Outline for Exploratory Analysis of: '*Accounting for Cluster Level Effects in Population Aggregates with Geospatial Models*'

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

April 7, 2020

1 Problem Setup

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

$$Z_{ic}|\mu_{ic} \sim \text{Binomial}(N_{ic}, \mu_{ic}) \quad (1)$$

$$\mu_{ic} = \text{expit}(u(s_{ic}) + \epsilon_{ic}), \quad (2)$$

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 that died. We will think of the spatial term $u(\cdot)$ as being a function that is continuously indexed in space, and estimated from the data, 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 be considering inference at the area level, where the number of EAs is known, and a partition of area i into subareas indexed by g, where the number of EAs is not known exactly. In particular, g could be used to denote the grid point index for a fine spatial grid, but it would also be possible to consider a different partitioning of area i where the exact number of EAs in each subarea is not known. The targets of inference in area i and subarea g will be:

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

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

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 subarea g respectively, and $q_{ic} \equiv \frac{N_{ic}}{N_i}$ and $q_{ic}^g \equiv \frac{N_{ic}}{N_{ig}}$ are the weight of cluster c in the population-weighted average for p_i and p_{ig} respectively. The set C^g gives the indices of the EAs that are in subarea g. Note that p_i, p_{ic}, and p_{ig} are empirical averages rather than parameters.

The main goal of this project will be to explore how best to aggregate from the cluster level to the subarea and area levels in a way that balances quality of the predictive distribution at multiple aggregation levels, and computational feasibility. In order to accomplish this, we will iteratively integrate out parts of (3) and (4), each time attempting to reduce the computational difficulty and simplify the model, but possibly introducing undercoverage in the process. An important side goal of this project will be to determine what are the most important sources of uncertainty when conducting inference at both area and subarea (most likely 5km pixel) levels.

2 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 the cluster index c varying over the values in C_i so that each of these five variables are vectors of length $|C_i|$, 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:		
CPBL	**CPBl**	**CPbl**	**Cpbl**	**cpbl**
p_i	$\approx E_{\boldsymbol{s}_{i:}}\left[p_{ic}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_i\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_i\right]$
p_{ig}	$\approx E_{\boldsymbol{s}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right]$	$\approx E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_{ig}\right]$

In the CPbl model, for instance, rather than using draws from the posterior of p_i, we use draws from the posterior of $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_i\right]$ 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 c/C, p/P, b/B, and l/L 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 model complexity and computational feasibility standpoint. We integrate out $\boldsymbol{s}_{i:}$ first since it requires the most complex model assumptions. By integrating it out, the result is more robust to assumptions on the joint distribution of $\boldsymbol{s}_{i:}$, instead relying only on its expectation in each area and subarea.

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.

CPBL: 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**
  $u_{ig}^{(j)} \leftarrow u_{i:}|s_{ig}, \mathcal{Y}$
  $N_{ig}^{(j)} \leftarrow \sum_{c \in C^g} N_{ic}^{(j)}$
  $Z_{ig}^{(j)} \leftarrow Z_{ig}|\mathbf{N}_{i:}, \mu_{ig}$
  $p_{ig}^{(j)} \leftarrow \sum_{c \in C^g} \frac{N_{ic}^{(j)}}{N_{ig}^{(j)}} \times \frac{Z_{ic}^{(j)}}{N_{ic}^{(j)}}$
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.

For the following models, we implicitly condition on the data when drawing samples:

CPBl: In this model, we include all forms of variation except for EA location uncertainty.

Recall (3) and (4):

$$p_i = \frac{1}{N_i} \sum_{c \in C_i} Z_{ic} = \sum_{c \in C_i} \frac{N_{ic}}{N_i} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C_i} q_{ic} \times p_{ic}$$
$$p_{ig} = \frac{1}{N_{ig}} \sum_{c \in C^g} Z_{ic} = \sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C^g} q^g_{ic} \times p_{ic}$$

Our goal is to get an expression for the expectation of these with respect to the EA locations, $\boldsymbol{s}_{i:}$. In order to do so, we must consider the following expectation:

$$E_{\boldsymbol{s}_{i:}}\left[p_{ic}\right] = E_{\boldsymbol{s}_{i:}}\left[\frac{Z_{ic}}{N_{ic}}\right] \tag{5}$$
$$= \frac{1}{N_{ic}} E_{\boldsymbol{s}_{i:}}\left[Z_{ic}\right] \qquad (N_{ic} \perp\!\!\!\perp \boldsymbol{s}_{i:}). \tag{6}$$

At the pixel level, we must evaluate:

$$E_{\boldsymbol{s}_{i:}}\left[\sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}}\right] \approx \sum_{n \in \{0,\ldots,K_g\}} m^g_n \cdot \sum_{c \in C^g,\ |C^g|=n} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}}. \tag{7}$$

Although it is possible to use (7) to draw from $E_{\boldsymbol{s}_{i:}}\left[p_{ig}\right]$ for each p_{ig} individually for subarea level predictions, getting area level predictions by drawing from the joint distribution with any reasonable resolution over all subareas could be unwieldy. It is easier to start from (6) directly. In order to draw from $E_{\boldsymbol{s}_{i:}}\left[p_{ic}\right]$ one can first draw from $E_{\boldsymbol{s}_{i:}}[Z_{ic}]$ by drawing ϵ_{ic} and N_{ic}, and then from the joint distribution $(\mu_{ic}(s_{ic}), s_{ic}) \mid \epsilon_{ic}$. Plugging into (6) yields the resulting posterior sample of $E_{\boldsymbol{s}_{i:}}\left[p_{ic}\right]$.

CPbl: In this model, we include variation in cluster effects ($\boldsymbol{\epsilon}_{i:}$) and also in the population denominator term ($\boldsymbol{N}_{i:}$), but integrate out binomial and EA location variation. Recall (3) and (4):

$$p_i = \frac{1}{N_i} \sum_{c \in C_i} Z_{ic} = \sum_{c \in C_i} \frac{N_{ic}}{N_i} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C_i} q_{ic} \times p_{ic}$$
$$p_{ig} = \frac{1}{N_{ig}} \sum_{c \in C^g} Z_{ic} = \sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C^g} q^g_{ic} \times p_{ic}.$$

Taking expectation over $\boldsymbol{Z}_{i:}$ and $\boldsymbol{s}_{i:}$, we get:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_i\right] = \sum_{c\in C_i}\frac{N_{ic}}{N_i}\times E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[\frac{Z_{ic}}{N_{ic}}\right]$$
$$= \sum_{c\in C_i}\frac{N_{ic}}{N_i}\times E_{\boldsymbol{s}_{i:}}\left[\mu_{ic}\right]$$

for the area level NMR, resulting in,

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_i\right] = \sum_{c\in C_i}\frac{N_{ic}}{N_i}\times\int_{A_i}\text{expit}(u(s)+\epsilon_{ic})\cdot\frac{q(s_{ig})}{\int_{A_i}q(s)\,ds}\,ds, \tag{8}$$

where this integral can be approximated on a numerical grid. Note that this is easier than the integral for the CPBl model, since in this case we are integrating a smooth function over space, whereas in the CPBl model we would need to simulate binomial variation as a function of $\mu(s_{ic})$ depending on the location of s_{ic}. Still, since it must be calculated for every EA and every posterior draw, it will require some techniques to simplify. For instance, if the above integral is calculated for a grid of possible values of ϵ_{ic}, one could make a spline approximation of the integral as a function of ϵ_{ic}, making it much easier to evaluate each of the terms in the summation (1 term for each EA). TODO: explore grid simplifications on grid, approximating the sum over C_i using CLT or other methods.

For the subarea level NMR we get:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_{ig}\right] = E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[\sum_{c\in C^g}\frac{N_{ic}}{N_{ig}}\times\frac{Z_{ic}}{N_{ic}}\right]$$
$$= E_{\boldsymbol{s}_{i:}}\left[\sum_{c\in C^g}\frac{N_{ic}}{N_{ig}}\times E_{\boldsymbol{Z}_{i:}}\left[\frac{Z_{ic}}{N_{ic}}\middle|\boldsymbol{s}_{i:}\right]\right]$$
$$= E_{\boldsymbol{s}_{i:}}\left[\sum_{c\in C^g}\frac{N_{ic}}{N_{ig}}\times\mu_{ic}\right] \qquad (\mu_{ic}\approx\text{constant for } c\in C^g).$$

To make the above summation feasible, we will need to approximate the distribution of $|C^g|$ with, say, $|\tilde{C}^g|$. If g is sufficiently small or has sufficiently small population density, we could use the approximation in the description of the CPBL model, with $P(|\tilde{C}^g| = n) = m_n^g$ for $n = 0, ..., K_g$ for some small K_g. For somewhat large g, however, this might become infeasible. For subareas g that are more likely to contain EAs, another option is to fix a

Monte Carlo approximation of the values of C^g, taking an equally weighted average over them. For sufficiently large subareas, C_i, the variation in the proportion of EAs per subarea g will decrease. It therefore might not be necessarily to take a very large number of Monte Carlo samples for a sufficient approximation. For any subarea g likely to receive a large number of EAs, binomial variation and EA location variation might matter less due to additional averaging. For such subareas it might even be sufficient to fix the number of EAs in g to their expectation. In this case, however, we will assume that g is small, and that the approximation given in the description of the CPBL model is sufficient. In that case, we have:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_{ig}\right] \approx \sum_{n\in\{0,\ldots,K_g\}} m_n^g \cdot \sum_{c\in C^g,\ |C^g|=n} \frac{N_{ic}}{N_{ig}} \times \mu_{ic}. \tag{9}$$

Since it might be possible for areas in the above approximation to accrue with aggregation, using (8) it is preferable for area level inference rather than aggregating posterior draws at the pixel level from (9) unless proven otherwise.

In order to generate posterior draws of these expectations at the area and pixel levels, (8) and (9) can be used after taking draws of $\boldsymbol{N}_{i:}|\mathcal{W}$ and $\boldsymbol{\epsilon}_{i:}$ as follows:

Algorithm 3 Draw $p_i^{(j)}$ from the posterior $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}[p_i]|\mathcal{Y}$

1: $\boldsymbol{u}_{i:}^{(j)} \leftarrow \boldsymbol{u}_{i:}|\mathbf{s}_{i:}, \mathcal{Y}$
2: $\boldsymbol{\epsilon}_{i:}^{(j)} \leftarrow \boldsymbol{\epsilon}_{i:}|\mathcal{Y}$
3: $\mathbf{N}_{i:}^{(j)} \leftarrow \mathbf{N}_{i:}|\mathcal{W}$
4: $N_i^{(j)} \leftarrow \mathbf{1}^T\mathbf{N}_{i:}^{(j)}$
5: **for all** $c \in C_i$ **do**
6: Use (8) to draw $p_{ic}^{(j)} \leftarrow E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}[p_{ic}]$
7: **end for**
8: $p_i^{(j)} \leftarrow \sum_{c\in C_i} \frac{N_{ic}^{(j)}}{N_i^{(j)}} \times p_{ic}^{(j)}$

Algorithm 4 Draw $p_{ig}^{(j)}$ from the posterior $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}[p_{ig}]|\mathcal{Y}$

1: $\boldsymbol{u}_{i:}^{(j)} \leftarrow \boldsymbol{u}_{i:}|\mathbf{s}_{i:}, \mathcal{Y}$
2: $\boldsymbol{\epsilon}_{i:}^{(j)} \leftarrow \boldsymbol{\epsilon}_{i:}|\mathcal{Y}$
3: $\mathbf{N}_{i:}^{(j)} \leftarrow \mathbf{N}_{i:}|\mathcal{W}$
4: $N_{ig}^{(j)} \leftarrow \sum_{c \in C^g} N_{ic}^{(j)}$
5: Use (9) to draw $p_{ig}^{(j)} \leftarrow E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}[p_{ig}]$

Cpbl: In this model, cluster variation is included ($\boldsymbol{\epsilon}_{i:}$), but we integrate out variation in the population denominator ($\boldsymbol{N}_{i:}$), binomial variation ($\boldsymbol{Z}_{i:}$), and EA location variation ($\boldsymbol{s}_{i:}$).

Recall (3) and (4):

$$p_i = \frac{1}{N_i}\sum_{c \in C_i} Z_{ic} = \sum_{c \in C_i} \frac{N_{ic}}{N_i} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C_i} q_{ic} \times p_{ic}$$
$$p_{ig} = \frac{1}{N_{ig}}\sum_{c \in C^g} Z_{ic} = \sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}} = \sum_{c \in C^g} q_{ic}^g \times p_{ic}.$$

Taking expectation over $\boldsymbol{Z}_{i:}$, $\boldsymbol{s}_{i:}$, and $\boldsymbol{N}_{i:}$, at the area level we get:

$$\begin{aligned}
E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right] &= E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[\sum_{c \in C_i} \frac{N_{ic}}{N_i} \times \frac{Z_{ic}}{N_{ic}}\right] \\
&= E_{\boldsymbol{N}_{i:}}\left[\sum_{c \in C_i} \frac{N_{ic}}{N_i} \times E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[\frac{Z_{ic}}{N_{ic}}\middle|\boldsymbol{N}_{i:}\right]\right] \\
&= E_{\boldsymbol{N}_{i:}}\left[\sum_{c \in C_i} \frac{N_{ic}}{N_i} \times E_{\boldsymbol{s}_{i:}}[\mu_{ic}]\right] \\
&= \sum_{c \in C_i} E_{\boldsymbol{N}_{i:}}\left[\frac{N_{ic}}{N_i}\right] \times \int_{A_i} \mu_{ic}(s_{ic}) \cdot \frac{q(s_{ic})}{\int_{A_i} q(s)\, ds}\, ds_{ic}.
\end{aligned}$$

At this point, it is important to note that we may know the particular stratum, say urban or rural crossed with area, of each EA. In that case, the distribution of $\boldsymbol{N}_{ic}|\mathcal{W}$ may

depend on whether EA c is urban or rural. We therefore split the above summation into a summation over rural EAs, and another over urban EAs. Letting C_i^{URB} and C_i^{RUR} be the number of urban and rural EAs an area i, we then get:

$$
\begin{aligned}
E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right] = &\sum_{c \in C_i^{\text{RUR}}} q^{\text{RUR}} \times \int_{A_i^{\text{RUR}}} \mu_{ic}(s_{ic}) \cdot \frac{q(s_{ic})}{\int_{A_i^{\text{RUR}}} q(s)\ ds}\ ds_{ic} \\
&+ \sum_{c \in C_i^{\text{URB}}} q^{\text{URB}} \times \int_{A_i^{\text{URB}}} \mu_{ic}(s_{ic}) \cdot \frac{q(s_{ic})}{\int_{A_i^{\text{URB}}} q(s)\ ds}\ ds_{ic}, \qquad (10)
\end{aligned}
$$

where q^{RUR} and q^{URB} are the values of the expectation $E_{\boldsymbol{N}_{i:}}\left[\frac{N_{ic}}{N_i}\right]$ in rural and urban EAs respectively.

Just as in the CPbl model, we can calculate the integrals (10) repeatedly by estimating them as a function of ϵ_{ic} using interpolating splines. TODO: look into the same CLT approximations as for the CPbl model.

At the subarea level, whatever stratum g is in, we assume all EAs in subarea g are also in that stratum so that the N_{ic} for any such EA are equal in expectation. We then have:

$$
\begin{aligned}
E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right] &= E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[\sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \frac{Z_{ic}}{N_{ic}}\right] \\
&= E_{\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[\sum_{c \in C^g} \frac{N_{ic}}{N_{ig}} \times \mu_{ic}\right] \\
&= E_{\boldsymbol{s}_{i:}}\left[\sum_{c \in C^g} \frac{1}{|C^g|} \times \mu_{ic}\right].
\end{aligned}
$$

To calculate this expectation, we will again need to approximate the distribution of $|C^g|$. Again, if g is small in area, then μ_{ic} will be constant over g, and we can approximate the distribution of $|C^g|$ by giving a small number of possible values probability mass. if g is large, then we could use a Monte Carlo sample for the approximation. For now, we will assume that g is small in area, noting that this will be the case for a fine pixelated grid over which estimates could be produced. In this case, we get:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right] \approx \sum_{n\in\{0,\ldots,K_g\}} m_n^g \cdot \sum_{c\in C^g,\ |C^g|=n} \frac{1}{n} \times \mu_{ic}. \tag{11}$$

We can take posterior draws from $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_i\right]$ and $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right]$ using similar algorithms as above

cpbl: In this model, we also integrate out the cluster effect, including only variation in the spatial function $u(\cdot)$.

Starting from where we left off above, we have:

$$\begin{aligned}
E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_i\right] &= E_{\boldsymbol{\epsilon}_{i:}}\Bigg[\sum_{c\in C_i^{\mathrm{RUR}}} q^{\mathrm{RUR}} \times \int_{A_i^{\mathrm{RUR}}} \mu_{ic}(s_{ic}) \cdot \frac{q(s_{ic})}{\int_{A_i^{\mathrm{RUR}}} q(s)\ ds}\ ds_{ic} \\
&\quad + \sum_{c\in C_i^{\mathrm{URB}}} q^{\mathrm{URB}} \times \int_{A_i^{\mathrm{URB}}} \mu_{ic}(s_{ic}) \cdot \frac{q(s_{ic})}{\int_{A_i^{\mathrm{URB}}} q(s)\ ds}\ ds_{ic}\Bigg] \\
&= \sum_{c\in C_i^{\mathrm{RUR}}} q^{\mathrm{RUR}} \times \int_{A_i^{\mathrm{RUR}}} E_{\boldsymbol{\epsilon}_{i:}}[\mu_{ic}(s_{ic})] \cdot \frac{q(s_{ic})}{\int_{A_i^{\mathrm{RUR}}} q(s)\ ds}\ ds_{ic} \\
&\quad + \sum_{c\in C_i^{\mathrm{URB}}} q^{\mathrm{URB}} \times \int_{A_i^{\mathrm{URB}}} E_{\boldsymbol{\epsilon}_{i:}}[\mu_{ic}(s_{ic})] \cdot \frac{q(s_{ic})}{\int_{A_i^{\mathrm{URB}}} q(s)\ ds}\ ds_{ic},
\end{aligned}$$

yielding:

$$\begin{aligned}
E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_i\right] &= |C_i^{\mathrm{RUR}}| \cdot q^{\mathrm{RUR}} \times \int_{A_i^{\mathrm{RUR}}} E_{\boldsymbol{\epsilon}_{i:}}[\mu_{i1}(s_{i1})] \cdot \frac{q(s_{i1})}{\int_{A_i^{\mathrm{RUR}}} q(s)\ ds}\ ds_{i1} \\
&\quad + |C_i^{\mathrm{URB}}| \cdot q^{\mathrm{URB}} \times \int_{A_i^{\mathrm{URB}}} E_{\boldsymbol{\epsilon}_{i:}}[\mu_{i2}(s_{i2})] \cdot \frac{q(s_{i2})}{\int_{A_i^{\mathrm{URB}}} q(s)\ ds}\ ds_{i2},
\end{aligned} \tag{12}$$

which is the formula we used in the survey paper (TO DO: simplify q^{URB} and q^{RUR}). Here, clusters 1 and 2 are used in $\mu_{i1}(s_{i1})$ and $\mu_{i2}(s_{i2})$ as arbitrary rural and urban clusters respectively. We could also use Jon's Logistic approximation to avoid needed to numerically integrate over ϵ_{ic}.

At the pixel level, we have:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_{ig}\right] = E_{\boldsymbol{s}_{i:},\boldsymbol{\epsilon}_{i:}}\left[\sum_{c\in C^g}\frac{1}{|C^g|}\times\mu_{ic}\right]$$
$$= E_{\boldsymbol{s}_{i:}}\left[\sum_{c\in C^g}\frac{1}{|C^g|}\times E_{\boldsymbol{\epsilon}_{i:}}\left[\mu_{ic}\mid s_{ic}\right]\right].$$

As above if g is large, we could use Monte Carlo samples to approximate the outer expectation, but we will assume that g is small. In that case, we can use the same approximation as previously to get:

$$E_{\boldsymbol{s}_{i:}}\left[\sum_{c\in C^g}\frac{1}{|C^g|}\times E_{\boldsymbol{\epsilon}_{i:}}\left[\mu_{ic}\mid s_{ic}\right]\right]\approx\sum_{n\in\{0,\ldots,K_g\}}m_n^g\cdot\sum_{c\in C^g,\ |C^g|=n}\frac{1}{n}\times E_{\boldsymbol{\epsilon}_{i:}}\left[\mu_{ic}\mid s_{ic}=s_{ig}\right].$$

Since $E_{\boldsymbol{\epsilon}_{i:}}\left[\mu_{ic}\mid s_{ic}=s_{ig}\right]$ is equal for each $c\in C^g$, we finally get:

$$E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:},\boldsymbol{\epsilon}_{i:}}\left[p_{ig}\right]\approx E_{\boldsymbol{\epsilon}_{i:}}\left[\mu_{ic}\mid s_{ic}=s_{ig}\right] \tag{13}$$

We can use the same method as for the area level to calculate this integral. A benefit of this model is that aggregations from the subarea to the area level are consistent. As written, this is not the case for the Cpbl and CPbl models, although perhaps there are different approximations that can be made for exact aggregation consistency. The reason for this lack of consistency is that, for there to be aggregation consistency, we would need a way to take joint draws over all subareas g from $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:}}\left[p_{ig}\right]$ and $E_{\boldsymbol{Z}_{i:},\boldsymbol{s}_{i:},\boldsymbol{N}_{i:}}\left[p_{ig}\right]$. But this would be possible using Monte Carlo draws from $\boldsymbol{s}_{i:}$ if those draws were shared for all g! Perhaps only a small number would be needed. Could always check variation over the Monte Carlo draws to see if we have enough. TODO: change the approximations in Cpbl and CPbl models to use Monte Carlo draws from $\boldsymbol{s}_{i:}$ with exceptions for g with small pop. density?

3 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.