR(dsn=dsn, layer="Taranto")
tarantoc<- readOGR(dsn=dsn, layer="Tarantoc")
comu<- readOGR(dsn=dsn, layer="ComuTa")
comuP<- readOGR(dsn=dsn, layer="ComuPuglia")

setwd("G:/Il mio Drive/Disag")
Regc<-read_xlsx("Regcod.xlsx")
Proc<-read_xlsx("Provcod.xlsx")
Comc<-read_xlsx("Comcod.xlsx")


# Regions Analysis - Application spatial model linear Lagsarlm

lqReg<-poly2nb(RegM, queen=TRUE)
WqReg<-nb2listw(lqReg , style="W", zero.policy=TRUE)
ltReg<-poly2nb(RegM, queen=FALSE)
WtReg<-nb2listw(ltReg , style="W", zero.policy=TRUE)
RegM<-merge(regions, Regc, by.x="COD_REG", by.y="CODREG")
Rem<-RegM$Rimp/RegM$P1
# spatial linear model Lagsarm on census viariables
regsar<-lagsarlm(Males~0 + Rem +P18+P2+P32+P44+P54+P64+P66,  data = RegM , listw =
WqReg, zero.policy = TRUE)
resreglm<-reglm$residuals
resregsar<-regsar$residuals
resreg<-cbind(reglm$residuals, regsar$residuals)
summary(WqReg$weights)
boxplot(resreg)
par(mfrow = c(1,2))
plot(reglm)
plot(regsar)

#province Italy - Analysis - Application spatial model linear Lagsarlm

ProM<-merge(provinces, Proc, by.x="COD_PROV", by.y="CODPRO")

Remp<-Proc$Rimp/Proc$P1
Prolm<-lm(Males~0 + Remp +P18+P2+P32+P44+P54+P64+P66+Rimp,  data = Proc)
lqProv<-poly2nb(provinces, queen=TRUE)
WqProv<-nb2listw(lqProv , style="W", zero.policy=TRUE)
Rempp<-ProM$Rimp/ProM$P1
# spatial model linear Lagsarlm
prosar<-lagsarlm(Males~0 + Rempp +P18+P2+P32+P44+P54+P64+P66,  data = ProM , listw =
WqProv, zero.policy = TRUE)
coefpr<-prosar$coefficients
res.prov.lm<-Prolm$residuals
res.pro.sar<-prosar$residuals
par(mfrow = c(1,2))
plot(density(Prolm$residuals))
```

```
plot(density(prosar$residuals))
fit.pro.sar<-prosar$fitted.values
difpro<-prosar$residuals
plot(Prolm$residuals)
boxplot(Prolm$residuals, prosar$residuals)
plot(difpro)
summary(Prolm)
summary(WqProv)

#Apulia Provinces Analysis - Application spatial model linear lagsarlm

ProMpuglia<-subset(ProM, COD_REG=="16")
RempP<-ProMpuglia$Rimp/ProMpuglia$P1
# spatial model linear Lagsarlm
Prolm<-lm(Males~0 + RempP +P18+P2+P32+P44+P54+P64+P66+Rimp,  data = ProMpuglia)
lqProvPuglia<-poly2nb(ProMpuglia, queen=TRUE)
WqProvPuglia<-nb2listw(lqProvPuglia , style="W", zero.policy=TRUE)
prosarPuglia<-lagsarlm(Males~0 + RempP +P18+P2+P32+P44+P54+P64+P66,  data =
ProMpuglia , listw = WqProvPuglia, zero.policy = TRUE)
coefprPuglia<-prosarPuglia$coefficients
valProLm<-Prolm$qr[1]
valProLm2<-Prolm$xlevels
lmProModel<-Prolm$model
lmProModel<-cbind(valProlmModel, Prolm$residuals, Prolm$fitted.values)
sarProModel<-NULL
sarProModel<-cbind(prosar$residuals, prosar$fitted.values)
#export data result
write.csv(lmproModel, "lmpromod.csv")
write.csv(sarProModel, "sarpromod.csv")


#Cities Analysis - Application spatial model linear lagsarlm

ComM<-merge(cities, Comc , by.x="PRO_COM", by.y="Codice Istat")
Remc<-ComM$Rimp /ComM$P1
# spatial model linear Lagsarlm
Comlm<-lm(Males~0 + Remc +P18+P2+P32+P44+P54+P64+P66,  data = ComM)
lqCom<-poly2nb(ComM, queen=TRUE)
WqCom<-nb2listw(lqCom , style="W", zero.policy=TRUE)
comsar2<-lagsarlm(Males~0 + Remc +P18+P2+P32+P44+P54+P64+P66,  data = ComM , listw
= WqCom, zero.policy = TRUE)
lmcom<-cbind(Comlm$model, Comlm$residuals, Comlm$fitted.values)
write.csv(lmcom, "lmcom.csv")
sarcomx<-comsar$X
sacomres<-cbind(comsar$residuals, comsar$fitted.values)
#export data result
write.csv(sarcomx, "sarcomx.csv")
write.csv(sacomres, "sarcomres.csv")

#cities Puglia - Analysis - Application spatial model linear Lagsarlm

ComMPuglia<-subset(ComM, COD_REG=="16")
```

```
RemcP<-ComMPuglia$Rimp /ComMPuglia$P1
ComMPuglia$P32*coefprPuglia[4]+ComMPuglia$P44*coefprPuglia[5]+ComMPuglia$P54*coef
prPuglia[6]+ComMPuglia$P64*coefprPuglia[7]+ComMPuglia$P66*coefprPuglia[8])
# spatial model linear Lagsarlm
ComlmP<-lm(Males~0 + RemcP +P18+P2+P32+P44+P54+P64+P66,  data = ComMPuglia)
lqComP<-poly2nb(ComMPuglia, queen=TRUE)
WqComP<-nb2listw(lqComP , style="W", zero.policy=TRUE)

# Estimed e° by Beta Provinces

Males_stim<-
ComM$Males*( ComM$Rimp/ComM$P1*coefpro[1]+ComM$P18*coefpro[2]+ComM$P2*coefpr
o[3]+ ComM$P32*coefpro[4]+ComM$P44*coefpro[5]+ComM$P54)
comsarP<-lagsarlm(Males~0 + Remc +P18+P2+P32+P44+P54+P64+P66,  data = ComM , listw
= WqCom, zero.policy = TRUE)
summary(Comlmp)
summary(comsarP)
mmer<-merge(ComMPuglia, comsarP$residuals, by.x=ComM$PRO_COM,
by.y=names(comsarP$residuals))
respro<-cbind(Prolm$residuals, prosar$residuals)
hist(respro[,1])
hist(respro[,2])
plot(density(respro[,1]))
plot(density(respro[,2]), add = TRUE)
boxplot(respro[1,], respro[2,])
boxplot(respro)
par(mfrow = c(1,1))
length(Comc)
intersec


#Linear Moran Test

moran.lm<-lm.morantest(Comlm, WqCom, alternative="two.sided", zero.policy = TRUE)
print(moran.lm)
LM<-lm.LMtests(Comlm, WqCom, test="all", zero.policy = TRUE)
print(LM)
impacts(ComM , listw=WqCom)
summary(Prolm)
summary(prosar)
respro<-cbind(Prolm$residuals, prosar$residuals)
hist(respro[,1])
hist(respro[,2])
plot(density(respro[,1]))
plot(density(respro[,2]), add = TRUE)
boxplot(respro[1,], respro[2,])
boxplot(respro)
```

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Declaration

Conflict of interest The authors have not disclosed any competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

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

Anselin, L.: Spatial externalities spatial multipliers and spatial econometrics. Int. Reg. Sci. Rev. **26**(2), 153–166 (2003)

Anselin, L., Bera, A.K.: Spatial dependence in linear regression models with an introduction to spatial econometrics. Stat.: Textb. Monogr. **155**, 237–289 (1998)

Arbia, G: Pairwise likelihood inference for spatial regressions estimated on very large datasets. Sp. Stat. **7**, 21–39 (2014)

Ayuga-Téllez, E., Contato-Carol, M., González, C., Grande-Ortiz, M., Velázquez, J.: Applying multivariate data analysis as objective method for calculating the location index for use in urban tree appraisal. J. Urban Plann. Dev. **137**(3), 230–237 (2011)

Bollino C.A., Polinori P.: Reconstructing value added at the municipal scale and growth paths at the micro-territorial level: the case of Umbria, *Rivista di Scienze regionali*, fascicolo 2 (2007).

Cervellera S., Cusatelli C.: Period life tables in suburban areas: the case of the Italian municipality of Taranto, *Rivista Italiana di Economia Demografia e Statistica,* Vol. LXXVI (1) gennaio-marzo 2022, CLEUP: Padova, ISSN: 0035–6832 (2022).

Cervellera S., Cusatelli C., Giacalone M.: Estimation of life expectancy in small areas using big data from the municipal registry. In: Pratesi, S.B., D'Ambrosio (Eds): SAE20/21 BIG4small - *Book of short papers, Edizioni Zaccaria,* Napoli, December 2021, ISBN: 9788899594138 (2021).

Chow, G.C., Lin, A.: Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. Rev. Econ. Stat. **53**(4), 372–375 (1971)

Drago C., Hoxhalli G.: *Measuring and Forecasting Job-Search in Italy Using Machine Learning* (2020).

Galeano, P., Peña, D., Tsay, R.S.: Outlier Detection in Multivariate Time Series by Projection Pursuit. J. Am. Stat. Assoc. **101**, 654–669 (2006)

Gennaro, V., Cervellera, S., Cusatelli, C., et al.: Use of official municipal demographics for the estimation of mortality in cities suffering from heavy environmental pollution: Results of the first study on all the neighborhoods of Taranto from 2011 to 2020. Environ. Res. (2022). https://doi.org/10.1016/j.envres.2021.112007

Giacalone, M.: Optimal forecasting accuracy using Lp-norm combination. Metron **80**(2), 1–44 (2021)

ISTAT: Infrastructure in Italy. A provincial analysis of the endowment and functionality, Rome (2006).

Jee, J.R.: Projection pursuit. Wiley Interdiscip. Rev. Comput. Stat. **1**, 208–215 (2009)

Levantesi, S., Nigri, A., Piscopo, G., Spelta, A.: Multi-country clustering-based forecasting of healthy life expectancy. *Qual. Quant.* 1–27 (2023).

Li, H., Appolloni, A., Dou, Y., Basile, V., Kopsakangas-Savolainen, M.: A parametric method to estimate environmental energy efficiency with non-radial adjustment: an application to China. *Ann. Oper. Res.* 1–27 (2022).

Loperfido, N.: Skewness-based projection pursuit: A computational approach. Comput. Stat. Data Anal. **120**, 42–57 (2018)

Loperfido, N.: Finite mixtures, projection pursuit and tensor rank: a triangulation. Adv. Data Anal. Classif. **13**(1), 145–173 (2019)

Loperfido, N.: Kurtosis-based projection pursuit for outlier detection in financial time series. Eur. J. Finan. **26**, 142–164 (2020)

Mazziotta, C., Vidoli, F.: La costruzione di un indicatore sintetico ponderato Un'applicazione della procedura Benefit of Doubt al caso della dotazione infrastrutturale in Italia. Italian J. Region. Sci. (2009a). https://doi.org/10.3280/SCRE2009-001002

Mazziotta, C., Vidoli, F.: Robustness and spatial stability of infrastructure endowment indicators: a test for Italian provinces.In: *XXX Conferenza Italiana di Scienze Regionali*, Florence (2009b).
Nigri, A., Barbi, E., Levantesi, S.: The relay for human longevity: Country-specific contributions to the increase of the best-practice life expectancy. Qual. Quant. **56**(6), 4061–4073 (2022)
Polasek W., Sellner R.: Spatial Chow-Lin methods: Bayesian and ML forecast comparisons. In: Rimini Centre for Economic Analysis (RCEA), working paper. pp. 38–08 (2008).
Sun, J.: Projection Pursuit. Encyclopedia of Statistical Sciences. Vol. 10 (2006).
Tang, Q., Wang, J., Jing, Z.: Tempo-spatial changes of ecological vulnerability in resource-based urban based on genetic projection pursuit model. Ecol. Ind. **121**, 107059 (2021)
Xia, X., An, H.Z.: Projection pursuit autoregression in time series. J. Time Ser. Anal. **20**(6), 693–714 (1999)
Yang, Y., Appolloni, A., Ding, X., Basile, V., Ma, H.: The influence of excessive consumption on residents' family thriving: the roles of intergenerational poverty transmission and educational cognition. Ann. Oper. Res. (2022). https://doi.org/10.1007/s10479-022-05106-3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Vincenzo Basile[1] · Stefano Cervellera[2] · Carlo Cusatelli[3] · Massimiliano Giacalone[4]

✉ Vincenzo Basile
vincenzo.basile2@unina.it

Stefano Cervellera
stefano.cervellera@uniba.it

Carlo Cusatelli
carlo.cusatelli@uniba.it

Massimiliano Giacalone
massimiliano.giacalone@unicampania.it

[1] Department of Economics, Management, Institutions, Federico II University of Naples, Via Cinthia 21, 80126 Naples, Italy

[2] Department of Computer Science, University of Bari "Aldo Moro", Via Edoardo Orabona 4, 70125 Bari, Italy

[3] Ionian Department, University of Bari "Aldo Moro", Via Duomo 259, 74100 Taranto, Italy

[4] Department of Economics, University of Campania "Luigi Vanvitelli", Corso Gran Priorato di Malta 1, Capua 81043, Caserta, Italy