A fully continuous spatial model for multi-level survey-based population inference and aggregation

John Paige
Department of Statistics, University of Washington,
Geir-Arne Fuglstad
Department of Mathematical Sciences, NTNU,
Andrea Riebler
Department of Mathematical Sciences, NTNU,
and Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

June 2, 2020

Abstract

The production of fine-scale, pixel level maps using demographic data from complex, multi-level household survey data have become increasingly prevalent in the current era of precision public health. For these types of datasets, no continuous spatial model has been developed to simultaneously account for the many sources of uncertainty important at such scales including enumeration area (EA) locations, cluster effects, spatial variation, binomial variation, and pixel and EA level population denominators. Instead, all continuous spatial models applied in this context routinely ignore these sources of uncertainty except for spatial variation and cluster effects. In addition, no study has explored how all these sources of uncertainty affect predictive uncertainty.

In this study, we propose a model and method of inference that accounts for these sources of uncertainty in a computationally efficient way. Our model can be fit and applied post-hoc after fitting a spatial model that does not directly account for these sources of uncertainty, and can therefore be used in conjunction with a wide variety of continuous spatial models, agreeing with the central predictions of the original model, and differing only in the uncertainty. In a simulation study, we find that our proposed model produces 80% credible interval (CI) widths at the pixel level on average roughly 3 times larger than those of a SPDE-based model as a result of accounting for the additional sources of uncertainty, and 80% CIs on average 10% larger at the county level. Finally, we use our model to predict neonatal mortality rates in Kenya from 2010-2014 with data from the 2014 Kenya demographic health survey, validating it at the pixel level. Achieved 80% CI coverage is still too low at only 70%, but much higher than the 30% pixel level coverages of the SPDE model. We believe cluster location jittering, and errors in estimates of urbanicity and population density are contributing to the proposed model's undercoverage.

1 Introduction

- Problems with pixel level mapping
- Problems with spatial analysis of survey data
- Marked point process models in ecology

2 Problem Setup

We assume the following model for neonatal mortality, indexed at the area (usually county), i, and the cluster, c:

$$\begin{aligned} Z_{ic}|\mu_{ic} &\sim \text{Binomial}(N_{ic}, \mu_{ic}) && (1) \\ \mu_{ic} &= \text{expit}(u(s_{ic}) + \epsilon_{ic}), && (2) \end{aligned}$$

where the counts are the number of neonatals that died, and $\epsilon_{ic} \sim N(0, \sigma_\epsilon^2)$ is a cluster level random effect that can allow for dependency between observations in the same cluster. Hence, within the area we have $|C_i|$ distinct prevalences, where C_i is the set of cluster indices in area i, and $|\cdot|$ denotes the set cardinality. The count Z_{ic} is the number of neonatals in EA c and area i, where s_{ic} is the spatial location of EA c in the area.

The main targets of inference are the proportion of neonatals that died at different aggregation levels. To begin, we can start by considering inference at the area level, where the number of EAs is known, and a partition of area i into a fine pixelated grid indexed by g, where the number of EAs in each pixel is not known, but the stratum associated with each grid cell (usually urban/rural classification) is known. The targets of inference in area i and subarea g will be:

$$p_i = \frac{1}{N_i} \sum_{c \in C_i} Z_{ic} = \sum_{c \in C_i} \frac{N_{ic}}{N_i} \times \frac{Z_{ic}}{N_{ic}} \qquad (3)$$

$$p_{ig} = \frac{1}{N_{ig}} \sum_{c \in C^g} Z_{ic} = \sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}}, \qquad (4)$$

where $N_i \equiv \sum_{c=1}^{C_i} N_{ic}$ and $N_g \equiv \sum_{c \in C^g} N_{ic}$ are the number of neonatals in area i and pixel g respectively. Note that the set C^g gives the indices of the EAs that are in subarea g, and that p_i and p_{ig} are empirical proportions rather than latent probabilities.

Our main goal will be to explore how best to aggregate from the cluster level to the pixel and area levels in a way that balances quality of the predictive distribution at multiple aggregation levels, and computational feasibility. An important side goal of this project will be to determine the most important sources of uncertainty when conducting inference at both area and pixel (most likely 5km resolution) levels. In the ideal scenario, we would validate this model at the pixel level, and show that the data is well characterized by the predictive distribution, unlike in existing models for these contexts.

3 Spatial Aggregation Framework and Models

In this analysis, we will assume a framework for accounting for different sources of variation when producing areal level results from continuously indexed spatial models. We consider five variables that respectively lead to different sources of variation: $\boldsymbol{u}_{i:}, \boldsymbol{\epsilon}_{i:}, \boldsymbol{N}_{i:}$, $\boldsymbol{Z}_{i:}$, and $\boldsymbol{s}_{i:}$, where ':' denotes varying over all indices to form a vector, and where $u_{ic} \equiv u(s_{ic})$. In order to determine which sources of variation are most important for central predictions and credible intervals at different spatial aggregation levels, one could imagine approximating the targets of inference, p_i and p_{ig}, with a series of simplifications, each successive approximation accounting for fewer of the five considered sources of variation as follows:

Models:

LCPB	**LCPb**	**LCpb**	**Lcpb**	**lcpb**
p_i	$\approx E_{\boldsymbol{Z}_{i:}}\left[p_{ic}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_i\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:},\boldsymbol{s}_{i:}}\left[p_i\right]$
p_{ig}	$\approx E_{\boldsymbol{Z}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:},\boldsymbol{s}_{i:}}\left[p_{ig}\right]$

In the LCpb model, for instance, rather than using draws from the posterior of p_i, we use draws from the posterior of $E_{\boldsymbol{Z}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right]$ to make central predictions and credible intervals for area i. The same is true for p_{ig}, and with equivalent expectations used as approximations for the other models. The symbols l/L, c/C, p/P, and b/B denote whether variation in cluster effects ($\boldsymbol{\epsilon}_{i:}$), population denominators ($\boldsymbol{N}_{i:}$), binomial variation ($\boldsymbol{Z}_{i:}$), and EA location uncertainty ($\boldsymbol{s}_{i:}$) is integrated out or not, where by 'integrating out' we mean taking expectation over as described above. Lowercase symbols signify integrating out the corresponding sources of variation when making predictions for area i and subarea g, as opposed to sampling over that variation instead, which is denoted by uppercase symbols.

Since we never integrate out $\boldsymbol{u}_{i:}$, there are 16 different orders in which we can integrate out sources of variation in the other four terms. However, we chose this hierarchy from a distributional dependence and computational feasibility standpoint. We integrate out $\boldsymbol{s}_{i:}$ last since all variables except $\boldsymbol{u}_{i:}$ are dependent on the existence and number of EAs in any pixel and area. We integrate out $\boldsymbol{Z}_{i:}$ last, because it is dependent on all the other variables.

It may seem as though integrating out four of the five considered sources of uncertainty reduces the level of uncertainty by too large an amount. However, this is similar to what is commonly done in practice, and accounting for additional sources of variation can be computationally difficult. Moreover, when producing estimates for large enough areas, we have found that some of these sources of variation, such as $\boldsymbol{Z}_{i:}$ and $\boldsymbol{\epsilon}_{i:}$, tend to average out. However, the required size an area must be for each source of variation to become insignificant is not known. Assessing the scenarios in which each source of variation becomes significant, their relative importance, and how to integrate them out when necessary in a computationally feasible way is therefore very important.

LCPB: In this case, we attempt to sample over all considered forms of variation. In this case, we must decide on a model for the EA locations, $\boldsymbol{s}_{i:}$. For simplicity, it would be simplest to assume that EA locations are independent of each other. Relaxing this assumption would require a much more complex model that we do not consider here. In addition to assuming independence, it would be justifiable to assume that the probability of an EA being at a certain location within area i is proportional to that location's population density. Under these two assumptions, the EA locations in area i *must* follow a Poisson process with intensity proportional to the continuously indexed population density surface, say $q(s)$ for spatial location s, conditioned on the information that there are $|C_i|$ EAs in total in area i. This is also known as a binomial process. Note that we are assuming for simplicity that $|C_i|$ is much greater than the number of sampled clusters, so that we can ignore the sampled cluster locations.

We will assume that $\boldsymbol{N}_{i:}$ is independent of $\mathbf{s}_{i:}$, instead entirely dependent on a separate census dataset, say $\mathcal{W}$, used only to determine the distribution of $\boldsymbol{N}_{i:}$.

The number of EAs in g, $|C^g|$, is binomial with probability equal to $\frac{q(s_{ig})}{\int_{A_i} q(s)\ ds}$, and has $|C_i|$ trials. Here, A_i is the spatial domain of area i. $|C^g|$ is also roughly Poisson with rate $|C_i|\frac{q(s_{ig})}{\int_{A_i} q(s)\ ds}$ for sufficiently small subareas.

To draw a sample from the posterior of p_i, we can draw a sample from the joint distribution of $\boldsymbol{N}_{ig}$ and p_{ig} over all grid cells $g \in A_i$, and then aggregate with the formula, $p_i = \sum_{g\in A_i} \frac{N_{ig}}{N_i} \times p_{ig}$. To accomplish this, we can use the following algorithm given the dataset, $\mathcal{Y}$:

Algorithm 1 Draw $p_i^{(j)}$, $p_{ig}^{(j)}$ from posterior $p_i, p_{ig}|\mathcal{Y}$

```
$\mathbf{s}_{i:}^{(j)} \leftarrow \mathbf{s}_{i:}|q(\cdot)$
$\boldsymbol{u}_{i:}^{(j)} \leftarrow \boldsymbol{u}_{i:}|\mathbf{s}_{i:}, \mathcal{Y}$
$\boldsymbol{\epsilon}_{i:}^{(j)} \leftarrow \boldsymbol{\epsilon}_{i:}|\mathcal{Y}$
$\mathbf{N}_{i:}^{(j)} \leftarrow \mathbf{N}_{i:}|\mathcal{W}$
$\mathbf{Z}_{i:}^{(j)} \leftarrow \mathbf{Z}_{i:}|\mathbf{N}_{i:}, \boldsymbol{\mu}_{i:}$
$N_i^{(j)} \leftarrow \mathbf{1}^T\mathbf{N}_{i:}^{(j)}$
for all $g \in A_i$ do
    $N_{ig}^{(j)} \leftarrow \sum_{c\in C^g} N_{ic}^{(j)}$
    $p_{ig}^{(j)} \leftarrow \sum_{c\in C^g} \frac{N_{ic}^{(j)}}{N_{ig}^{(j)}} \times \frac{Z_{ic}^{(j)}}{N_{ic}^{(j)}}$
end for
$p_i^{(j)} \leftarrow \sum_{g\in A_i} \frac{N_{ig}^{(j)}}{N_i^{(j)}} \times p_{ig}^{(j)}$
```

For sufficiently fine g, u will not change significantly over A^g, the spatial domain of subarea g. In that case, we can simplify the process by considering only the values of u at the centroid of each subarea g, and conditioning on those values when drawing $\boldsymbol{Z}_{i:}$. This would provide a considerable computational advantage if, for instance, u is represented as a linear combination of basis functions, since the basis matrix would not need to be recomputed for each draw of $\boldsymbol{s}_{i:}$:

Algorithm 2 Draw $p_i^{(j)}$, $p_{ig}^{(j)}$ from posterior $p_i, p_{ig}|\mathcal{Y}$

```
$\mathbf{s}_{i:}^{(j)} \leftarrow \mathbf{s}_{i:}|q(\cdot)$
$\boldsymbol{\epsilon}_{i:}^{(j)} \leftarrow \boldsymbol{\epsilon}_{i:}|\mathcal{Y}$
$\mathbf{N}_{i:}^{(j)} \leftarrow \mathbf{N}_{i:}|\mathcal{W}$
for all $g \in A_i$ do
    $u_{ig}^{(j)} \leftarrow u_{i:}|s_{ig}, \mathcal{Y}$
    $N_{ig}^{(j)} \leftarrow \sum_{c \in C^g} N_{ic}^{(j)}$
    $Z_{ig}^{(j)} \leftarrow Z_{ig}|\mathbf{N}_{i:}, \boldsymbol{\mu}_{i:}$
    $p_{ig}^{(j)} \leftarrow \sum_{c \in C^g} \frac{N_{ic}^{(j)}}{N_{ig}^{(j)}} \times \frac{Z_{ic}^{(j)}}{N_{ic}^{(j)}}$
end for
$N_i^{(j)} \leftarrow \mathbf{1}^T \mathbf{N}_{i:}^{(j)}$
$p_i^{(j)} \leftarrow \sum_{g \in A_i} \frac{N_{ig}^{(j)}}{N_i^{(j)}} \times p_{ig}^{(j)}$
```

A disadvantage of this approach, aside from being more computationally intensive than the others, is that subareas with very small population densities will likely get very few EAs drawn in them on average. Hence, many posterior draws would be required in order to get a finer estimate of the posterior distribution. If any such subareas existed, then the posterior $|C^g||\mathcal{Y}$ can be well-approximated by placing probability mass only on a small number of possible values of $|C^g|$, say K values. Call this approximation $|\tilde{C}^g|$, with probability mass function $P(|\tilde{C}^g| = n) = m_n^g$ for $n = 0, ..., K_g$. Then one could condition on $|\tilde{C}^g|$ being equal to $0, \ldots, K_g$, averaging over the draws with weights m_n^g. However, we do not need to get into the mathematical details of this until we actually run into this problem.

4 Plan of Action

First, I will simulate a fixed population and set of surveys for testing purposes with parameters similar to the Kenya secondary education ELK-T model fits as follows:

- Simulate the EAs (locations, households, neonational populations) as in the survey paper
- Simulate NMRs and deaths at EAs using the SPDE model fitted to the NMRs
- Draw surveys from the EAs as in the survey paper. Generate 100 stratified and unstratified, but we will only use 1 stratified survey to start with, and the unstratified might not be necessary.
- Evaluate differences between the chosen models at multiple aggregation levels

Could try this again, but using the ELK-T model to simulate NMRs.

When we discussed this project in person, we considered proposing a variation of the Cpbl model (sampling over cluster level overdispersion) as an alternative to the cpbl model (the same, but integrating out cluster level overdispersion). I mainly bring up the other model in case the difference between the cpbl and Cpbl models is not large. Here are a few reasons I can think of why that might be the case:

1. There are many EAs in a given area, and averaging over all of them might not make a huge difference
2. We might be comparing the models at the wrong aggregation level. There are fewer EAs the smaller the areal level, and by the time we get to pixels, averaging over variation in EAs will make much more of a difference
3. Even at small areal aggregation level, binomial variation might make more of a difference than cluster (EA) level overdispersion. In that case, accounting for EA level overdispersion would be great, but not necessarily an exciting result.

4. There are a lot of different approximations possible to use for these models, which could get confusing quickly. Testing them will be important as long as focus is kept on the larger goals

Since the BCPL model would be very easy to code up at the pixel level (other levels would be harder, but doable using Algorithm 1 given above, I was hoping to include it in the exploratory phase to see how much variation at each aggregation level is due to each of these effects.

As alluded to above, the performance of these methods will likely vary considerably depending on the aggregation level, so it would be prudent for us to, at least in the exploratory phase, evaluate performance at the region, county, and 5km pixel levels. For simplicity, we could start by including BCL only at the pixel level, and including other models at all levels, for a single survey.

5 Appendix

5.1 Drawing from $\boldsymbol{N}_{i:}|\mathcal{W}$