

Spatial Composite and Disaggregate Indicators: Chow-Lin Methods and Applications

Francesco VIDOLI¹, Claudio MAZZIOTTA²

Abstract. The present paper aims to verify a statistical procedure that allows for the transition from aggregate indicators to disaggregate indicators, whereas in this case, the terms aggregate/disaggregate refer to territorial areas and not to categories. More specifically, an attempt was made to reconstruct the indicators of the infrastructural endowment at the provincial level (in Italy) on the basis of analogous indicators that are available at the regional level. For such purposes, Chow-Lin's approach was followed, and some regressors representing the territorial demand for infrastructure, related to productive, demographic and tourist aspects were used. The comparison between the results obtained through the application of the mentioned approach and the "real" data (available at the provincial level) in effect allows for the verification of the coherence among factors of demand and the real distribution of the infrastructure in Italy.

Key words: Spatial Data Analysis, Composite Indicators Disaggregation, Chow-Lin approach.

1. Introduction

The present paper aims to verify if and to what extent the methods of spatial disaggregation, in the form derived from Chow-Lin's approach, allow for the reconstruction in an appreciably precise way of the infrastructural index at the disaggregate level on the basis of the corresponding indicator at the superior territorial level. This essentially means verifying if the infrastructural levels of territorial units (the Italian provinces) may be "explained" by what should be their natural factors of

¹ Francesco Vidoli, Università degli Studi Roma Tre, Dipartimento di Istituzioni pubbliche, Economia e Società, e-mail: f.vidoli@sose.it

² Claudio Mazziotta, Università degli Studi Roma Tre, Dipartimento di Istituzioni pubbliche, Economia e Società, e-mail: c.mazziotta@uniroma3.it

generation, also referred to as “demand”, in particular the factors of a demographic and economic nature.

More specifically, having obtained an estimate of the provincial indicators of infrastructure on the basis of the aforementioned methods, such indicators are then compared with the “real” ones, available on the basis of the statistics provided by ISTAT (2006). As soon as this comparison presents a good proximity between the “real” data and the estimated ones, it may be deduced that the infrastructural provincial endowment conforms to corresponding demographic and economic factors of the demand; otherwise conclusions to the contrary shall be drawn.

2. Methodological approach

In order to derive indicators at the disaggregate level from the ones at the aggregate level we propose a model based on the approach presented by Chow-Lin (1971), that is founded on three fundamental hypotheses:

1. *Structural similarity*: the aggregate model and the disaggregate model are structurally similar. This implies that the relationships between the variables observed at the aggregate level are the same as those at the disaggregate level; that is the regression parameters in both models remain the same.
2. *Error similarity*: the spatially correlated errors present the same structure both at the aggregate level and the disaggregate level; that is to say that the spatial correlations are not significantly different.
3. *Reliable indicators*: the interpolating variables have sufficiently large predictive power at both the aggregate and disaggregate level, or R^2 (or the F test) of the regression model significantly differs from zero.

It is noted that violation of hypothesis 1 leads to attaining systematically biased estimates; violating hypothesis 2, instead, involves *spillover* effects that largely contribute to the estimates, and violating hypothesis 3 implies that the disaggregate estimates mirror the simple proportion of the aggregate ones.

Such models have been prevalently used for the construction of monthly or quarterly series, starting with the annual series, but in recent years these have been used in some applications in the spatial field as well. As for Italy, it is worth mentioning the work of Bollino and Polinori (2007), where they present the reconstruction of the Value Added at the municipal level in Umbria from the point of view of the convergence between suburban townships and urban municipalities that benefit from a higher growth, explained by contiguity factors and by agglomeration mechanisms.

The model is characterized both by an econometric relationship between the indicator at the provincial level and a series of explanatory variables observable at the disaggregate level (and also obviously at the aggregate one), and by a methodology of inferring unknown parameters. Mazziotta and Vidoli (2009b) tested a first application of the model to infrastructural data.

The model is based on the assumption that at the disaggregate level a linear econometric relationship is valid:

$$y_d = X_d \beta_d + \varepsilon_d \quad [1]$$

where: y_d represents a vector ($n*1$) of observations of the composite indicator at the disaggregate level, X_d is a matrix ($n*k$) of observations of k explanatory variables observable at the disaggregate level and n is the number of municipalities.

It is assumed that C is a matrix of dimension ($n*N$), where n is the number of the Italian provinces, capable of transforming the disaggregate observations into aggregate ones; such transformation may obviously be obtained through any operator.

In particular, if you choose the sum operator, provincial estimates are obtained for comparison with the corresponding regional values ($y_a = \sum y_d$) and the generic element $C_{i,j}$ is constructed as:

$$C_{i,j} = \begin{cases} 1, & \text{if province } i \in \text{region } j \\ 0, & \text{elsewhere} \end{cases}$$

If you choose the arithmetic mean operator, on the contrary, C is created as:

$$C_{i,j} = \begin{cases} \frac{1}{k}, & \text{if province } i \in \text{region } j, \text{ where } k = \text{number of provinces belonging to the region } j \\ 0, & \text{elsewhere} \end{cases}$$

and regional estimates are reconstructed through the average of provincial estimation³ [$y_a = E(y_d)$].

Therefore, assuming the hypotheses of *structural similarity* ($\beta_d = \hat{\beta}_a$), it is possible to write:

$$y_a = X_a \beta_d + \varepsilon_a \quad [2]$$

under the following aggregation constraints $y_a = C y_d$, $X_a = C X_d$ and $\varepsilon_a = C \varepsilon_d$.

Recently Polasek and Sellner (2008) presented an advancement, or better a very interesting generalization on the model, introducing a spatial auto-correlation term. Indeed, assuming that effects of spatial correlation exist in competitive levels among the provinces, but also and especially within the very similar provinces, then (see for example Anselin, 1988), given a matrix of spatial weights W_N and a spatial lag parameter $\rho \in [0,1]$, it is possible to hypothesize at the disaggregate level a “*mixed regressive spatial auto regressive*” relationship:

$$y_d = \rho_d W_N y_d + X_d \beta_d + \varepsilon_d \text{ with } \varepsilon_d \sim N[0, \sigma_d^2 I_N] \quad [3]$$

The reduced form of equation [3] lets us better appreciate the spatial component in which the contribution of X_d has been filtered through the spatial component.

$$y_d = (I - \rho_d W_N)^{-1} X_d \beta_d + (I - \rho_d W_N)^{-1} \varepsilon_d \quad [4]$$

More specifically, such *spatial filter* is applied proportionally to the distance; if in fact $(I - \rho_d W_N)^{-1}$ develops in series (similarly to an inverse matrix of Leontief), the following is obtained:

$$E(y_d | X_d) = (1 + \rho_d W_N + \rho_d^2 W_N^2 + \dots) X_d \beta_d \quad [5]$$

From equation [5], it is noted more easily how all the contiguous areas are involved in the estimate of y_d and that this occurs through a coefficient proportional to the distance (*distance decay*).

³ In this specific case, the arithmetic mean is used.

It is possible, therefore, to rewrite the reduced form of equation [4] with $\mathbf{R}_N = (\mathbf{I} - \rho_d \mathbf{W}_N)$.

$$y_d = \mathbf{R}_N^{-1} \mathbf{X}_d \beta_d + \mathbf{R}_N^{-1} \varepsilon_d, \quad \varepsilon_d \in N[0, \boldsymbol{\Sigma}_d] \quad [6]$$

with the $\boldsymbol{\Sigma}_d$ matrix of the covariance equal to:

$$\boldsymbol{\Sigma}_d = \sigma_d^2 (\mathbf{R}_N' \mathbf{R}_N)^{-1} \quad [7]$$

The unknowns terms of the models at the disaggregate level are therefore, the ρ_d , β_d and the σ_d^2 covariance. To estimate these unknowns, it is possible, according to the basic hypotheses, to exploit the relationship between y and $\mathbf{X}$ at the aggregate level and to estimate the mixed autoregressive model (see, for example LeSage, 1998) at the aggregate level, in the form:

$$y_a = \rho_a \mathbf{W}_N y_a + \mathbf{C} \mathbf{X}_d \beta_a + \varepsilon_a, \quad \varepsilon_a \sim N[0, \sigma_a^2 \mathbf{I}_N] \quad [8]$$

obtaining $\hat{\rho}_a$ and $\hat{\sigma}_a^2$.

As far as *structural similarity* ($\rho_d = \hat{\rho}_a$, $\beta_d = \hat{\beta}_a$) and *error similarity* ($\sigma_d^2 = \hat{\sigma}_a^2$) hypotheses, it is possible to substitute the estimated parameters in equations [3] and [6].

As far as the estimate of β_a like Chow-Lin's classic method, the following is obtained⁴:

$$\hat{\beta}_{a, GLS} = (\mathbf{X}_a' (\mathbf{C} \hat{\boldsymbol{\Sigma}}_d \mathbf{C}')^{-1} \mathbf{X}_a)^{-1} \mathbf{X}_a' (\mathbf{C} \hat{\boldsymbol{\Sigma}}_d \mathbf{C}')^{-1} y_a \quad [9]$$

and the estimate of y_d at the disaggregate level can be constructed as:

$$\hat{y}_d = \underbrace{\hat{\mathbf{R}}_N^{-1} \mathbf{X}_d \hat{\beta}_a}_{1^\circ \text{term}} + \underbrace{\hat{\boldsymbol{\Sigma}}_d \mathbf{C}' (\mathbf{C} \hat{\boldsymbol{\Sigma}}_d \mathbf{C}')^{-1} (y_a - \mathbf{C} \hat{\mathbf{R}}_N^{-1} \mathbf{C}' \mathbf{X}_a \hat{\beta}_a)}_{2^\circ \text{term}} \quad [10]$$

The first term of equation [10] therefore represents the *naïve* estimate of the unknown vector y_d , while in the second part of the equation the estimate error at the aggregate level is distributed through the “*gain projection matrix*” $\mathbf{G}$ (Goldberger, 1962).

$$\mathbf{G} = \hat{\boldsymbol{\Sigma}}_d \mathbf{C}' (\mathbf{C} \hat{\boldsymbol{\Sigma}}_d \mathbf{C}')^{-1} \quad [11]$$

This gain more crucially depends on the *spatial lag* parameter $\hat{\rho}_a$ at the aggregate level; note that if $\hat{\rho}_a = 0$, the $\hat{\boldsymbol{\Sigma}}_d$ matrix is equal to the matrix identity and it is reduced to the *projection matrix*: $\mathbf{G} = \mathbf{C}' (\mathbf{C} \mathbf{C}')^{-1}$ as in the base model provided in equation [1]; the ρ_d parameter and the $\mathbf{W}_N$ matrix therefore let the $1/N$ part of the residual at the aggregate level not be assigned equally to all the municipalities; instead it is filtered through the spatial weights matrix.

⁴ Please note that $\hat{\beta}_{a, GLS}$ are not dependent on $\hat{\sigma}_a^2$, but rather on $\hat{\rho}_a$.

3. An application to the infrastructure indicators

The application of the model illustrated above assumes the availability of the following information: i) synthetic infrastructural indicators at the regional level (and provincial, for the subsequent verification of the model's accuracy); ii) demographic and economic variables correlated with infrastructural needs, at the provincial and regional levels. The first ones are derived from the application of a particular method of synthesizing elementary indicators (source: ISTAT) applied by the authors in a previous work (Mazziotta and Vidoli, 2009a). The second ones are reported in Table 1.

On the basis of such data, we estimate a spatial simultaneous autoregressive lag and a mixed model at the aggregate level (regional) $y_a = \rho_a W_N y_a + C X_d \beta_a + \varepsilon_a$ in order to obtain a good level of efficiency, according to hypothesis 3 of the reliable indicators.

The estimated regression shows that the model is satisfactory (Table 2 shows the results for the coefficients of independent variables are all significant, the value of ρ equal to 0,37 (p-value 0,043) and goodness of fit estimated by R^2 equal to 0,43) both for the meaning of the variables included as regressors (*proxy* of the economic development, of demographic density and of the supply of qualified tourism) and for the verified statistical properties. Having once obtained $\hat{\beta}_a^2$, the estimated standard deviations $\hat{\sigma}_a^2$ and the parameter $\hat{\rho}_a$, thanks to the hypothesis of *structural similarity* and of *error similarity*, such parameters have been substituted in the equation [10] for the purpose of obtaining the estimated infrastructural indicator at the disaggregate level.

Table 1: Generation factors of infrastructural endowment

<i>Variable</i>	<i>Source</i>	<i>Year</i>
Total resident population	ISTAT	2007
Extra-agriculture Value Added (share on total)	Ist. Tagliacarne	2003
Share of population residing in municipalities with more than 30 thousand inhabitants	ISTAT	2007
Share of population residing in municipalities with more than 50 thousand inhabitants	ISTAT	2007
Accommodation capacity of low to mid range hotels per inhabitant	(⁵)	2006
Accommodation capacity of luxury hotels per inhabitant	(⁵)	2006
Average stay of foreign visitors	(⁵)	2006

⁵ Such indicators were calculated by F. Vidoli and L. Taffara for a research project financed by Istituto Tagliacarne, aimed at recreating levels of competitiveness at urban and territorial levels (2008 and 2009).

Table 2: Estimated model* results at the regional level, $\rho = 0,37$, $R^2 = 0,43$

Variable	Estimation	Std. Error⁶	z value	Pr(> z)
Gross domestic product	7,893E-06	0,000	2,1024	0,0355
Share of the population residing in municipalities with more than 50 thousand inhabitants	3,916E-01	0,181	2,1587	0,0309
Accommodation capacity of low to mid range hotels per inhabitant	-1,182E-02	0,005	-2,2568	0,0240

* See equation [8].

The results obtained indicate a considerable gap when compared to the “real” data available at the provincial level; in other words it is deduced both from the graphic examination of the distributions (Figure 1) and from the application of a specific indicator of spatial robustness (called ISR, see Table 3). As far as regards the latter, it involves a spatial robustness indicator applied to the ranks (in this case, the positions held by individual provinces on the bases of infrastructural indicators – the “real” ones and those resulting from the model) that varies from 0 to 1 and is created in such a way as to highlight not only the average differences of rank, but also among which units territorially identified such differences have manifested themselves.

For greater detail, it is worthwhile referring to a previous work (Mazziotta and Vidoli, 2009b); here it is sufficient considering that such indicator is the result of the product of two matrices: the first (contiguity matrix, **W**) identifies the territorial contiguity of the units (the provinces, in our case) among them; the second (transition matrix, **T**) highlights both the differences in the ranks and from which unit and towards which other unit this difference has manifested itself.

Multiplying the **I-W** matrix by **T** yields an index that, comparing the ranks held by the province in the two situations considered (“real” infrastructural indicators and the calculated ones using the model), highlights changes in the ranks that only involved units spatially not contiguous.

The indicator of spatial robustness (IRS) between two ranking distributions, in the algebraic form, may be expressed as follows:

$$ISR_{R_0, R_1} = \frac{\sum_{i,j} T_{i,j}(1 - w_{i,j})}{Max_I} \quad [12]$$

It should be noted that the maximum of the proposed index (Max_I) equals the worst situation from the perspective of conformity of the two rankings, that is to say the one in which the unit i that was in the first position in the R_0 ranking finds itself in the last position in the R_1 ranking and so on, for as many non contiguous units $(n-1) * (n-2) * \dots$, or as many times as there is a value greater than zero in the matrix **T(I-W)**.

As in Table 3 and Figure 1, there are many marked differences between the two rankings, given that the ISR is equal to 0,38 with an average change in the ranks among different areas (in this specific case, regions) that is particularly high (24,9).

⁶ Numerical Hessian approximate standard errors.

Figure 1: Ranking of Italian provinces, based on “real” indicators (on the left) and “calculated” indicators (on the right).

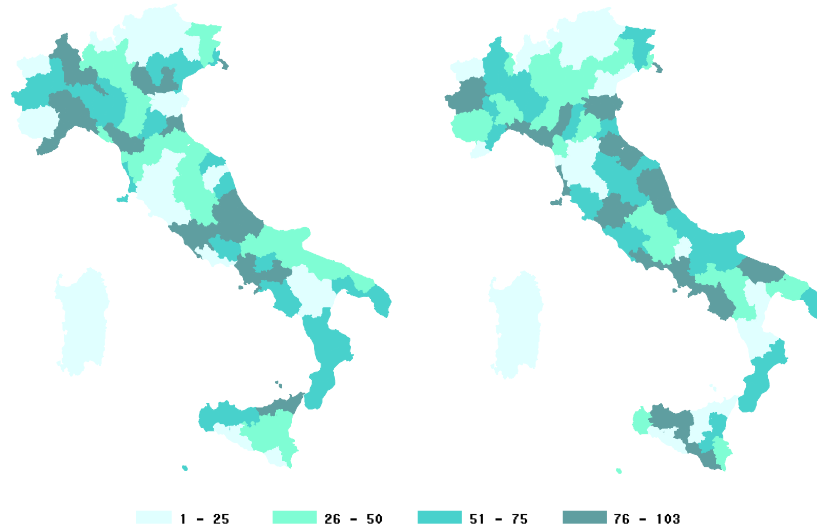


Table 3: Index of spatial robustness (ISR) of “real” indicators vs. the results from the model

	ISR	N° of extra-area ranking shifts	Extra-area mean ranking difference
Provincial indicator	0,384	75	24,9

4. Conclusions

Because of the variables used in the model application at the provincial level, it may be sustained that the results obtained tend to identify the levels of infrastructure that the factors of demand in each province require.

The objective of the work was to recreate the infrastructural indicators (of land transportation, in this specific case), at a level of high territorial detail (provincial) through the disaggregation of territorially superior levels (regional), obtained through the application of a model derived from the Chow-Lin approach, applied according to the version supplemented by Polasek and Sellner.

Actually, the use in such models of socio-economic variables like regressors (infrastructural demand factors), confers to the comparison of the two rankings – the one created on the basis of “real” provincial infrastructural data and the one “recreated” with the model – the meaning of comparison between the supply and demand of infrastructure at the territorially disaggregate level. Obviously, this meaning as much is founded as the model presents a high level of statistical fitness. In our application, the used goodness-of-fit measures indicate that the model results can be estimated

satisfactorily from a statistical point of view, considering the cross-section application and the territorial units. A better (or larger) selection of variables would be an important improvement in the model's application. At the present, in any case, a considerable gap between the estimated levels and those assumed to be "real" may be interpreted, even conservatively, as confirmation of mismatch existing between the demand and the supply for infrastructure expressed by the territory. And this seems, in the latter analysis, to be the result that is statistically more evident and economically more interesting: the territorial distribution of the infrastructural endowment of transport is not in line with the "theoretic" factors of generation that are present in the Italian provinces.

References

- Anselin L.: *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht (1988)
- Bollino C. A., Polinori P.: Ricostruzione del valore aggiunto su scala comunale e percorsi di crescita a livello micro-territoriale: il caso dell'Umbria, *Rivista di Scienze regionali*, fascicolo 2 (2007)
- Chow G. C., Lin, A.: Best linear unbiased interpolation, distribution, and extrapolation of time series by related series, *The Rev. of Economics and Statistics*, 53(4): 372-375 (1971)
- Goldberger A. S.: Best linear unbiased prediction in the generalized linear regression model, *American Statistical Association J.*, 57: 369-375 (1962)
- ISTAT.: *Le infrastrutture in Italia. Un'analisi provinciale della dotazione e della funzionalità*, Roma (2006).
- LeSage J. P.: *Spatial econometrics*, Technical report, University of Toledo (1998)
- 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. of Regional Science*, vol 8, n°1, Franco Angeli (2009a)
- Mazziotta C, Vidoli F.: Robustezza e stabilità spaziale di indicatori di dotazione infrastrutturale: una verifica per le province italiane, XXX Conferenza Italiana di Scienze Regionali, Firenze (2009b)
- Polasek W., Sellner R.: Spatial Chow-Lin methods: Bayesian and ML forecast comparisons, Rimini Centre for Economic Analysis (RCEA), working paper 38-08 (2008)