Estimating the County Health Rankings Using Bayesian Hierarchical Modeling

Ksenia Lepikhina
Masters of Science Candidate
Department of Applied Mathematics
University of Colorado Boulder
ksenia.lepikhina@colorado.edu
May 4, 2021

Public health officials can greatly benefit their communities by taking into account the results from the County Health Rankings. The County Health Rankings rank counties within each state based on a number of health outcomes and health factors. Though these rankings are valuable, they remain limited by their uncertainty quantification. This paper aims to use Bayesian hierarchical models to rank each county and measure differences between the County Health Rankings and the estimated ranks. Estimated rankings are produced for each of the health outcomes using two models. One of the model uses state and county level random effects. The other model uses state and county level random effects and includes demographic fixed effects (race, ethnicity, sex) and urbanization classification fixed effects. A successful implementation can provide a framework for estimating uncertainty in the rankings.

Keywords: County Health Rankings; Bayesian; Ranking

1 Introduction

The County Health Rankings, published by the University of Wisconsin Population Health Institute, are publicly available population and demographic based data that emphasize differences in health across all United States counties. While the County Health Rankings provide counties with valuable health information, they are limited by not having strong uncertainty quantification. This study will focus on quantifying the uncertainty of the 2020 County Health Rankings using Bayesian hierarchical models and spatial statistics techniques [1].

In this paper, I intend to reproduce and extend the results from the 2013 paper by Athens, et al. [2] "Using Empirical Bayes Methods to Rank Counties on Population Health Measures" used County Health Rankings data from 2010, demographic Census data from 2008, and Urbanization classification data from 2006. The data that will be used in this paper will include County Health Rankings data from 2020, demographic Census data from 2018, and Urbanization classification data from 2013. I plan on implementing similar hierarchical models to the paper by Athens, et al. using new data with the intention to replicate their results. Because all of the data is spatial, I intend to also look for spatial relationships among US counties. Using both the hierarchical models and spatial analysis, I also intend on predicting health outcomes for counties with missing data.

This research can help highlight health inequities across urbanization classes and demographics across United States counties. The Wisconsin Health Rankings allow policy makers and public health officials make decisions for county improvements based on the outcome of the rankings and the factors that went into receiving a certain ranking.

2 Related Work

The paper "Using Empirical Bayes Methods to Rank Counties on Population Health Measures" by Athens, et al. [2] uses 5 health outcome measures (premature death, self-reported health, mean physically unhealthy days per month, mean mentally unhealthy days per month, and percent of live births with low birth weight) from the 2010 County Health Rankings data and applies empirical Bayesian hierarchical models to find county rank estimates. Athens, et al. constructed two models – one with and one without demographic fixed effects – to determine which one improved rank precision. The paper used hierarchical Poisson models to quantify the uncertainty for the "mortality rates underlying premature mortality" [2], binomial models to model uncertainty for low birth weight and self-reported health, and log-normal models to estimate uncertainty for poor physical and poor mental health days.

Charles Courtemanche, et al. write in their paper "Modeling Area-Level Health Rankings" that "The [County Health Rankings] also do not account for uncertainty, despite their use of sample data for some components and an imputation process for missing data" [3] which makes it challenging to assess statistical differences between counties. Not accounting for error instills a false confidence in the Wisconsin

County Health Rankings which then leads to misinformed or inaccurate decisions. "Modeling Area-Level Health Rankings" uses a Bayesian hierarchical model for factor analysis to address problems pertaining to factor weighting and uncertainty. The paper then evaluates uncertainty by finding probability intervals for each of the rankings.

The paper "Measuring the spatial distribution of health rankings in the United States" by Will Davis, et al. write about a "method [that] relies on a factor analysis model to directly compute weights for [the] rankings, incorporate county population sizes into the variances, and allow for spillovers of health stock across county lines." [4] Their hierarchical models assumed Gaussian distributions for each of the health outcomes and they used posterior imputation to populate missing data. Their work finds that county health variation is highly dependent on demographic and economic variation.

3 Data

In order to perform the goals outlined in this paper, a number of data sources are necessary. First and foremost, the primary data source for this project is data from the 2013 County Health Rankings [5]. The dataset consists of two primary sections: health factors, and health outcomes. These following data is broken up by each United States county with sufficient data.

The health factors consist of features such as smoking rate, obesity rate, mortality rate, chlamydia rate, access to recreation, access to physicians/dentists, percent of population that is uninsured, percent of medicare enrolled mammography screenings, unemployment rate and many more. The health factors are presented in a number of different manners. Some is presented as a percentage (e.g. percentage of population with limited access to healthy food), some is presented as a count (e.g. annual violent crimes), and some is presented as an average (e.g. average daily PM25). The factors include 95% confidence intervals and Z-scores.

There are five primary health outcomes: premature deaths, poor or fair health, poor physical health days, poor mental health days, and low birth weight births. Premature death is measured by years of potential life lost before age 75 per 100,000 people and weights younger deaths as heavier than older deaths (e.g. a death at age 55 counts twice as much as a death at age 65, and a death at age 35 counts eight times as much as a death at age 70 [5]). Years of potential life lost is more useful than mortality because it is a metric that reflects a counties intention to care about preventable deaths.

Poor/fair health is measured by the percentage of adults who consider themselves to be in poor or fair health (self-reported and age adjusted) over a seven year period (2005-2011). Measuring self-reported health quality helps quantify challenges of disabilities and chronic diseases in a population. This health outcome is highly correlated with mortality – "people with 'poor' self-rated health had a twofold higher mortality risk than persons with 'excellent' self-rated health" [5].

Poor physical health days is measured as the average number of poor physical health days self-reported in the past 30 days. Similarly, poor mental health days is defined as the average number of poor mental health days in the past 30 days. This metric is useful for estimating how healthy people are while alive and is a reliable estimate of recent health. Average number of unhealthy days is also correlated with higher unemployment, poverty, percentage of adults who did not complete high school, mortality rates, and prevalence of disability than counties with fewer unhealthy days [5].

Low birth weight births is measured as a percentage of live births where the baby weighs less than 2500 grams (or just under 8 lbs) over a seven year period (2004-2010). This metric can give a good representation of infant morbidity. Low birth weight is a valuable public health indicator of health risks in all categories of the Country Health Rankings health factors.

Notably, self-reported outcomes may differ by race/ethnicity.

See Figure 1 for a summary of health factors and outcomes. The length of life health outcome (mortality) includes premature death and the quality of life outcome (morbidity) includes poor or fair health, poor physical health days, poor mental health days, and low birth weight births.

For information on how the rankings are calculated and further background on the data, visit the University of Wisconsin Population Health Institute's County Health Rankings website.

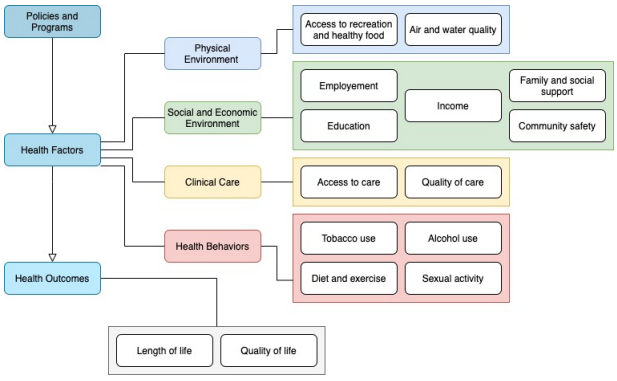

Fig. 1: County Health Ranking Model

The dataset also includes various demographics for each county with sufficient data in the US. The demographics dataset consists of the total estimated county population, percent of the population under 18 and over 65, percent of the population that is non-Hispanic African American, American Indian and Alaska, Asian, Native Hawaiian/ Pacific Islander, Hispanic, and non-Hispanic white, number of people with not proficient in English, and percent of the population that are female. These values were derived from the 2011 Census and the 5-year (2007-2011) American Community Survey estimates.

The final dataset that is used for this analysis is the National Center for Health Statistics (NCHS) Urban-Rural Classification Scheme for Counties from 2013. The urbanization classification are based off of "the Office of Management and Budget's (OMB) February 2013 delineation of metropolitan statistical areas (MSA) and micropolitan statis-

tical areas (derived according to the 2010 OMB standards for defining these areas) and Vintage 2012 postcensal estimates of the resident U.S. population" [6]. The classification values range from 1 to 6 where 1 is a large metropolitan area and 6 is "non-core". See Figure 2 for a map demonstrating these classifications.

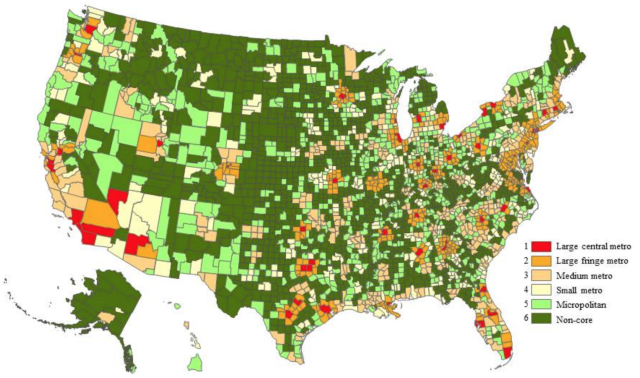

Fig. 2: Urbanization Classification. Map from [2]

4 Methods

For each of the five health outcomes (y – number of events) for each state (j) and county in each state (k), I intend to use the same hierarchical model presented by Athens, et al. [2] In each of these models, *AA* represents the African American demographic for a given county, *As* represents the Asian demographic for a given county, *AI* represents the American Indian demographic for a given county, *La* represents the Latino demographic for a given county, Urb represents the urbanization classification [6] for a given county (or Figure 2) and F represents the female demographic for a given county.

4.1 Bayesian Hierarchical Models

Premature deaths is modeled using a Poisson hierarchical model with a log link as premature deaths follows Poisson distribution. Premature deaths are reported as the total number of premature deaths and as years of potential life lost. Parameter $n_{j,k}$ is the county population (used for the offset) and parameter $\lambda j,k$ is the event rate for a county (e.g. premature death rate). e_j is the state-specific random effect parameter and $e_{j,k}$ is the county-specific random effect parameter. s_j and $s_{j,k}$ are the standard deviation of state-specific random effects and standard deviation of county-specific random effects respectively. In Model 1, the prior for the intercept was set to be distributed as a Gamma(7.5,1) and in Model 2, the prior for the intercept was the same and the prior for the β values was set to be Normal(0,10).

$$
\begin{aligned}
y_{j,k} &\sim \text{Poisson}(\lambda_{j,k}; n_{j,k}) \\
\text{Model 1} : \log(\lambda j,k) &= \beta_0 + e_j + e_{j,k} \\
\text{Model 2} : \log(\lambda j,k) &= \beta_0 + \beta_1 \text{AA} + \beta_2 \text{As} + \beta_3 \text{AI} + \beta_4 \text{La} + \\
&\quad + \beta_5 \text{Urb} + \beta_6 \text{F} + e_j + e_{j,k} \\
e_j &\sim N(0, s_j^2) \\
e_{j,k} &\sim N(0, s_{j,k}^2)
\end{aligned}
$$

A Binomial hierarchical model with a logit link is used to model low birth-weight births and self-reported health. Low birth-weight births are reported as the total number of low birth-weight births and as a percentage of the births in the county. Self-reported health is reported as a percentage. In Model 3, the prior for the intercept was set to be distributed as a Normal(1,1) and in Model 4, the prior for the intercept and the β values were set to be Normal(1,1). In Model 5, the prior for the intercept was set to be Normal(1,1) and in Model 6, the prior for the intercept, σ and the β values were set to be Normal(0,1).

$$
\begin{aligned}
y_{j,k} &\sim \text{Binomial}(p_{j,k}; n_{j,k}) \\
\text{Model 3,5} : \text{logit}(p_{j,k}) &= \beta_0 + e_j + e_{j,k} \\
\text{Model 4,6} : \text{logit}(p_{j,k}) &= \beta_0 + \beta_1 \text{AA} + \beta_2 \text{As} + \beta_3 \text{AI} + \beta_4 \text{La} + \\
&\quad + \beta_5 \text{Urb} + \beta_6 \text{F} + e_j + e_{j,k}
\end{aligned}
$$

Poor physical health days and poor mental health days are modeled using a Gaussian hierarchical model with an identity link. Both poor physical health days and poor mental health days are reported as averages. In Model 7, the prior for the intercept and σ were set to be distributed as a Normal(3,10) and in Model 8, the prior for the intercept, σ and the β values were set to be Normal(0,10). In Model 9, the prior for the intercept and σ values were set to be Normal(0,10) and in Model 10, the prior for the intercept, σ and the β values were set to be Normal(0,1).

$$
\begin{aligned}
y_{j,k} &\sim \text{Normal}(\mu_{j,k}, \sigma_{j,k}; n_{j,k}) \\
\text{Model 7,9} : \mu_{j,k} &= \beta_0 + e_j + e_{j,k} \\
\text{Model 8,10} : \mu_{j,k} &= \beta_0 + \beta_1 \text{AA} + \beta_2 \text{As} + \beta_3 \text{AI} + \beta_4 \text{La} + \\
&\quad + \beta_5 \text{Urb} + \beta_6 \text{F} + e_j + e_{j,k} \\
\sigma_{j,k} &\sim N(0, S_{j,k}^2)
\end{aligned}
$$

4.2 Missing Data

For the entire dataset, approximately 47.2% of data had at least one missing value in at least one of the columns (health factors, health outcomes, demographics, urbanization classification). The entire dataset included 3141 rows with one row for each county in each state. The reduced dataset contained only 1661 observations. In order to use the full 3141 observations, an imputation method using `R`'s `mice` package was implemented. In particular, for this dataset,

the CART (Classification and Regression Tree) [7] method was implemented. CART is a desirable method for imputation because it is resilient to outliers, does well with multicollinearity and skewed data, and can fit nonlinear functions and handle interactions. The general procedure for CART is as follows [8]:

1. Recursively partition the data
2. Fit a classification or regression tree
3. For each missing y value (often denoted as y_{mis}), find its final node in the fitted tree
4. Randomly draw from the members in the node and take the observed random draw value as the imputation result

In particular, for each model described in further detail below, I ran the `mice` CART imputation method with three multiple imputation and a single iteration.

The results of the imputation were then pooled using the `brms` (Bayesian Regression Models using Stan) library.

4.3 Sampling

Using the `brms R` package, the imputed data was passed into each of the following models. For each imputed dataset, the hierarchical Bayesian model was computed. The Bayesian estimates (such as the posterior mean) for each parameter were then averaged to return a final estimate. Each model is fit using a `Stan` back end.

`Stan` uses the Hamiltonian Monte Carlo (HMC) algorithm – a "Markov chain Monte Carlo (MCMC) method that uses the derivatives [(gradient)] of the density function being sampled to generate efficient transitions spanning the posterior" [9] – to provide Bayesian inference over a model conditioned on the data. In particular, `Stan` uses a sampler called the "No-U-Turn" (NUTS) sampler which is a form of HMC sampling.

The traditional Metropolis-Hastings random walk algorithm used to construct the Markov transitions in implementations of the MCMC fails in high-dimensional spaces. The HMC algorithm is capable of making larger jumps away from the initial point and thus better at exploring higher dimensional spaces. [10]

The goal of sampling is to draw samples for our parameters from a posterior distribution (probability of the parameters given the data). In order to do so, the HMC introduces an auxiliary momentum parameter (denoted as ρ) and the joint density between ρ and the models hyperparameters (θ) defines the Hamiltonian which can be separated into kinetic energy (the conditional probability) and potential energy (density of hyperparameters). The momentum parameter is drawn independently from a multivariate normal for each iteration. The joint density $p(\rho, \theta)$ is evolved using Hamilton's equations:

$$\begin{aligned}\frac{d\theta}{dt} &= -\frac{\partial}{\partial\rho}\log p(\rho,\theta) = -\frac{\partial}{\partial\rho}\log p(\rho|\theta)\\ &= -\log p(\rho)\\ \frac{d\rho}{dt} &= \frac{\partial}{\partial\theta}\log p(\rho,\theta) = \frac{\partial}{\partial\theta}\log p(\rho|\theta) + \frac{\partial}{\partial\theta}\log p(\theta)\\ &= \frac{\partial}{\partial\theta}\log p(\theta)\end{aligned}$$

Then with a new draw of a momentum parameter, the current value of the hyperparameters are updated using a leapfrog integrator (numerically solving the differential equations). Finally, to account for numerical error in the leapfrog integrator, a Metropolis acceptance step is applied where the probability of keeping the proposed new parameters (ρ^*, θ^*) (generated by moving from the previous parameters – ρ, θ) is:

$$\min(1, \exp(-\log p(\rho,\theta) + \log p(\rho^*,\theta^*)))$$

If the proposed parameters are not accepted, the previous parameters are then used to initialize the next iteration once again. [9]

4.4 Ranking

4.4.1 MSE Loss Ranking

In order to rank all US counties within each state for each health outcome, the models described in section 4.1 were used to find estimates of the hyper parameters $(\beta_0, ..., \beta_6, e_j, e_{j,k})$. The estimates were derived from the posterior means of each posterior distribution. The parameters were then used to find estimates $\hat{y}_{j,k}$. The $\hat{y}_{j,k}$ values were then grouped by state and then ranked (R_{est}).

The rankings were then compared to the "true" 2013 County Health Rankings (R_{true}) using Mean Squared Error (MSE) Loss [11].

$$\hat{L} = \frac{1}{k}\sum_{i=1}^{k}(R_{est} - R_{true})^2$$

5 Evaluation and Results

5.1 Evaluation

In order to diagnosis each of the hierarchical models, "Shiny Stan", a user interface provided by `Stan`, was used. Thinning is typically used to reduce the final autocorrelation in the dataset [12] and to benefit memory computational constraints. Another method that is commonly used to improve the results of MCMC chains is the concept of "burn-in". When running an MCMC, some initial starting points are better than others. Burn-in attempts to solve the problem of a poor start point. Essentially, burn-in is used to remove iterations that start in the tail of the equilibrium distribution.

For this paper, the Markov Chains were not thinned and a burn-in period was not used as the machine this work was implemented on had sufficient memory and the iteration run time was already quite long.

"Shiny Stan" does not include information about thinning or burn-in, however it provides a number of useful figures that can help assess convergence. Trace plots are an example of a tool that can be used to identify problem areas with a model. For example, multimodality in a trace plot can be identified by jumps between different distributions and regions where the sampler reaches and has difficulty returning to the main distribution in the trace plot can indicate wide posterior tails [12].

Another metric is the MCSE (The Monte Carlo standard error). It is an estimate of the error of the estimate for the posterior mean based on the posterior standard deviation estimate and the number of effective samples.

$$MCSE(\bar{\theta}) = \frac{s}{\sqrt{N_{eff}}}$$

where θ is the posterior mean of the parameter, s is the estimated posterior standard deviation and N_{eff} is the effective sample size. Because an MCMC is typically not independent, the Monte Carlo standard error will often be higher than if it was independent [12].

For the Hamiltonian Monte Carlo method implemented specifically in `Stan` there are a couple of additional diagnostics. The three specific HMC diagnostics are: checks for divergent transitions, checks for maximum tree-depth, and information on the Bayesian fraction of missing information. In general, stepsize and tree-depth are the two tuning parameters for adjusting the three diagnostic outputs. `Stan` tries to find the optimal stepsize and tree-depth during warm up but does not always find the best settings.

If `Stan` notifies that there are divergent transitions then "the sampler is not drawing samples from the entire posterior and inferences will be biased." [12] Reducing the step-size helps mitigate this however it also means that it will require more steps to explore the space. Getting the "maximum tree depth reached" warning means that more steps need to be taken but are capped at the maximum. A larger tree depth will help with model efficiency and faster space exploration. Bayesian fraction of missing information can be understood as – if you have a heavy tail distribution, you need large momentum to go to the tail, and a Gaussian momentum proposal (which is common for HMC) is not enough momentum to get you there.

Han Liu and Larry Wasserman wrote in their textbook "Statistical Machine Learning" (2014) provided additional guidelines on how to analyze trace plots. In particular, the paper provides guidelines on how to specify the variance parameter for Gaussian models. Gaussian models with too low variances tend to have acceptance rates that are too high and models with too high variances tend to have acceptance rates that are too low [13]. Gelman (1996) wrote that the ideal acceptance rate should be around 0.25 [14].

5.2 Results

The County Health Rankings are presented as two primary health outcomes 1. The two health outcomes are length of life and quality of life. Length of life (mortality) is measured by premature death and quality of life (morbidity) is measured by the average number of poor physical health days out of 30 days, the average number of poor mental health days out of 30 days, low birth weight births and the percentage of adults reporting poor or fair health. Mortality and morbidity are equally weighted in terms of the final county ranking.

After fitting each of the hierarchical Bayesian models, I extracted the posterior mean for each of the parameters and predicted each of the elements that go into the health outcomes. For mortality, each of the estimated values for premature death were ordered within each state and ranked. These ranked values were then compared with the "true" ranks from the County Health Rankings using the mean squared error loss described 4.4.1. The mean squared error loss was quite poor and ranged from 0 to approximately 10,000 depending on the state (Figure 3). The District of Columbia had the smallest mean squared error loss because there is only one county there and therefore the predicted rank was 1 and the true rank was 1.

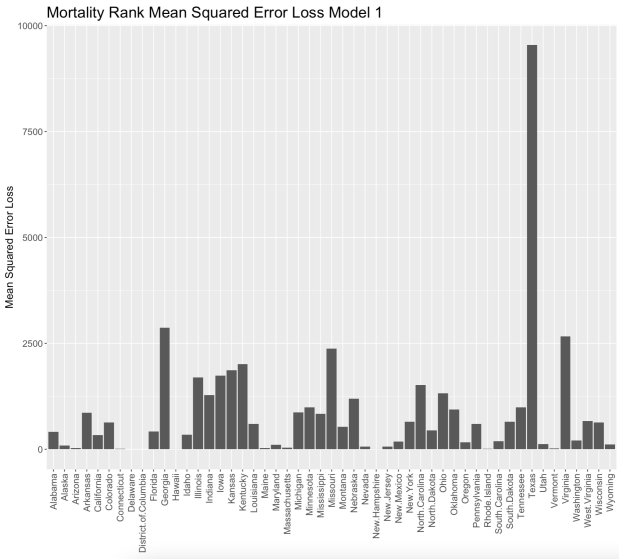

Fig. 3: Mortality Mean Squared Error Loss By State – Model 1

For the second model (includes demographics), each of the estimated values were estimated, ordered within each state and ranked. These ranked values were then also compared with the "true" ranks from the County Health Rankings using the mean squared error loss described 4.4.1. The mean squared error loss for this mortality score was poor (very high) as well and ranged from 0 to approximately 10,000 depending on the state (Figure 4). Once again, the District of Columbia had the smallest mean squared error loss.

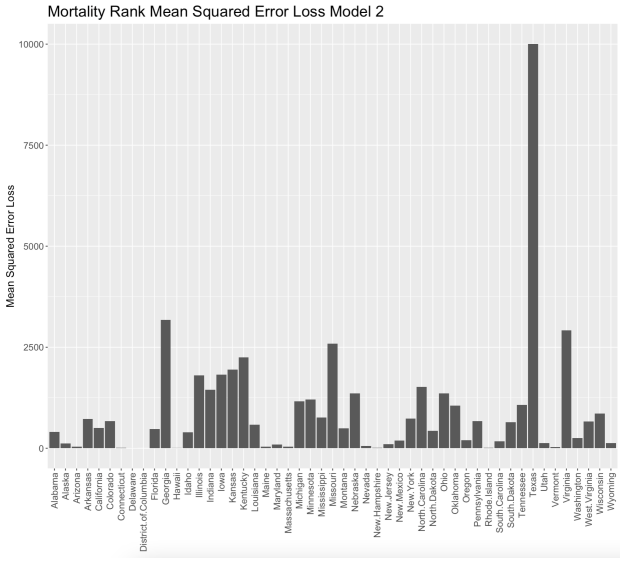

Fig. 4: Mortality Mean Squared Error Loss By State – Model 2

Comparing Figure 4 to Figure 3, there is no substantial difference. Figure 5 show the relationship between the mean squared error loss between Model 1 and Model 2 for mortality. The R^2 value is 0.997 with the regression noting that the relationship is significant with a p-value significantly smaller than the standard α of 0.05. This implies that there is no significant difference between the two models.

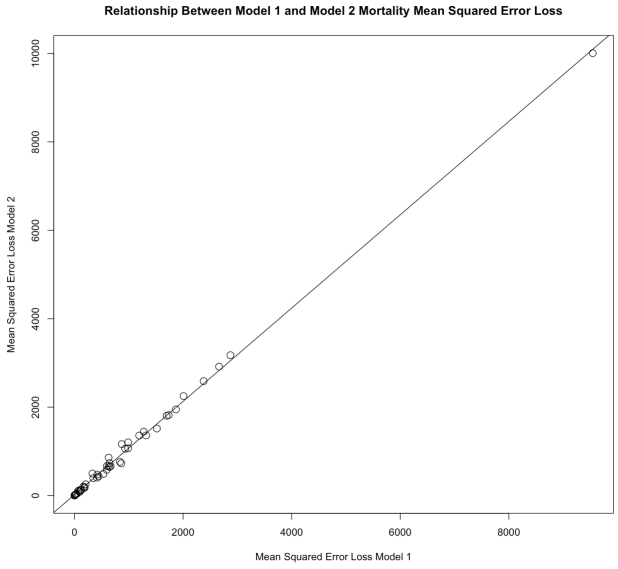

Fig. 5: Mortality Model Comparison

Similarly, the mean squared error follows a similar trend to the two mortality models in the two morbidity models. Figures for morbidity mean square error by state, see Figures 14 and 15 in the appendix. Figure 6 shows the correlation between the first morbidity model (with just county/state random effects) and the second morbidity model (county/state random effects and demographic fixed effects). In this figure, the correlation is nearly perfect, once again implying that there is no significant difference between the two morbidity models.

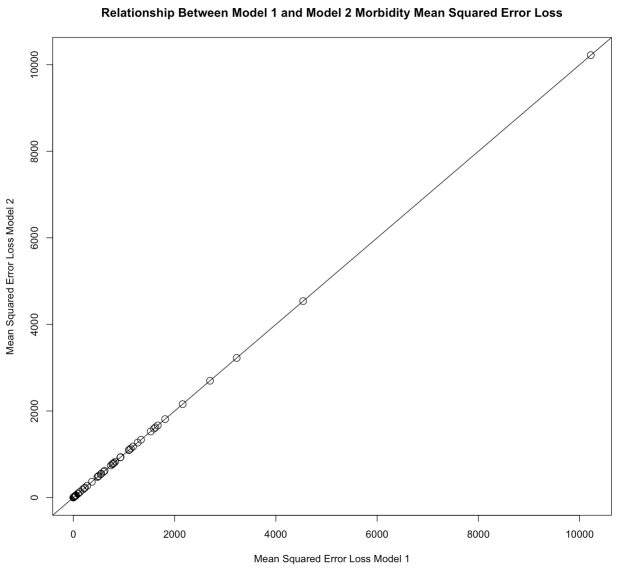

Fig. 6: Morbidity Model Comparison

Notice that in the past few figures, the mean squared error for some of the states is exceptionally large. A substantial amount of the error can be accounted for by studying trace plots, the posterior distribution, and the mean Metropolis acceptance values. For the analyses in this paper, some of the difficulties arose from not having sufficient resources to find "good" posterior estimates. Some of the Bayesian estimates for health outcomes did fairly well. For example, Figure 7 shows the trace plot for the county random effect for poor health percentage health outcome. The models implemented here used 12 chains with 4000 iterations. No burn in or thinning was done. The trace plot shows no signs of divergence and shows that the MCMC is exploring the full space.

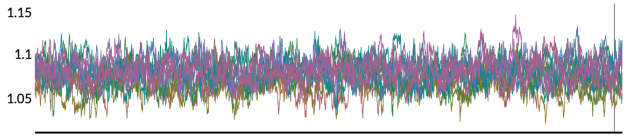

Fig. 7: MCMC trace plot of the estimate for the county random effect of the percentage of adults who consider themselves to be in poor or fair health (self-reported and age adjusted) over a seven year period (2005-2011)

Figure 8 shows the posterior distribution for the county random effect for poor health percentage health outcome. The figure shows that the posterior distribution is likely unimodal and not skewed.

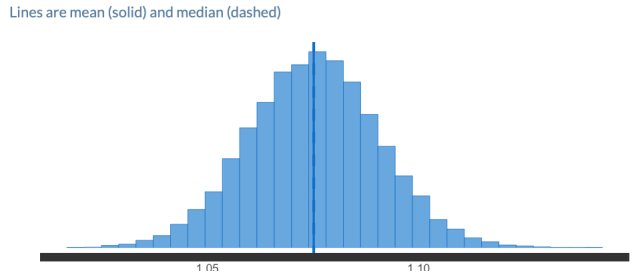

Fig. 8: Posterior distribution of the estimate for the county random effect of the percentage of adults who consider themselves to be in poor or fair health (self-reported and age adjusted) over a seven year period (2005-2011)

Figure 9 shows the mean Metropolis acceptance for the county random effect for poor health percentage health outcome. The figure shows that around the posterior mean, the Metropolis acceptance is around 0.45.

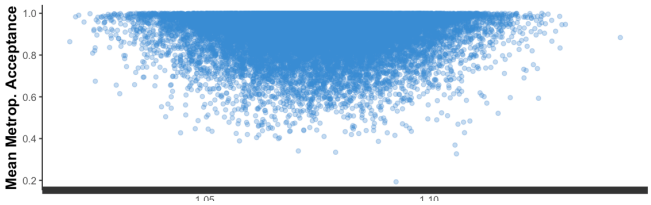

Fig. 9: Mean Metropolis Acceptance of the estimate for the county random effect of the percentage of adults who consider themselves to be in poor or fair health (self-reported and age adjusted) over a seven year period (2005-2011)

An example of a diverging model can be seen with the poor physical days model. Figure 10 shows the trace plot for the county random effect for poor physical days (model 1). Similarly, this model was implemented using 12 chains with 4000 iterations and no burn in or thinning. The trace plot shows substantial signs of divergence and shows that the MCMC can not settle on a particular mean value.

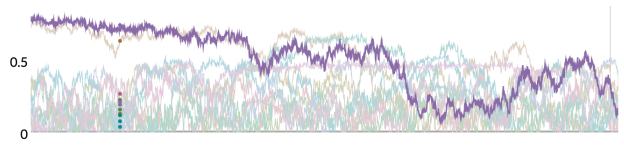

Fig. 10: MCMC trace plot of the estimate for the county random effect of the average number of poor physical health days self-reported in the past 30 days

Figure 11 shows the posterior mean for the county random effect for poor physical days. The figure shows that the posterior distribution is likely multimodal and suggests that the posterior mean may not be the best parameter estimate.

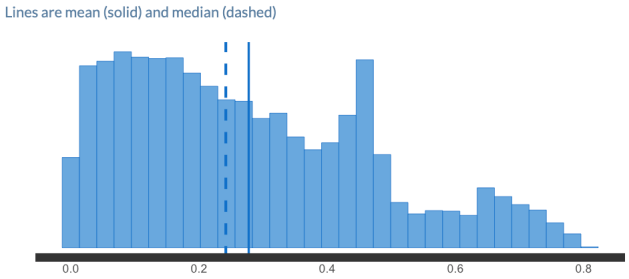

Fig. 11: Posterior distribution of the estimate for the county random effect of the average number of poor physical health days self-reported in the past 30 days

Figure 12 shows the mean Metropolis acceptance for the county random effect for poor physical days. The figure shows that around the posterior mean, the Metropolis acceptance is all over the place with the red points indicating divergence.

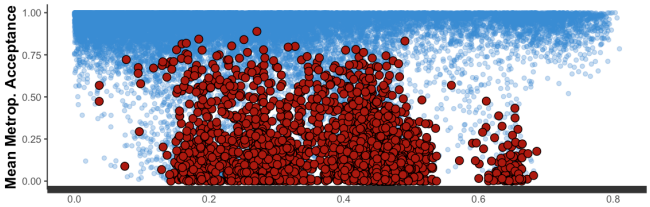

Fig. 12: Mean Metropolis Acceptance Rate of the estimate for the county random effect of the average number of poor physical health days self-reported in the past 30 days

Figure 13 shows a notable trend that the squared error loss for the morbidity model that excludes demographics increases when the number of counties within a state increase. Texas has the highest squared error loss (9544.91) and has the highest number of counties (254). This trend can also be seen in the second morbidity model and the two mortality models (Figures 16, 17, 18).

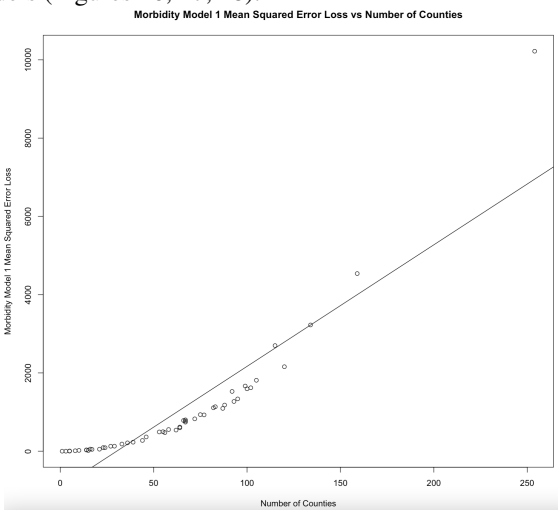

Fig. 13: Morbidity Model 1 MSE Loss and Total Number of Counties in a State in the County Health Rankings

6 Future Work

One area to explore further is alternative ranking methods. The paper "Loss Function Based Ranking in Two-Stage, Hierarchical Models" [11] by R. Lin, et al. provides an excellent overview of how ranking using Bayesian hierarchical models can be improved. The paper wrote that for effective ranking, the loss function should consist of a comparison between estimated and true ranks. R. Lin, et al. showed that this "produces optimal or near optimal performance" [11]. Future work for this paper could include an investigation into additional loss functions (i.e. weighted mean squared error loss, $100(1-\gamma)\%$ loss, summed unit specific loss – ratio of misclassifications, etc.)

The paper by Athens, et al. [2] estimated the parameters in the hierarchical models using Linear Mixed-Effects Modeling (`lme4` in R – see example of linear mixed effect model here [15]). This paper did not dive into frequentist methods and largely focused on Bayesian techniques. Another area to look into is comparing the frequentist methods and the Bayesian methods implemented in this paper. It would be interesting to quantify the differences in performance between

the two statistical schools of thought. In particular, studying the mean squared error loss differences and ranking accuracy compared to the "true" data source from the Wisconsin County Health Rankings.

Another interesting area to look into is performing a spatial analysis across all counties and ranking across the US. The rankings are performed within states because states have the power to allocate resources to their counties. A ranking across the US is challenging due to the lower chance of change within a county. However, investigating the rankings across the US could point to larger clusters of "problem" areas (areas with worse rankings).

Finally, it is super important to extend this research by looking at how different ranking methods rank the counties within each state. It's possible that distinct methods highlight stark differences in either urbanization or demographic. It would be fascinating to study the evolution of the rankings over time across different ranking methods to highlight public health growth areas.

7 Conclusions

The purpose of this paper was to estimate the County Health Rankings using Bayesian hierarchical modeling to provide a framework for quantifying the error in the estimates. The methods presented in this paper find that Bayesian hierarchical modeling has the potential of satisfying that goal. The models for morbidity and mortality were able to estimate the "correct" rankings for states with fewer counties. Additionally, this paper implies that there is no significant difference between the model that includes demographics as fixed effects and the model that simply uses state and county level random effects.

Many of the models for the county health outcomes suffered from divergence. Dedicating more time to hyperparameter tuning and more computational resources can lead to better convergence and better posterior estimates. Better posterior estimates could potentially also present more substantial differences between models that include demographic fixed effects and models that do not. With more resources, Bayesian hierarchical modeling can successfully quantify error in the County Health Rankings.

Acknowledgements

The author would like to thank Dr. Debashis Ghosh, Dr. Vanja Dukic, Dr. Judith Law, and Dr. Brian Zaharatos for their endless support on this project. Thank you to Dr. Ghosh (primary advisor) for providing excellent guidance, suggestions, and additional readings. Thank you to Dr. Dukic for the connection with Dr. Ghosh and for offering additional help and literature suggestions. Thank you to Dr. Law and Dr. Zaharatos for continuous support throughout the Applied Math Masters program.

References

[1] Chi, G., and Zhu, J., 2008. "Spatial regression models for demographic analysis". *Population Research and Policy Review,* **27**(1), pp. 17–42.

[2] Athens, J. K., Catlin, B. B., Remington, P. L., and Gangnon, R. E., 2013. "Peer reviewed: Using empirical bayes methods to rank counties on population health measures". *Preventing chronic disease,* **10**.

[3] Courtemanche, C., Soneji, S., and Tchernis, R., 2015. "Modeling area-level health rankings". *Health services research,* **50**(5), pp. 1413–1431.

[4] Davis, W., Gordan, A. D., and Tchernis, R., 2020. Measuring the spatial distribution of health rankings in the united states. Tech. rep., National Bureau of Economic Research.

[5] The county health rankings. https://www.countyhealthrankings.org/.

[6] Nchs urban-rural classification scheme for counties. https://www.cdc.gov/nchs/data_access/urban_rural.htm#2013_Urban-Rural_Classification_Scheme_for_Counties.

[7] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A., 1984. *Classification and regression trees*. CRC press.

[8] 3.5 classification and regression trees.

[9] Team, S. D. 15.1 hamiltonian monte carlo.

[10] Betancourt, M., 2017. "A conceptual introduction to hamiltonian monte carlo". *arXiv preprint arXiv: 1701.02434*.

[11] Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G., 2006. "Loss function based ranking in two-stage, hierarchical models". *Bayesian Analysis (Online),* **1**(4), p. 915.

[12] Arnold, J. B. 8 mcmc diagnostics.

[13] Liu, H., and Wasserman, L., 2014. *Ch 12: Bayesian Inference*. p. 299–251.

[14] Gelman, A., Roberts, G., and Gilks, W., 1996. "Efficient metropolis jumping hules". *Bayesian statistics*.

[15] Magnusson, K. Using r and lme/lmer to fit different two- and three-level longitudinal models.

[16] Social explorer. http://www.socialexplorer.com/.

[17] , 2018. A practical guide to mcmc part 1: Mcmc basics, Jan.

[18] Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., and Bürkner, P.-C., 2019. "An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian". *Journal of Speech, Language, and Hearing Research,* **62**(5), pp. 1225–1242.

8 Appendix

The following two figures show the mean square error loss for each state for both morbidity models. These figures are included to show that the trend is common across all four models.

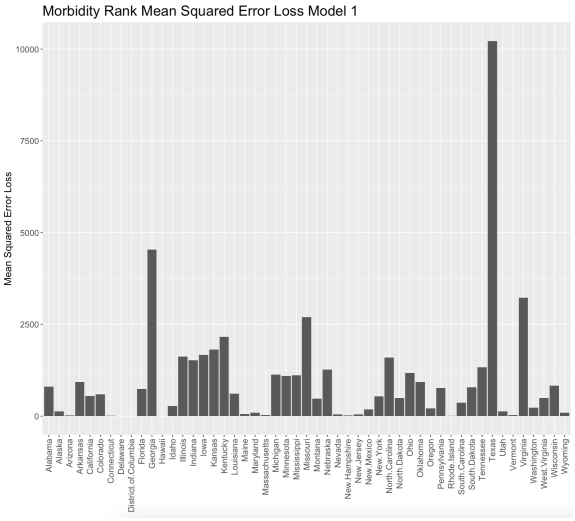

Fig. 14: Morbidity Model 1 MSE Loss by State

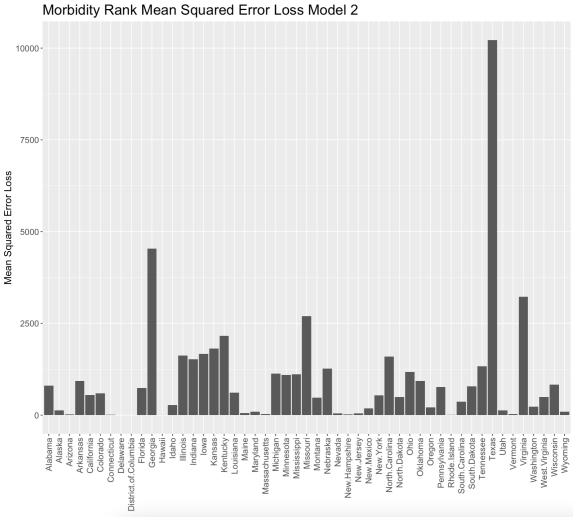

Fig. 15: Morbidity Model 2 MSE Loss by State

The following three figures show the mean square error loss with respect to the total number of counties. These figures are included to show that the trend is common across all four models.

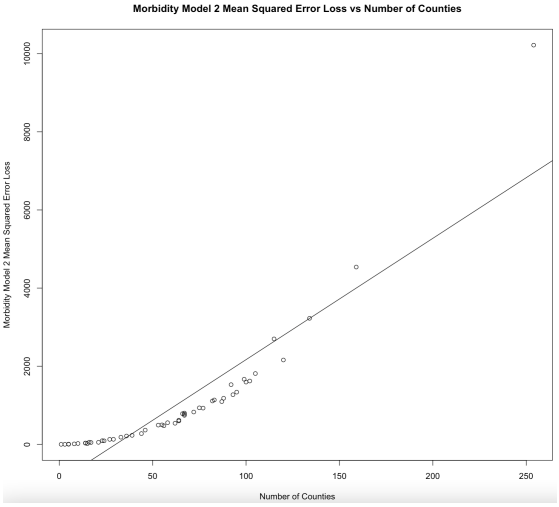

Fig. 16: Morbidity Model 2 MSE Loss and Total Number of Counties in a State in the County Health Rankings

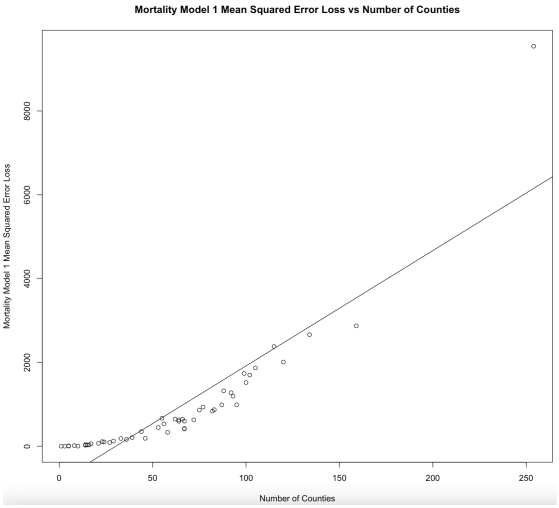

Fig. 17: Mortality Model 1 MSE Loss and Total Number of Counties in a State in the County Health Rankings

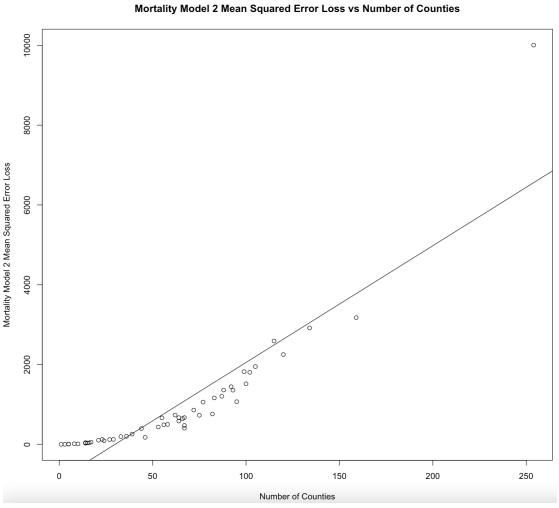

Fig. 18: Mortality Model 2 MSE Loss and Total Number of Counties in a State in the County Health Rankings

The following two figures show the mean square error loss with respect to the population of the states.

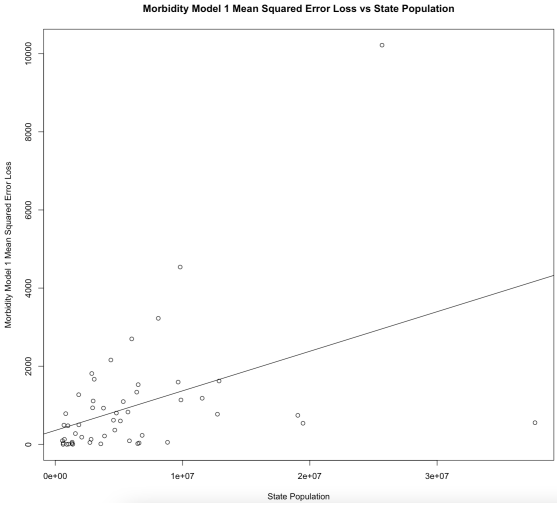

Fig. 19: Morbidity Model 1 MSE Loss and Total County Populations

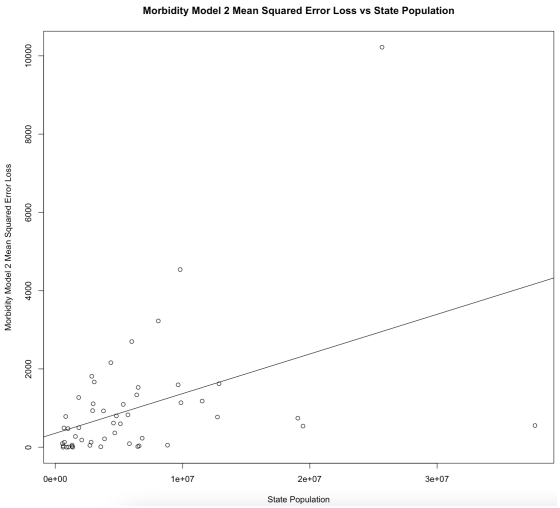

Fig. 20: Morbidity Model 1 MSE Loss and Total County Populations

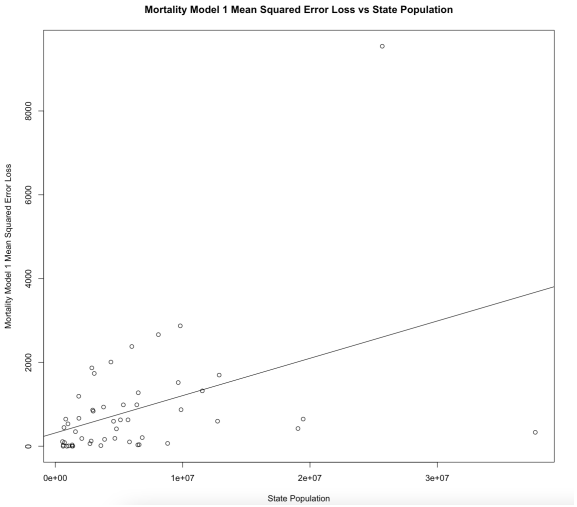

Fig. 21: Mortality Model 1 MSE Loss and Total County Populations

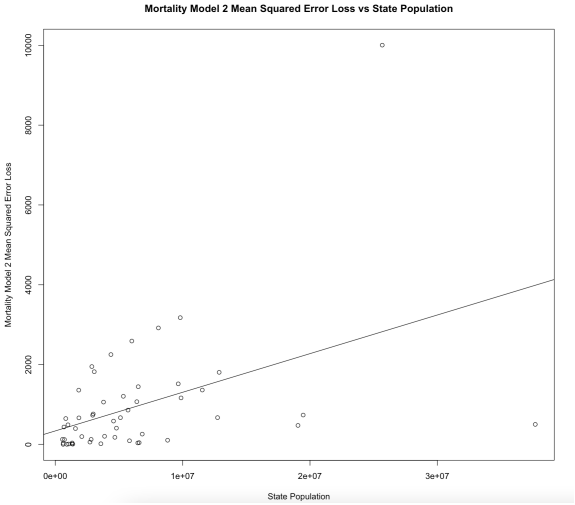

Fig. 22: Mortality Model 2 MSE Loss and Total County Populations