

Atom-based models for misaligned data integration and benchmarking in small area estimation

Matilde Trevisani and Alan Gelfand

Abstract We propose a class of misaligned data models for addressing typical small area estimation (SAE) problems. We extend recently proposed atom-based models for spatial misalignment to the SAE context enabling use of auxiliary covariates, which are available on areal partitions non nested with the small areas of interest (misaligned areal regression), along with planned domains survey estimates also misaligned with these small areas (misaligned areal interpolation). We model the latent characteristic of interest at atom level as a Poisson variate with mean arising as a product of population size and incidence. We introduce spatial random effects either using a CAR model or a process specification. For the latter, incidence is a function of a Gaussian process model for the spatial point pattern over the entire region. Atom counts are driven by integrating the point process over atoms. In the proposed class of models benchmarking to large area estimates is automatically satisfied.

Key words: small area estimation; spatial misalignment; hierarchical Bayesian methods; atom-based models

1 Introduction

This paper considers small area estimation (SAE) when multiple data sources, possibly misaligned with small areas, are available for inference. Supplementary sources of relevant information may consist either in auxiliary data (from census or admin-

Matilde Trevisani

Department of Economics, Administration, Mathematical and Statistical Sciences, University of Trieste, 34100 Trieste, e-mail: matildet@econ.units.it

Alan Gelfand

Department of Statistical Science, Duke University, Durham, NC 27708-0251, e-mail: alan@stat.duke.edu

istrative archives) or in survey data collected on planned or major domains. Both these types of data may be available at a level of spatial resolution incompatible or misaligned with that of small areas. Hence, developing models to combine most appropriately and efficiently any source of potentially important information becomes of crucial concern for research in SAE.

Under the hierarchical Bayesian (HB) framework, misaligned data models have been recently proposed to address the issue of making inference for domains (target zones) different from the ones (source zones) for which data are available. In particular, we consider the so-called *atom-based models* and extend them to originally address typical SAE questions: (i) combining auxiliary covariates which are available on nonnested areal partitions (misaligned areal regression problem); (ii) providing small area estimates by using planned domain data (misaligned areal interpolation problem).

We offer two versions of the proposed class of atom-based models. The first one, which we refer to as *Small $\odot$ Large area estimation model* (S $\odot$ LAE for short), introduces area-specific effects (not accounted for by auxiliary covariates) as heterogeneous and/or spatial random effects (these last e.g. by a conditionally autoregressive prior specification), as is customarily done in mixed effects models. The second, the so-called *Small $\odot$ Large area estimation model* (S $\odot$ LAE) fits, over the entire region, an intensity surface which drives a latent spatial point pattern of locations within atoms representing the “cases” of interest (e.g. of unemployment as in our example). Expected atom counts are obtained by integrating the intensity over atoms.

To illustrate our approach we consider the problem of estimating the number of unemployed at Local Labour Market (LLM) area level by using two misaligned data sources: auxiliary information available on different administrative partitions; reliable estimates of unemployed for Labour Force Survey (LFS) planned domains. A simulation study examines the capability of the proposed models to improve on traditional SAE model estimates.

2 Regression-only atom models for SAE

Before proposing a general methodology to address small area estimation problems, we consider a more familiar setting. Namely, we start with a traditional SAE model and suitably extend it in order to perform a regression analysis also on covariates available over areal partitions misaligned with the small areas set.

Suppose that $i = 1, \dots, I$ index the small areas (target zones) on which we want to estimate the characteristic θ (a total or a mean) of interest. A crude but still challenging HB model takes the form

$$d_i | \theta_i \sim N(\theta_i, \sigma_i) \tag{1}$$

$$\theta_i | \mu_i \sim \text{Po}(S_i \mu_i) \quad \log(\mu_i) = \mathbf{x}_i' \beta + v_i \tag{2}$$

$$v_i | \tau \sim N(0, \tau) \tag{3}$$

where d_i and σ_i are a direct estimate of θ_i and, respectively, a design variance estimate (of d_i). S_i is a regional synthetic estimate (of θ_i) which takes the place of an expected count as in standard Poisson mean formulations.

Instead, let $j = 1, \dots, J$ index the areas (source zones) on which a possibly important covariate $\mathcal{X}$ is available, and, finally, $k = 1, \dots, K$ index the intersection areas (atoms) arising from crossing each target zone with each source zone. Attractively, for us, there are no “edge” atoms; the target and source zones partition the same region. Accommodating the general case is straightforward (see [1] for modeling edge atoms).

Then, some sort of realignment is needed to carry out a regression analysis using the X ’s to produce better small area estimates θ_i . A number of techniques exist to obtain estimates for misaligned data at the required area level (usually called the *modifiable areal unit problem* in statistical literature, or *areal interpolation* in geography). None of these methods allows for a fully inferential approach to the problem of areal interpolation. We follow instead a fully inferential philosophy and adopt a HB approach allowing for a full uncertainty integration and propagation through the model. In particular we adopt the so-called atom-based models borrowed from the most recent HB literature on spatial misalignment ([1], [4]). The strategy behind atom-based models is to realign the misaligned data onto a common “intersection-partition”, whose areal unit is the “atom”, then conveniently set the regression model at the atom level, and, finally, build the estimates of interest by suitable aggregation over atoms.

In particular, with reference to our motivating example, we address the case in which both the small area characteristic of interest θ and the possibly important covariate $\mathcal{X}$ are count data (e.g. number of unemployed and of unemployment list enrolled respectively). Thus, the skeleton of an atom-based model as we have framed it in the SAE context, is as follows. At atom level, small area quantity and nonnested auxiliary variable are modeled as

$$\begin{aligned} \theta_k | \eta_k &\sim \text{Po}(S_k e^{\eta_k}) & X_k | \omega_k &\sim \text{Po}(N_k e^{\omega_k}) & (4) \\ \eta_k &= \beta_0 + \beta_1 X_k + \beta_2 z_k + \gamma_{i(\ni k)} & \omega_k &= \zeta_0 + \zeta_1 w_k + \psi_{j(\ni k)} & (5) \end{aligned}$$

where θ_k , X_k and (z_k, w_k) are, respectively, the latent (or unknown) total of interest, the latent auxiliary variable and further known covariates on atom- k ; S_k and N_k are known synthetic estimate and, respectively, population count for atom- k ; e^{η_k} and e^{ω_k} represent the incidence rate respectively associated with θ and X on atom- k ; γ_i and ψ_j are random effects inherited—by atom- k —from small area- i and auxiliary source area- j , respectively (which atom- k belongs to, indicated in (5) by “ $\ni k$ ” notation); β ’s and ζ ’s are unknown regression coefficients.

Then, on regional partitions, we assume $d_i | \theta_i \sim \text{Norm}(\theta_i, \sigma_i)$, as in (1), for direct survey estimates d_i ’s for small areas, with

$$\theta_i = \sum \theta_{k:k \in i} \quad X_j = \sum X_{k:k \in j} \quad (6)$$

which amount to imposing the sum-constraints (6) on each atom set making up either a small area or an auxiliary source zone.

Our methodology assumes Poisson counts at the atom level for both the unknown of interest and the latent auxiliary variable. In other cases, these variable may have other distributions, e.g., Bernoulli trials with probit or logit links. Also, in principle, the linking model for θ might be set at small area level. In this case, atom values X_k should be aggregated so as to reconstruct small area value X_i (i.e. $X_i = \sum X_{k:k \in i}$) which is to be imputed into the regression equation set for θ_i . However, a regression analysis is preferably carried out at atom or finer level whenever further covariates are available on regional partitions finer than the small area set. In (5), z_k as well as w_k denote known values of auxiliary variables whose source is nested within atoms. Furtherly, linking models (5) are defined on the log-linear scale, although other functional forms, such as the identity-linear or the mixed types, can be considered (see e.g. [2]). And, in the regression analysis for θ or X , random effects, generically denoted by γ_i and ψ_j in (5), can be set either as spatially-structured (e.g. by specifying a Gaussian conditionally autoregressive (CAR) prior) or simply as heterogeneous effects (by customarily assuming an exchangeable Gaussian prior), or perhaps a sum of such effects.

3 A point level version

An alternative to the class of misaligned data models proposed in Section 2 is based upon Gaussian process specifications. This approach envisions an underlying intensity surface which is block-averaged to the areal units relevant to the study. Such intensities are commonly used for explaining spatial point patterns. For us, the intensity surface is associated with the unobservable spatial point pattern of cases of interest (e.g., we can think of “cases” as the unknown point locations for the unemployed individuals). Here the surface is assumed to be a realization of a logGaussian process model.

Suppose we revise the model for the expected incidences in (4) with (5) using integration of latent intensity surfaces over atoms, that is as

$$\int_{A_k} S(s) e^{\eta(s)} ds \quad \int_{A_k} N(s) e^{\omega(s)} ds, \quad (7)$$

for process- θ and X respectively, where A_k denotes the atom- k (with area $|A_k|$). Atoms are assumed small enough to approximate $S(s)$ and $N(s)$ by a constant function over atom, $S(s) \approx S_k/|A_k|$, $N(s) \approx N_k/|A_k|$. Moreover, $\eta(s) = \beta_0 + \beta_1 X_k + \beta_2 z_k + \gamma(s)$ and $\omega(s) = \zeta_0 + \zeta_1 w_k + \psi(s)$ where, differently from (5), random effects $\gamma(s)$ and $\psi(s)$ are given Gaussian process priors. Hence, the expected incidences (7) become

$$S_k e^{\beta_0 + \beta_1 X_k + \beta_2 z_k} \int_{A_k} e^{\gamma(s)} / |A_k| \quad N_k e^{\zeta_0 + \zeta_1 w_k} \int_{A_k} e^{\psi(s)} / |A_k| \quad (8)$$

respectively. Note the difference between these expressions and those in (4) with (5). In the latter $\gamma_{i(\cdot \ni k)} = \int_{A_k} \gamma(s) ds / |A_k|$ and $\psi_{j(\cdot \ni k)} = \int_{A_k} \psi(s) ds / |A_k|$. But $\exp(\gamma_{i(\cdot \ni k)}) \neq \int_{A_k} \exp(\gamma(s)) ds / |A_k|$ and if $\gamma(s)$ is a Gaussian process, the resulting aggregated random effect is not normally distributed, similarly for $\psi_{j(\cdot \ni k)}$. Under this construction, we see that (4) with (5) induces an ecological fallacy (see [9] for reference). So, working at the point level achieves the highest spatial resolution, introduces a covariance function rather than proximities, and avoids an ecological fallacy. But, we see that the desired relationships with X_k and w_k , respectively, are unaffected in this regard and so, we can fit the simpler areal level model without much concern.

4 SAE via S $\odot$ LAE/S $\odot$ LAE atom models

We proceed by addressing the problem of the additional integration of large area estimates i.e. direct estimates evaluated on survey planned domains. It is reasonable that small area estimates will further improve given that large estimates are generally reliable. We will conveniently refer to the final model as *Small $\odot$ Large* area estimation model, S $\odot$ LAE for short, for the discrete-space version, as S $\odot$ LAE for the continuous-space approach of Section 3.

To this aim we generalize the spatial structure of the foregoing section by adding a further layer, i.e. large areas $p = 1, \dots, P$. Then, survey information related to large areas is conveyed by an additional stage, $d_p | \theta_p \sim \text{Normal}(\theta_p, \sigma_p)$, that is by assuming an appropriate sampling model with the large area estimates subject to the sum-constraint $\theta_p = \sum \theta_{k:k \in p}$, as θ_i 's are in (6). d_p and σ_p have the customary interpretation of direct estimates and, respectively, sampling error variances for survey planned domains (e.g. for provinces in our application).

We point out that inference coherency synthesized by the so called benchmarking propriety (small area estimates sum up to large area estimates) is automatically satisfied (since $\sum_k \theta_k = \sum_i \sum \theta_{k:k \in i} = \sum_i \theta_i = \sum_p \sum \theta_{k:k \in p} = \sum_p \theta_p$) in the final model version of the present section.

5 A simulation study

In this section we present only a subset of results from a larger plan of simulation studies. However, in all experiments, survey data simulation was carried out on the basis of the Italian LFS. In particular, we generate a number of LFS samples for the sole Veneto Region using 2001 census data on population counts available at municipality level, and following a two-stage stratified sampling scheme with varying probability at stage-1 and simple random sampling without replacement at stage-2. Then, in the specific scenario here presented, $\mathcal{X}$ (unemployment list enrolled) and θ (unemployed) numbers are model-based simulated at municipality level. In particular, they were generated according to $X_l \sim \text{Po}(N_l \omega_l)$ with $\omega_l = \zeta_0(1 + \zeta_1 w_l) e^{\phi_j}$

and $Y_l \sim \text{Po}(S_l \eta_l)$ with $\eta_l = \beta_0(1 + \beta_1 x_l + \beta_2 z_l) e^{v_l}$ models, respectively, where: l is the municipality index; $\zeta_0, \zeta_1, \beta_0$ and β_1 were set to the estimates resulting from fitting an analogous regression model on real data; w_l and z_l are covariates available at municipality level (numbers of employed in industry and number of registered immigrants respectively); moreover, ϕ_j and v_i have been specified as heterogeneous effects. More information on the real context where our simulation was set can be found in [6].

Notice the the mixing (additive in covariates, multiplicative in random effects) link, instead of the customary log-link, was chosen for this particular simulation. The reason why we preferred such a link was its property of maintaining the same form when aggregated on higher levels. Thus, it makes the comparison between traditional (1)-(3) and atom-based models “fair” (otherwise possible limits of a traditional model might depend on how data have been generated, yet might not on its intrinsic capability). Besides, such a choice allow us avoiding ecological bias issues which would otherwise arise also for model equations (5) and (8) so far specified.

To compare the performance of $\hat{\theta}_i^{HB}$ estimators, we compute a series of standard measures out of $R = 100$ simulated samples: relative bias (*rb*) and absolute relative bias (*arb*), absolute relative error (*are*) and relative root mean squared error (*re*), efficiency (*eff*) and reliability (*rel*) of the standardly used efficiency indicator. Details for their computation can be found in [7].

The results shown here (Table 1) refer to a simulation scenario where both the misaligned variable $\mathcal{X}$ and the nested covariates at finer area level (z and w) are important auxiliary sources of information. Then, a SAE traditional model, like (1)-(3), might be unfit for it neglects the auxiliary information $\mathcal{X}$ or, at best, merely uses it by a “plug-in” method. Moreover, the unit of analysis of a traditional model is set to a level not finer than that defined by small areas. Last but not least, it fails to integrate direct large area information.

Best performance values are boxed though if misleading they are grey-filled. The other values have been grey-coloured with intensity growing with worsening performance (by dividing, for each measure, the range of expectation/median values into five equal intervals).

Main outcomes of the outlined simulation experiment are concisely listed below, with possible explanations according to Authors’ opinion. (See the legend in Table 1.)

- (As expected) The T model without / with $\mathcal{X}$ plugged-in (1) gets the worst performance scores.
- The T model with $\mathcal{X}$ plugged-in (2) gets far better scores (see bias), since the areal unit of analysis is set to a finer level than small area, yet it performs worse than RA model (see either accuracy or reliability), for not allowing all uncertainty integration.
- Both the S $\odot$ LAE and S $\odot$ LAE models have the best performance, also better than the RA version as they integrate large direct estimates as well. Moreover, they show equivalent results here being information available on levels not finer than atoms.

Table 1 Simulation study

estimator	BIAS				ACCURACY				EFFICIENCY		RELIABILITY	
	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{re}$	re^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
T without $\mathcal{X}$	4	2	26	22	30	28	33	32	13	13	57	50
<i>non-na</i>	2	2	23	21	26	25	31	30	15	14	60	59
T with $\mathcal{X}$ plug-in(1)	3	2	20	18	23	21	28	28	13	13	61	60
<i>non-na</i>	2	2	18	16	20	19	27	28	14	14	66	64
T with $\mathcal{X}$ plug-in(2)	1	1	11	9	17	16	23	20	14	14	80	75
<i>non-na</i>	1	0	10	8	16	15	22	21	15	15	89	91
RA	0	1	9	7	14	13	17	16	15	15	96	96
<i>non-na</i>	0	0	8	6	13	11	16	15	17	16	104	108
S $\odot$ LAE	0	0	5	3	12	11	15	14	16	15	102	103
<i>non-na</i>	0	0	3	3	11	10	14	13	18	16	110	109
S $\odot$ LAE	-1	0	5	2	11	11	15	14	16	15	101	104
<i>non-na</i>	0	0	3	3	10	11	15	13	17	16	106	108

Overlined measures are mean values, whereas M superscript denotes median values.

T is traditional model (1)-(3)

- without $\mathcal{X}$: the misaligned covariate X is not a regreesor;

- with $\mathcal{X}$ plug-in(1): with X plugged-in by “population weighting” interpolation;

- with $\mathcal{X}$ plug-in(2): with X plugged-in by using aggregated atom-based model estimates.

RA is the regression-only atom-based model (Section 2)

S $\odot$ LAE is the discrete-space version of the atom-based model (Sections 2 and 4)

S $\odot$ LAE is the continuous-space version of the atom-based model (Sections 3 and 4)

non-na rows refer to measures calculated letting out survey non-sampled areas.

6 Concluding remarks

We conclude by pointing out some relevant innovative aspects of the proposed class of misaligned data models (in both versions) with respect to a traditional model for SAE. The proposed HB framework builds small area estimates (θ_i) through aggregation (over θ_k 's), hence in a way similar to that by which survey estimates are generally obtained. However, it integrates any source of potentially important information within a full probability model, thus formally combining the multiple sources of variation within a global analysis. Sources of information are those traditionally incorporated into a SAE model, such as small area direct estimates (d_i), auxiliary variables (z_k), (possibly) synthetic estimates (S_k). Furthermore, the atom-based model here proposed allows for misaligned supplementary variables (X_k) and large area direct estimates (d_p) being incorporated as well, properly accounting for the uncertainty related to this integration. Finally, traditional model-based estimators do not generally benchmark to reliable direct survey estimates for large areas (however, the construction of benchmarked estimators is currently an active thread of SAE research, see e.g. [8], [3] and [5]). On the other hand benchmarking is automatically satisfied in the proposed atom-based model.

References

1. Banerjee, S., Carlin, B. P., Gelfand, A. E.: Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall (2004)
2. Best, N. G., Ickstadt, K., Wolpert, R. L.: Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions. *J. Amer. Statist. Assoc.* **95**, 1076–1088 (2000)
3. Lu, L., Larsen, M. D.: Small area estimation in a survey of high school students in Iowa. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2627–2634 (2007)
4. Mugglin, A. S., Carlin, B. P., Gelfand, A. E.: Fully Model-Based Approaches for Spatially Misaligned Data. *J. Amer. Statist. Assoc.* **95**, 877–887 (2000)
5. Pfeiffermann, D., Tiller, R.: Small-area estimation with state-space models subject to benchmark constraints. *J. Amer. Statist. Assoc.* **101**, 1387–1397 (2006)
6. Torelli, N., Trevisani, M.: Labour force estimates for small geographical domains in Italy: problems, data and models. *International Review of Social Sciences* **4**, 443–464 (2008)
7. Trevisani, M., Torelli, N.: Hierarchical bayesian models for small area estimation with count data. *DiSES Working Papers* **115**, University of Trieste (2007)
8. Ugarte, M. D., Militino, T. G., Goicoa, T.: Benchmarked estimates in small areas using linear mixed models with restrictions. *Test*, 1–23, *online first* (2008)
9. Wakefield, J. C.: Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183 (2007)