Modeling Area-Level Health Rankings

Charles J. Courtemanche\textsuperscript{1}  Samir Soneji\textsuperscript{2}  Rusty Tchernis\textsuperscript{3}

\textsuperscript{1}Georgia State University & NBER, \textsuperscript{2}Dartmouth College
\textsuperscript{3}Georgia State University, IZA & NBER

July, 2014
Overview

- Population Health measures
- UWPHI methods and ranks
- Our contribution
- Spatial factor analysis model
- Results
- Extensions
Population Health

- The goal is to rank areas based on health of the population
- The problem is that “health” is not observable directly
- There are two levels of population health measures available - state and county level
- Those are used to compare health levels across areas and to identify disadvantaged areas (mobilizing action vs. resource allocation)
- RWJF funded www.countyhealthrankings.org
County Health Ranks

- We will only concentrate on “Mortality and Morbidity” ranks
- The variables used are county level averages of
  1. YPLL (2005-2007 NCHS)
  2. % reporting fair or poor health (2003-2009 BRFSS)
  3. # physically unhealthy days a month (2003-2009 BRFSS)
  4. # mentally unhealthy days a month (2003-2009 BRFSS)
  5. % low birth weight (2001-2007 NCHS)
UWPHI ranks

- Each variable $j$ is standardized:
  $$Z_{ij} = \frac{Y_{ij} - \mu_j}{\text{St.d.}(Y_{ij})}$$

- For each county an overall $Z$ — score is computed
  $$Z_i = .5 * Z_{i1} + .1 * Z_{i2} + .1 * Z_{i3} + .1 * Z_{i4} + .2 * Z_{i5}$$

- Each county is assigned a rank based on its $Z$ — score.
- $Z$ — scores are ranked and ranks are reported/mapped
- Example of maps
Potential Limitations

1. Add-hoc weights
2. No measure of uncertainty
3. County population sizes are not incorporated
4. Counties are assumed to be spatially independent
Our contribution

We propose to use Hogan and Tchernis (2004) spatial factor analysis model

1. Data-dependent weights
2. Results can be summarized using distribution of county ranks (representation of uncertainty)
3. County population sizes are incorporated into the variance of both factors and errors
4. Incorporated spatial correlation of factor scores
UWPHI Framework

County Health Rankings Deterministic Model

- Low Birthweight Births
- Self-Reported Health
- Premature Death
- Mentally Unhealthy Days
- Physically Unhealthy Days

Health Outcomes

0.20 0.10 0.50 0.10 0.10
Our Framework

Latent Variable Model

Health

Low Birth-weight Births
Self-Reported Health
Premature Death
Mentally Unhealthy Days
Physically Unhealthy Days
Proposed Model - step 1

\[ Y_{ij} = \mu_j + \lambda_j \delta_i + e_{ij} \]

\[ \mu_j = E(Y_{ij}) \]
\[ = \text{average of variable across areas} \]

\[ \delta_i \sim N(0, 1) \]
\[ = \text{latent county health level for area } i \]

\[ \lambda_j = \text{factor loading} \]
\[ = Cov(Y_{ij}, \delta_i) \]

\[ e_{ij} \sim N(0, \sigma_j^2) \]
\[ = \text{unexplained error} \]
Proposed Model - compact form

**Level I.** Within-area variation

\[ Y_i | \delta_i \sim N(\mu + \lambda \delta_i, \Sigma), \sigma_{jk} = 0 \forall j \neq k \]

or in stacked form

\[ Y | \delta \sim N(\mu + \Lambda \delta, I_n \otimes \Sigma) \]

**Level II.** Between-area variation

\[ \delta \sim N(0, I_n) \]

Requires normality and fixed variance for identification
Proposed Model - steps 2 and 3

**Step 2:** Incorporating population sizes:

\[ M = \text{diag}\{m_i\} \]

\( m_i \) is population of county \( i \)

\[ Y|\delta \sim N(\mu + \Lambda \delta, M^{-1} \otimes \Sigma) \]

\[ \delta \sim N(0, M^{-1}) \]

**Step 3:** Incorporating spatial correlation:

\[ Y|\delta \sim N(\mu + \Lambda \delta, M^{-1} \otimes \Sigma) \]

\[ \delta \sim N(0, M^{-1/2}\Psi M^{-1/2}) \]

**Implication:** \( E(\delta_i|Y) \) may depend on variables from other counties - spatial smoothing.
Spatial Correlation

- We work with the Conditional Autoregressive (CAR) specification (Besag, 1974; Sun et al., 1999)

\[
\delta_i | \{ \delta_j : j \in R_i \} \sim N\left( \sum_{j \in R_i} \beta_{ij} \delta_j, \nu / \alpha_i \right),
\]

or marginally

\[
\delta \sim N(0, (I - \omega R)^{-1})
\]

where \( \beta_{ij} = \omega \), and \( \nu / \alpha_i = 1 \),

- \( R_{ij} = 1 \) if a county \( j \) is adjacent to county \( i \) and \( R_{ii} = 0 \).
- Thus \( \Psi = (I - \omega R)^{-1} \) is a full matrix inducing the correlation between variables between counties.
The model is estimated using MH within Gibbs Sampler, separately for each state.

We obtain 5,000 draws from the posterior distribution of parameters.

At each iteration we rank the draws of posterior mean of factor scores which produces a draw from the posterior distribution of county ranks.

In addition, we can compute probability that each county is in a certain range, e.g. top quintile of ranks (the worst health).
Data

- We use data from WI and TX
- WI has been researched in a number of contexts and has less missing data (only 2 observations out of 72 counties)
- TX is a larger state, but has a lot of missing data (out of 254 counties 31 are not ranked by UWPHI, and 116 of remaining counties have at least one missing observation)
- For easier comparison we will OLS predictions for missing observations (CHR uses means), but later we will incorporate missing data imputation
UWPHI and Posterior Ranks - Wisconsin

Wisconsin

UWPHI Rank

Posterior Rank

Courtemanche & Soneji & Tchernis
Modeling Area-Level Health Rankings
20th IPDC July, 2014
Weights and Squared Correlations

<table>
<thead>
<tr>
<th></th>
<th>UWPHI $w$</th>
<th>$\rho^2$</th>
<th>95% CI</th>
<th>Texas $\rho^2$</th>
<th>95% CI</th>
<th>Wisconsin $\rho^2$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premature Death</td>
<td>0.50</td>
<td>0.14</td>
<td>(0.09,0.19)</td>
<td>0.27</td>
<td>(0.17,0.38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health Status</td>
<td>0.10</td>
<td>0.24</td>
<td>(0.20,0.29)</td>
<td>0.21</td>
<td>(0.12,0.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phys. Unhealthy Days</td>
<td>0.10</td>
<td>0.41</td>
<td>(0.34,0.48)</td>
<td>0.21</td>
<td>(0.11,0.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ment. Unhealthy Days</td>
<td>0.10</td>
<td>0.15</td>
<td>(0.10,0.20)</td>
<td>0.17</td>
<td>(0.08,0.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Birthweight Births</td>
<td>0.20</td>
<td>0.06</td>
<td>(0.02,0.10)</td>
<td>0.15</td>
<td>(0.05,0.24)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Maps of Probability of Being in the Top Quintile
Maps of Probability of Being in the Top Quintile

UWPHI Ranks

Posterior Ranks

Courtemanche & Soneji & Tchernis
Modeling Area-Level Health Rankings
20th IPDC July, 2014
How Far Apart to be Different

![Graph showing health outcome rank and percentage overlap in Harris County, TX, Blanco County, TX, Texas, and Wisconsin.](image-url)
Dealing with Missing Data

- So far we used naive predictions of missing data and did not incorporate the uncertainty of prediction.
- Our method extends to predicting missing data in one simple step.
- Conditional on parameters, the distribution of the missing data is given by the model:

\[
Y_{ij}^{\text{miss}} = \mu_j + \lambda_j \delta_i + e_{ij},
\]

\[
e_{ij} \sim N(0, \sigma_j^2)
\]

- We rerun the model for all counties used so far as well as a subset of counties with at most one variable missing.
Naive Imputation

Posterior Rank

Texas UWPHI Rank

Courtemanche & Soneji & Tchernis  Modeling Area-Level Health Rankings  20th IPDC July, 2014  27 / 35
Posterior Imputation

Plot showing the relationship between Texas UWPHI Rank and Posterior Rank.
Conclusions

- Same data can lead to different rankings
- While there is a lot of agreement in WI we contribute in terms of uncertainty
- In TX we agree on fewer counties
- However, even in TX after accounting for missing data imputation there are a few counties we agree on
- These methods can be used for many other applications
Future Work

- Add other variables - multiple factors
- Should the model be estimated for the entire country or state by state?
- Other applications: quality of cancer care, hospital quality, county childhood obesity rates, etc.
- Investigate reliability of rankings (in the spirit of Arndt et al., 2013)
- Run a horse race between methods
Thank you!