Bayesian hierarchical models for misaligned data: a simulation study

Modelli gerarchici Bayesiani per dati non-allineati: uno studio su dati simulati

Giulia Roli and Meri Raggi

Abstract In this paper, the problem of combining information from different sources of data is considered. We focus our attention on spatially misaligned data, where available information (typically counts or rates from administrative sources) refers to spatial units that are different from the ones of interest. A hierarchical Bayesian perspective is considered, as firstly proposed by Mugglin *et al.* in 2000, to provide a fully model-based approach in an inferential, and not only descriptive, sense. In particular, explanatory covariates are arranged to be modeled according to spatial correlation through a conditionally autoregressive prior structure. In order to assess model performance and its robustness we generate artificial data basing on a real study and a simulation exercise is then carried out.

Abstract *In questo lavoro si considera il problema dell'integrazione di dati provenienti da diverse fonti. In particolare, vengono presi in esame dati spaziali non-allineati, ovvero informazioni (solitamente conteggi o tassi provenienti da fonti amministrative) riferite a unità spazialmente differenti da quelle di interesse. Con questo obiettivo, i modelli gerarchici Bayesiani, così come proposti da Mugglin et al. nel 2000, forniscono un valido strumento statistico in termini sia descrittivi sia, soprattutto, inferenziali. Tale approccio prevede di utilizzare una serie di variabili esplicative attraverso una struttura di autoregressione che spieghi la correlazione spaziale e, di conseguenza, aiuti nell'integrazione delle informazioni. Il modello proposto viene testato anche in termini di robustezza su dati simulati facendo riferimento ad uno studio realmente svolto.*

Key words: Bayesian analysis, Misaligned data, Linking spatial information

Giulia Roli
Dipartimento di Scienze Statistiche, Università di Bologna, e-mail: g.roli@unibo.it

Meri Raggi
Dipartimento di Scienze Statistiche, Università di Bologna, e-mail: meri.raggi@unibo.it

1 Introduction

A challenging topic of spatial analysis is certainly represented by linking data collected at different scales, locations and dimensions. The great interest on this area is mainly due to practical reasons, drawing on works from geographic, ecological, environmental, agricultural and geological fields. In statistical literature several terms have been introduced to denote the different aspects of the overall problem of "incompatible" spatial data and various solution to it have been proposed. The mainly used terms are the ecological inference problem, the Modifiable Areal Unit Problem (MAUP), the change of support problem (COSP), areal interpolation ([5]; [1]). The proposed solutions depend on the study objectives and on the field of application (i.e. block and co-kriging, multiscale modeling, Bayesian hierarchical modeling). Gotway and Young ([4]) provided a complete multidisciplinary history of this topic with a description of early solutions in both geographical and statistical literature.

Here, we focus the attention on spatially misaligned data, where available information (typically counts or rates from administrative sources) refers to spatial areas that are different from the ones of interest. A hierarchical Bayesian interpolation, estimation and spatial smoothing is considered, as firstly proposed by Mugglin *et al.* in 2000 ([6]), over misaligned data grids both defining regions too large to be considered marked points and where one grid contains the other. The main aim is to convert the source information to target zones by exploiting a set of covariates on both grids. In order to assess performance of this method and its robustness we generate artificial data basing on a real study. Preliminary results are presented, different scenarios will be compared as further developments of the simulation exercise.

2 Motivating example

European projects require often ad hoc data collection for socio-economic or environmental analysis. Case studies usually refer to areas defined more on the basis of problem features than to administrative boundaries. Onerous effort in terms of money and time consuming is needed to collect original data in many case limited both by the area of interest and by very specific topics. This implies that the use of secondary data (for example administrative or census data) as covariates would be desirable and sometimes necessary, rising up the misaligned data problem.

As an example to motivate the needing of statistical methods aiming at integrating misaligned data, we refer to a real study at European level granted by European Commission in 7FP. This project, namely CLAIM (www.claimproject.eu) aims to provide the knowledge base to support an effective Common Agricultural Policy design in the direction of improved landscape management. The specific target is the assessment of the capacity of landscape in providing added value for society in rural areas. This objective is realized through a deeper insight on the relation between agriculture, landscape features and the socio-economic development of rural areas by means of 9 case studies in different European countries. In Italy the selected area

is Po Delta area where data were collected through two ad hoc surveys. Obviously the area of interest does not coincide with municipalities boundaries making quite inaccurate the use of administrative data as covariates. Then, on one hand secondary administrative data are easily available, also geographical referred, and very useful as explanatory variables; on the other hand they refer to different and larger areas than selected case study sites.

3 Modeling framework

Let consider two misaligned spatial grids, S_B and S_C, where the first one defines the so-called *source zones*, i.e. areas on which data of interest are available, and the second grid partitions regions for which data are to be imputed, namely *target zones*. Thus, the first grid is often referred to the *response grid* and the second one to the *explanatory grid*. We consider a special case with respect to the more general framework proposed by Mugglin *et al.* in 2000, where one data grid contains the other, that is, following the motivating example, the administrative data coverage is higher than the whole area of interest, say $S_C \subset S_B$. In this case, there will not exist C cells with portions lying outside of S_B.

In the first grid regions are indexed by i (with $i = 1, \ldots, I$) and denoted by B_i; similarly, for the second grid we have regions C_j, with $j = 1, \ldots, J$. The intersection of the two grids creates *atoms*, i.e., cells identifying partitions of both B_i and C_j regions. Atoms can be referenced relative to an appropriate B cell and denoted by B_{ik} (with $k = 1, \ldots, K_i$) or, equally, to an appropriate C cell by C_{jl} (with $l = 1, \ldots, L_j$). Thus, we can formally define the function f such that $f(B_{ik}) = C_{jl}$ and the inverse function g such that $g(C_{jl}) = B_{ik}$. Atoms with no intersection across the grids are *edge atoms*. In our case, only B-edge atoms could exist, say B_{iE}.

For each source zone B_i, we can observe the response Y_i. Referring to the notation introduced above, the main aim is then to convert Y_i to Y'_j, i.e. imputing values of Y to the target zones, by exploiting covariates which can be observed on the explanatory grid, X_j, and/or to the response grid, W_i. In this first attempt, let us consider only the availability of an X covariate, which is assumed to be an aggregated measurement similarly to the response Y. Under this hypothesis the observed values of Y_i can be regarded as $\sum_k Y_{ik}$, where Y_{ik} are latent values for the atoms associated with B_i. Similarly, $X_j = \sum_j X_{jl}$, where X_{jl} are unobserved according to the atoms associated with C_j.

In order to specify the model, we assume Poisson distributions for the observed measurements

$$X_j \sim Poi(e^{\omega_j}|C_j|) \tag{1}$$

$$Y_i \sim Poi\left(e^{\mu_i} \sum_k |B_{ik}| h(X'_{ik}/|B_{ik}|; \theta_{ik})\right) \tag{2}$$

where $|A|$ denotes the area of a generic region A, ω_j and μ_i are random effects capturing spatial association among the X_j's ad the Y_i's, respectively, $h(\bullet)$ is a selected parametric function depending on θ_{ik} and adjusting an expected proportional-to-area allocation according to X'_{ik}, which are the values of X_{jl} associated to the response grid. As a result, the conditional distribution of the latent variables X_{jl} and Y_{ik} given the observed is a product multinomial

$$\left(X_{j1}, X_{j2}, \ldots, X_{jL_j}\right) \sim mult(X_j; q_{j1}, q_{j2}, \ldots, q_{jL_j}) \tag{3}$$

$$(Y_{i1}, X_{i2}, \ldots, X_{iK_i}) \sim mult(Y_i; p_{i1}, p_{i2}, \ldots, p_{iK_i}) \tag{4}$$

where $q_{jl} = \frac{|C_{jl}|}{|C_j|}$ and $p_{ik} = \frac{|B_{ik}|h(X'_{ik}/|B_{ik}|;\theta_{ik})}{\sum_k |B_{ik}|h(X'_{ik}/|B_{ik}|;\theta_{ik})}$.

For B-edge atoms, B_{iE}, there is not a corresponding C_{jl}, thus a latent X'_{iE} is introduced whose distribution is defined by the adjacent non-edge atoms

$$X'_{iE} \sim Poi(e^{\omega_i^*}|B_{iE}|) \tag{5}$$

where ω_i^* are additional spatial random effects to be associated to the others ω_j.

In a fully Bayesian setting, prior distributions for each involved parameter need to be specified. To capture the spatial nature of B_i, a Markov random field prior for the μ_i's can be adopted ([2]). In particular, if the adjacency form is considered, we obtain a conditional autoregressive (CAR) prior ([3]); in the most simplest form, we refer to an exchangeable prior that captures heterogeneity but not local clustering across areas. Similar considerations can be referred to the set of spatial random effects for C_j regions, namely $\left\{\omega_j \omega_i^*\right\}$. For other (hyper-) parameters proper and vague priors are generally adopted (see [6] for more details).

4 Simulation study

In order to test the method described above, we generate an artificial data set basing on the practical purposes of the motivating example. As a first attempt, we consider a restricted number of areas for both grids, with the potential of being extended by using data from Global positioning systems (GPSs) and geographical information systems (GISs), as soon as available. We select a number of 3 source zones B_i and 4 target zones C_j, with $S_C \subset S_B$. The intersection of the two grids generates 3 B-edge atoms, B_{1E}, B_{2E} and B_{3E} and a total of 9 non-edge atoms, B_{ik} (with $k = 1, \ldots, K_i$ and $K_i = 3, 5, 4$) or C_{jl} (with $l = 1, \ldots, L_j$ and $L_j = 3, 2, 2, 2$). Figure 1 graphically resumes the grids.

Areas of regions and atoms are randomly generated. These are then employed to simulate data for the covariate X and the response Y for each atom through the multinomial distribution, given the parameters fixed for the distribution of $\{\mu_i\}$ and $\left\{\omega_j \omega_i^*\right\}$. Here, we first consider a simple case, where both the spatial random effects follow exchangeable priors.

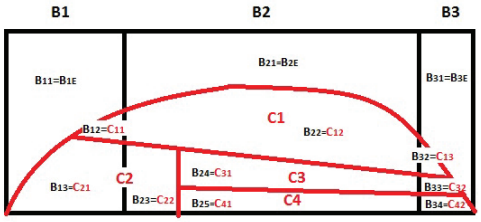

Fig. 1 Representation of Areal Data Misalignment of the simulation.

Parameter estimation is then implemented by considering the model specification introduced above adopting a MCMC approach and using the software WinBUGS ([7]). Referring to the seminal work of Mugglin *et al.* ([6]), we choose a function h for the computation of the p_{ik}'s ensuring non-null estimates of the response Y on the atoms, i.e. $h(X'_{ik}/|B_{ik}|;\theta_{ik}) = X'_{ik}/|B_{ik}| + \theta_{ik}$ with $\theta_{ik} = \frac{\theta}{K_i|B_{ik}|}$. We first set $\theta = 1$, considering different specifications in future developments.

Preliminary results are reported in Table 1. We consider the parameters of interest on response Y imputed for both target zones, Y'_j, and non-edge atoms associated to the explanatory grid, Y'_{jl}. We show and compare true and estimated parameters, together with 95% credibility intervals (CI).

Table 1 True parameters and estimates from hierarchical Bayesian model.

Parameters	True	Estimates	(95% CI)
Y'_1	27650	27590	(27320 - 27820)
Y'_2	2451	2426	(2367 - 2492)
Y'_3	8246	8263	(8192 - 8352)
Y'_4	53793	53690	(53190 - 54280)
$Y'_{1,1}$	2447	2375	(2309 - 2441)
$Y'_{1,2}$	13303	13160	(12800 - 13470)
$Y'_{1,3}$	11900	12050	(11850 - 12280)
$Y'_{2,1}$	75	73	(69 - 77)
$Y'_{2,2}$	2377	2353	(2295 - 2418)
$Y'_{3,1}$	5014	4969	(4836 - 5102)
$Y'_{3,2}$	3233	3294	(3227 - 3365)
$Y'_{4,1}$	34415	33960	(33200 - 34750)
$Y'_{4,2}$	19378	19730	(19380 - 20050)

The estimates seem to be quite precise, with acceptable relative biases (varying from a minimum of -2.9% to a maximum of 1.9%) and non-coverages of 95% CI only for two values.

5 Final remarks

In this work, we considered the problem of combining information from different sources of data focusing on spatially misaligned data. A hierarchical Bayesian modeling is used to convert the source information to target zones by exploiting a set of covariates on both grids. We applied this method to simple simulated data generated to resemble a real study. The estimates we obtained result to be quite precise, with acceptable relative biases.

As a future development, we would assess method robustness with respect to model misspecifications (e.g., with respect to prior assumptions) and further considering a larger number of areas in both grids. Finally, we would be interested into the application of the method to real data.

References

1. Banerjee, S., Carlin, B. P., and Gelfand, A. E.: Hierarchical Modeling and Analysis for Spatial Data. London: Chapman and Hall (2004)
2. Bernardinelli, L. and Montomoli, C.: Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Statist. Med. **11**: 983–1007 (1992)
3. Besag, J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Statist. Soc., Ser. B, **36**, 192–236 (1974)
4. Gotway, C.A., Young, L.J.: Combining Incompatible Spatial Data. J. Am. Statist. Ass. **97**(458): 632–648 (2002)
5. Lawson, A.B.:Bayesian disease mapping, CRC press (2009)
6. Mugglin, A.S., Carlin, B.P., Gelfand, A.E.: Fully Model-Based Approaches for Spatially Misaligned Data. J. Am. Statist. Ass. **95**(451): 877–887 (2000)
7. Spiegelhalter, D., Thomas, A., Best, N., Lunn, D.: WinBUGS User Manual, Version 1.4 (2003)