Simultaneous Clustering and Ranking of County-Level Health Outcomes

Ron Gangnon & Cora Allen-Coleman

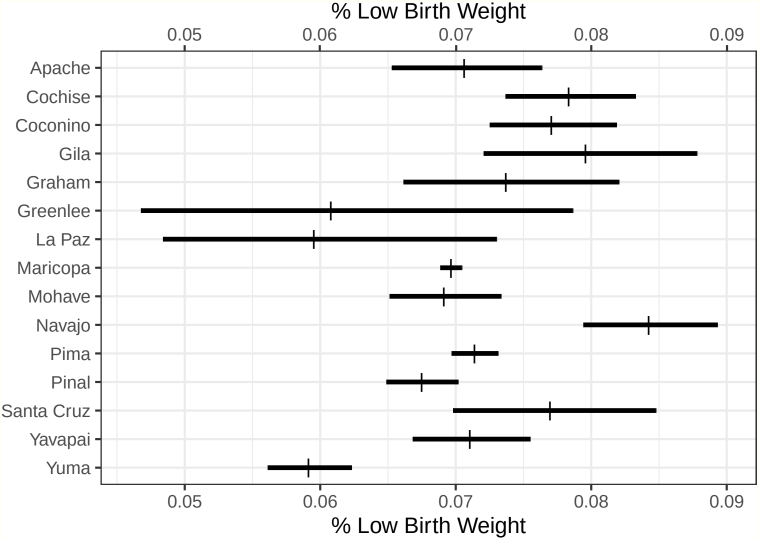
University of Wisconsin-Madison

7 January 2020

Background

- Ranking of counties (or other geopolitical regions) based on health indices is a common inferential goal in public health.
- Bayes (and empirical Bayes) methods provide (joint) posterior distributions for county health indices and corresponding county ranks.
- Inferences (point and distributional) regarding ranks should be guided by appropriate loss function/inferential goal.
- Illustrative example: % of live births with low birthweight ($<$2,500 g) in Arizona counties, 2017.

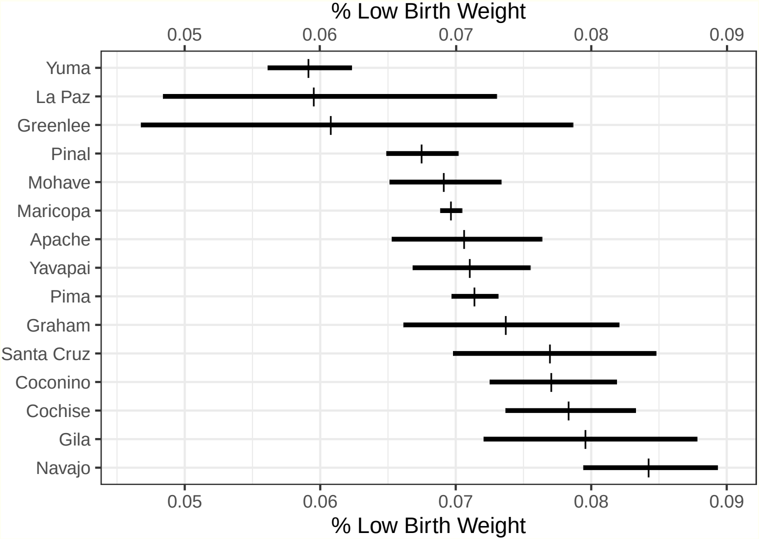
Arizona % Low Birth Weight



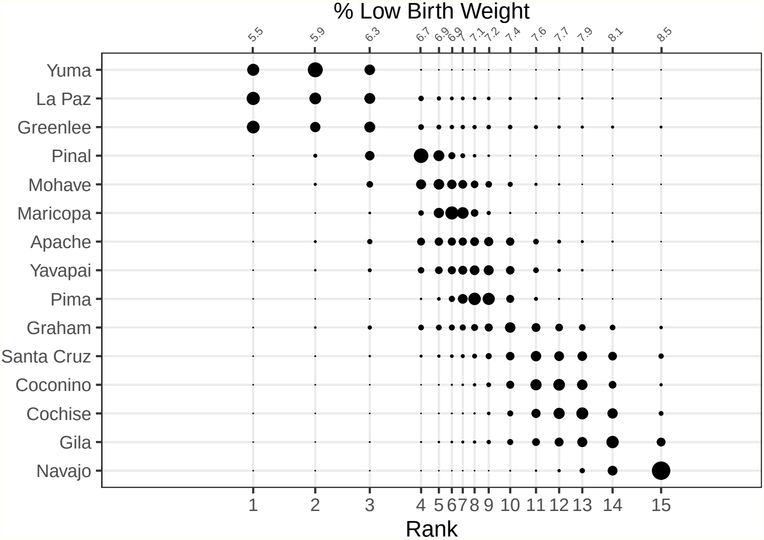
Inference Regarding Ranks

- Optimal point estimates for ranks minimize squared error loss on the proportion (% LBW) scale.
- Posterior distributions for county-specific ranks can either be displayed using dot plot (with dot size proportional to posterior) or summarized with highest posterior density (HPD) intervals.

Arizona % Low Birth Weight in (Optimal) Rank Order



Posterior Distribution of Ranks



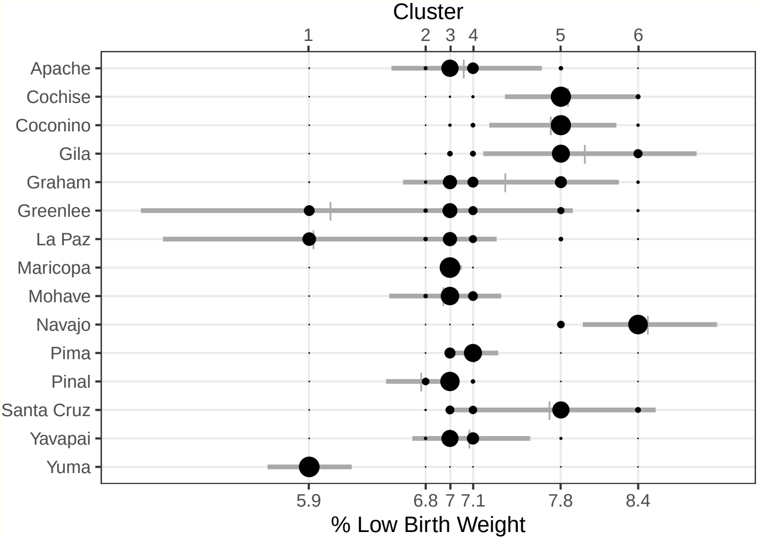
Concerns

- High degree of uncertainty regarding most ranks with the exception of large counties at the extremes of the distribution or outliers.
- High variability in one county necessarily influences uncertainty of ranks for all other counties.
- Hard to identify meaningful distinctions between ranks given the high level of noise.
- Potential (partial) solution based on discussions with County Health Rankings staff: Clustering

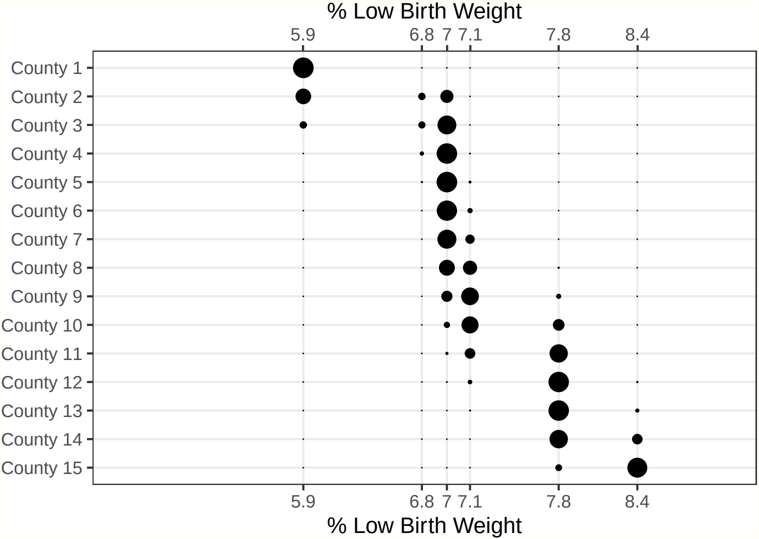
Nonparametric Mixture Model

- Likelihood: $y_i \sim \text{Binomial}(n_i, p_i), i = 1, 2, \ldots, N$
- Nonparametric prior: $\Pr(p_i = \theta_j) = \gamma_j, j = 1, 2, \ldots, m \leq N$
- Maximum likelihood estimates for number of mixture components m, support points $\theta_1, \theta_2, \ldots, \theta_m$ and (prior) probability $\gamma_1, \gamma_2, \ldots, \gamma_m$ from EM algorithm.
- Empirical Bayes posterior distributions for county-specific probability $p_1, p_2, \ldots, p_N$ (exact formula) and order statistics $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$ (simulated).

Posterior Distribution for % Low Birth Weight



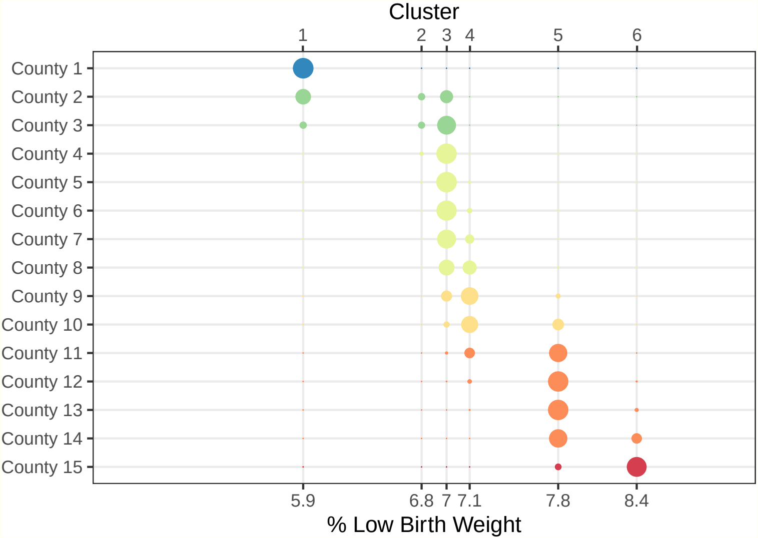
Posterior Distribution for Order Statistics



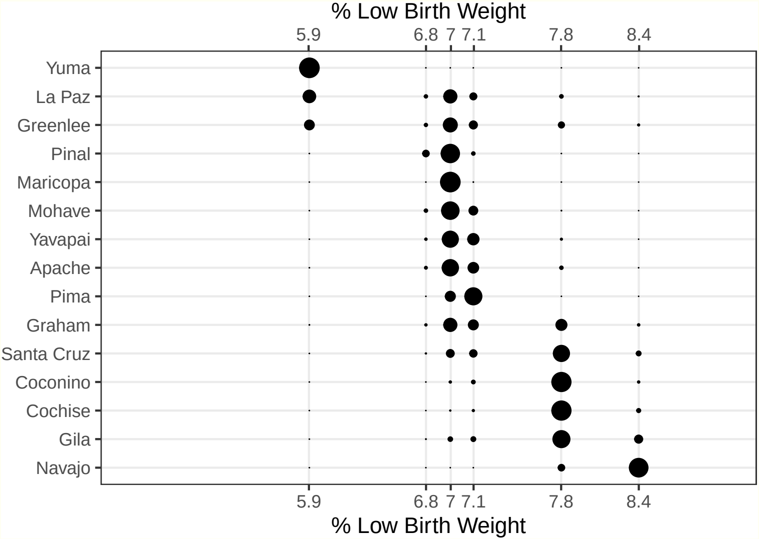
Cluster and Rank Assignments

- Assign cluster to order statistic/rank position using squared error loss on proportion (% LBW) scale.
  - Unconstrained minimization.
  - Compares $p_{(i)}$ with θ_j.
- Assign county to rank using integrated squared error loss on proportion (% LBW) scale.
  - Constrained minimization to avoid duplicate ranks using Hungarian algorithm.
  - Compares p_i with $p_{(j)}$.
- Assign cluster to county based on assigned rank position.
- Similar in spirit to triple-goal estimates from Shen and Louis (1998).

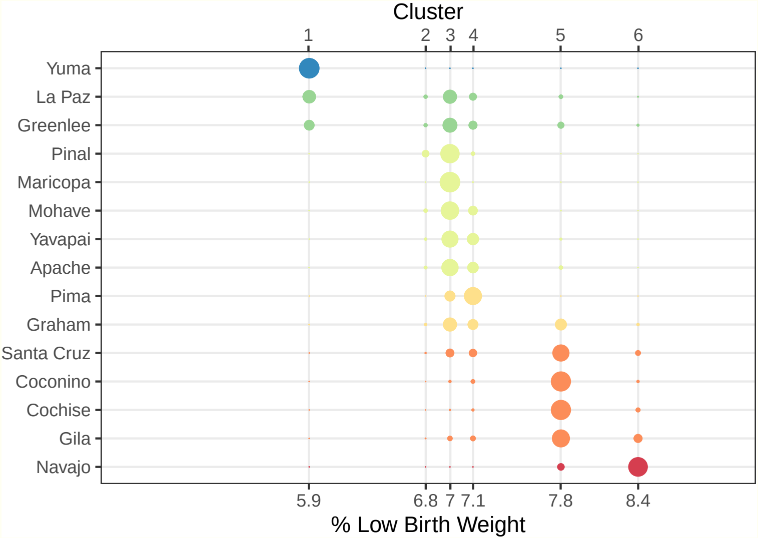
Optimal Cluster Assignments for Order Statistics/Rank Positions



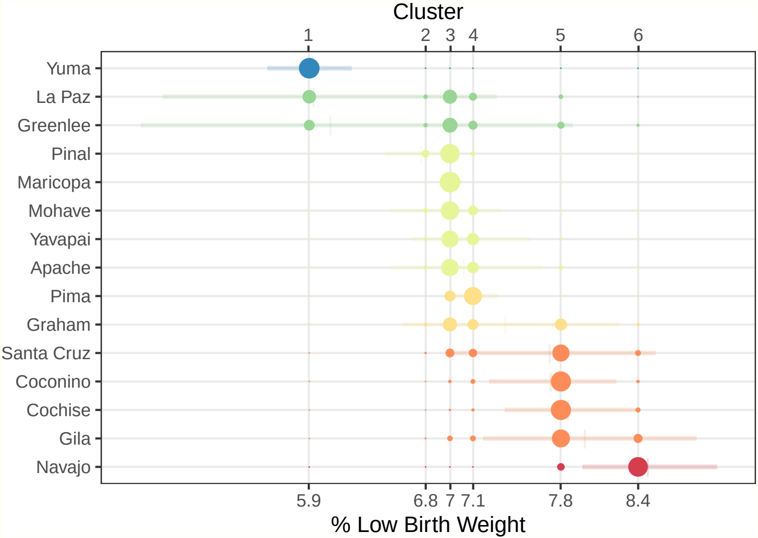
Optimal Rank Assignments for Counties



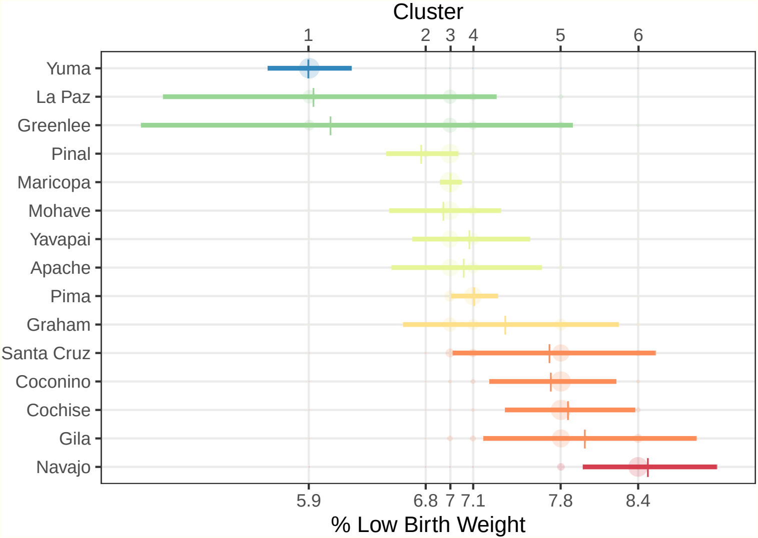
Simultaneous Clustering and Ranking of Counties



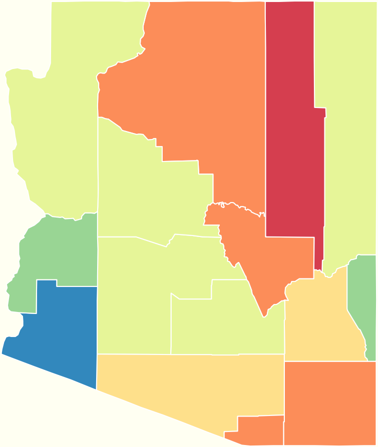
Final Display Emphasizing Cluster Assignments



Final Display Emphasizing Credible Intervals



Map with Cluster Assignments



Discussion

- Ranking (with uncertainty) corresponds to multiple inferential goals.
- High degree of uncertainty regarding most ranks leads to challenges in interpretation and messaging, even for very sophisticated end users.
- Use of (discrete) mixture models allows for simultaneous clustering and ranking of county health indices.
- Adding clustering to optimal ranking greatly facilitates interpretation and messaging, particularly for traditional target audiences for rankings.