Wuwei Zhang
Mentor: Taylor Okonek
2021/6/16

Disease Mapping Abstract

This quarter, under the guidance of my mentor, I learned a lot about disease mapping. We first spent some time talking about what different data types and maps are used for disease mapping. For **unit-level data or unit-level maps**, exact locations of the observations are recorded. So, the data is plotted as points on the map with these exact locations. It has the benefit that since geospatial information is likely to be encoded in terms of latitude and longitude, we know the location of each observation. However, if we just look at the raw data or the map, it is difficult to make conclusions about higher-level estimates. For **area-level data or area-level maps**, there are different aggregation levels we could consider. However, there are not specific benefits to having a larger or smaller aggregation level in the map we use, because it depends on the problems we want to address. In general, area-level maps makes it easy to see the overall distribution of the disease. However, if we only have area-level data, we don't know any more information within the area. Besides unit-level maps and area-level maps, **pixel-level maps** are quite popular nowadays, which need unit-level data or cluster level data. Often, if we want to fill in every pixel, we need statistical models to help us to process the raw data and generate data for every pixel. So, the drawbacks for pixel-level maps are obvious. Pixel-level maps have high uncertainty, because statistical models are used. Also, since the uncertainty is also at the pixel-level, it is difficult to see and distinguish uncertainty for each pixel and make general conclusions about the uncertainty.

After having an overall idea of what data and maps are like for disease mapping, we also talked about two common ways to represent data on the maps: **proportion and count**. Proportion shows the disease *prevalence* of an area in a more accurate way, because proportion takes the effect of population size into account. However, the government may need exact count to make public health policies. Thus, count data may be better for the central government to make these kind of public health policies.

Since most research in disease mapping uses **Bayesian data analysis**, Taylor introduced this world of Statistics to me, which is quite different from the frequentist inference I learned in intro Statistics class. Bayesian inference models uncertainty by a probability distribution over a prior belief (hypothesis), and then uses new data (likelihood) to update the distribution (posterior). So, we need probabilities for both hypotheses and data, and often need to construct a "subjective prior" before applying Bayesian data analysis.

$$\text{Posterior} = \frac{\text{Probability of the data} \times \text{Prior}}{\text{Average probability of the data}}$$

We also spent time reading and discussing an interesting paper, *Modeling and presentation of vaccination coverage estimates using data from household surveys* (by Tracy Qi Dong, Jon Wakefield). In this paper, the authors describe **two types of uncertainty within a**

vaccination coverage map: uncertainty within a state, which can be measured by spread of posterior distribution; and uncertainty between states, which is shown by overlaps of posterior distributions. By using ridgeline plots to show the posterior distribution of coverage estimate for each area, we can conclude that the last two states in this paper are separated from the rest, which means that they have the lowest vaccination coverage rate. But, for the rest areas, since they have too much overlap, we cannot make a solid conclusion.

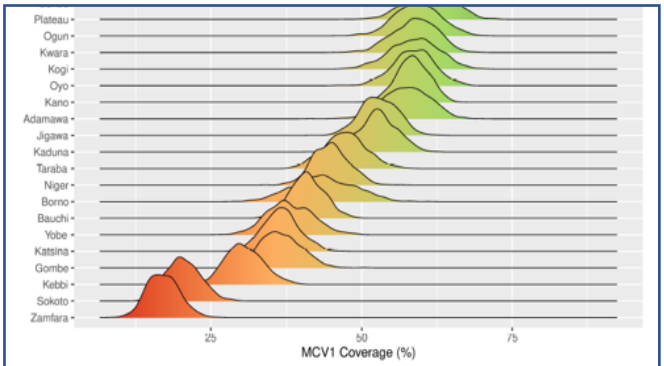

Figure 5: Ridgeline plots of the posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, ordered by posterior median, based on 1000 posterior samples from the *Binomial TS* model that includes the urban/rural strata variable.

Besides comparing uncertainty within a map, the new approach they proposed in this paper allows them to compute **uncertainty for the map as a whole**, which has the following steps. First, use a discrete color scale to represent a partition of the vaccination coverage rate. Then, use the posterior distribution of the coverage estimate for an area to assign that area to the interval that contains the greatest posterior probability (True classification probability, TCP). The last step is to calculate the average of the TCPs across all areas in a map (Average true classification probability, ATCP). This number, ATCP, measures the uncertain for the whole map. In conclusion, by comparing ATCPs of pixel-level map (ATCP=0.57), local-government-area-level map (ATCP=0.87), and state-level map (ATCP=0.94), we can conclude that coverage estimates at a finer spatial resolution tend to have larger associated uncertainty, and hence poorer precision. In particular, the pixel maps are often associated with huge uncertainties.