The Annals of Applied Statistics
2018, Vol. 12, No. 2, 915–939
https://doi.org/10.1214/18-AOAS1173

CLUSTERING THE PREVALENCE OF PEDIATRIC CHRONIC CONDITIONS IN THE UNITED STATES USING DISTRIBUTED COMPUTING[1]

BY YUCHEN ZHENG AND NICOLETA SERBAN

Georgia Institute of Technology

This research paper presents an approach to clustering the prevalence of chronic conditions among children with public insurance in the United States. The data consist of prevalence estimates at the community level for 25 pediatric chronic conditions. We employ a spatial clustering algorithm to identify clusters of communities with similar chronic condition prevalences. The primary challenge is the computational effort needed to estimate the spatial clustering for all communities in the U.S. To address this challenge, we develop a distributed computing approach to spatial clustering. Overall, we found that the burden of chronic conditions in rural communities tends to be similar but with wide differences in urban communities. This finding suggests similar interventions for managing chronic conditions in rural communities but targeted interventions in urban areas.

Tribute. *Steve was a mentor and a friend, with a big heart. He promoted and supported the new generation of researchers in all ways possible. Steve has inspired many of us in thinking "impact." The submitted paper is a tribute to his imprint he has had on my career as an applied statistician. It is an honor to contribute to this special issue commemorating Steve. This paper is collaborative with a Ph.D. student, Richard Zheng, I have mentored since he was an undergraduate student. I take pride in his achievements as much as Steve took pride in the achievements of the new generation. I am grateful to Steve being part of my career development during a time when I needed the support the most.* Nicoleta Serban

1. Introduction. The Medicaid public insurance program covers more than 36 million children in the United States yearly [Center for Medicare and Medicaid Services (2017a)]. Children covered under this program are from low-income families or/and with severe health disabilities. Disparities in health outcomes for Medicaid-enrolled children are substantive and of great concern nationally [Center for Medicare and Medicaid Services (2017b)]. A first step in addressing such disparities is measurement and evaluation of the health outcomes for this population.

Received November 2017; revised April 2018.

[1]Supported in part by the Coca Cola Professorship and the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R56HL126761.

Key words and phrases. Distributed computing, Medicaid, pediatric chronic conditions, spatial clustering.

Towards this objective, in this paper, we study the burden of chronic conditions in the Medicaid-enrolled child population, which can vary across communities within each state and across states. Characterizing the burden of chronic conditions can help in identifying communities with most need for interventions for improving health outcomes.

We compiled a unique (in-treatment) prevalence data on multiple chronic conditions common in children. The prevalence data are derived from patient-identifiable medical claims from the 2011 Medicaid Analytic eXtract (MAX) files acquired from the Centers for Medicare and Medicaid Services (CMS). The prevalence data are census tract estimates of the percentage of Medicaid-enrolled children diagnosed with a chronic condition, with a total of 64,873 census tracts across the United States, and 25 chronic conditions. The objective in this study is to characterize the burden of chronic conditions in communities by using a clustering or segmentation of the population of children based on the level of prevalence of their chronic conditions. This clustering approach reduces the information content in such large prevalence data into simple data clustering summaries by borrowing information across all census tracts (proxies of communities) and across prevalent childhood chronic conditions. The end point is to create a clustering map of the burden of chronic conditions, which can be used in informed decision making and targeted healthcare interventions.

An important challenge in deriving a clustering for the prevalence data is the presence of strong spatial dependence. Spatial dependence arises because proximal communities will have similar levels of chronic conditions; proximal communities will have similar demographics, social-economics and environmental factors, which can influence the development and the severity of chronic conditions [The World Health Organization (2005), Cockerham, Hamby and Oates (2017)]. These types of spatial effects have been widely modeled in disease mapping. Reviews of methodology for spatial epidemiological data in general may be found in Elliot et al. (2000), Lawson et al. (1999), Wakefield (2006), Waller and Gotway (2004). Most models were developed in spatial smoothing and regression settings. Incorporating spatial dependence is vital in understanding geographical patterns in disease incidence and mapping.

One first attempt for detecting spatial point clusters and hotspots using exploratory methods is the Geographical Analysis Machine (GAM) developed by Openshaw et al. (1987) and later improved by Besag and Newell (1991). Another popular method is to encode the spatial dependence (or other form of dependence) into the feature space. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al. (1996)], Ordering Points To Identify the Clustering Structure (OPTICS) [Kriegel et al. (2011)], and other related variations [Birant and Kut (2007), Wang, Wang and Li (2006)] are the most well-known density based clustering algorithms for correlated data, where distance based measures are used to allocate data to clusters. Carson et al. (2002) transformed raw pixel data in images into a joint color-texture-position feature space, and used the

Expectation Maximization for Gaussian Mixtures to construct a small set of image regions that are coherent in color and texture. Jiang and Serban (2012) introduced a model-based method for clustering random time-varying functions under spatial interdependence. Green and Richardson (2002) extended the hidden Markov models to the spatial domain with a finite-mixture model for Poisson rates, where the mixture component follows a spatially correlated process, the Potts model. This model is flexible in terms of assumptions but may be cumbersome to implement on extremely large data.

In this paper, we apply a general spatial clustering approach to cluster high-dimensional data assuming spatial dependence in the observed response data, while not restricting the spatial effect to be the same across the spatial domain. The main challenge in implementing this spatial clustering method to the nationwide prevalence data is the computational effort; the method is not scalable to a large number of spatially-dependent responses. To address this challenge, in this paper, we develop a distributed-computing spatial clustering method.

Distributed computing has become a much needed alternative modeling approach in many research domains, particularly in statistical learning, due to the advent of large size and complex datasets. The size of the data collected are sometimes too large to be stored at a central location, and the level of computation needed for statistical learning may not scale up to the data dimensionality. In addition, data in some cases are naturally collected in a decentralized fashion at a local level, and communication between local servers and a central machine is expensive and wasteful. The data are usually assumed to be independent to alleviate the computational burden, since data in each node can be calculated separately in a distributed fashion. Chu et al. (2007) indicated that algorithms applied to independent data are easily parallelizable on multicore computers, in a Map Reduce framework. Wolfe, Haghighi and Klein (2008) provided a general framework for distributing expectation-maximization algorithms under independence of the response data in which not only is the computation distributed, but the storage of parameters and expected sufficient statistics is also fully distributed. However, when strong are present among the response variables, the independence assumption is therefore violated, which can produce misleading inferences.

One contribution of this paper is thus the derivation and development of a distributed computing solution to the estimation of the clustering model under spatially interdependence. The estimation approach requires innovation in the decomposition of the log-likelihood function in a way that its maximization can be distributed across multiple computing cores. A second contribution in this paper is that we not only derive the distributed estimation approach but also implement it within the applied problem, specifically, identifying geographic clusters of the burden of pediatric chronic conditions, where each cluster can be characterized by different prevalence levels of chronic conditions and by different groups of the conditions.

In the following section, we present the approach for deriving the prevalence data. In the section that follows, we will continue with the introduction of the general form of the expectation-maximization (EM) algorithm in solving Gaussian Mixture Models, then we relax the independence assumption by re-formulating the E-step and M-step, and proposing an efficient parallel EM Algorithm. We apply the proposed algorithm to deriving the clustering map of the chronic disease prevalence among children enrolled in Medicaid using the large-scale prevalence data. We conclude with a discussion on the implications of the clustering map towards targeted healthcare interventions.

2. Chronic condition prevalence for the Medicaid-enrolled children.

2.1. *Data source*. We analyze the patient-level claims from the 2011 Medicaid Analytic eXtract (MAX) files obtained from the Centers for Medicare and Medicaid Services (CMS). The research in this study was approved by CMS (Data Use Agreement #23621) and by the Institutional Review Board of Georgia Tech (protocol #H11287). All data derived from the MAX files meet a minimum cell size of 11 in terms of number of patients according to the Data Use Agreement with CMS. We focus on children age 0 to 17.

2.2. *Prevalence estimation*. Prevalence estimation is an important research topic in health services research; prevalence estimates can be used for targeted interventions to improve health outcomes for a specific condition. Approaches for prevalence estimation range from micro-simulation models [Davila-Payan et al. (2015), Cameron et al. (2015), Kopec et al. (2010)] to geostatistical models [Diggle and Giorgi (2016)]. Such models can be applied when information on the population diagnosed and/or treated for a condition is sparse, in other words, for only a small subset of the population. In this study, information on children treated for a specific condition is available at the individual level across the entire Medicaid population and thus our prevalence estimates are derived as population rates of the Medicaid population treated for a specific condition, called treated prevalence.

We derive the prevalence estimates using the 3M Clinical Risk Grouping software [Neff et al. (2002)]. Episode Diagnostic Categories (EDCs) are derived for each child enrolled in Medicaid using the child's diagnosis codes, procedure codes, and national drug codes (NDCs) found in the recorded medical claims in the MAX files. EDCs are used to determine a patient's Primary Chronic Disease, which is the most significant chronic disease actively being treated, and its severity for each organ system.

We consider EDCs for the following 25 conditions: Acute Bronchitis and Bronchiolitis, Acute Respiratory Diagnoses—Moderate, Acute Skin Diagnoses, Acute Stress and Anxiety, Attention Deficit Hyperactivity Disorder (ADHD), Allergies, Asthma, Autism, Bipolar, Chronic Mental Health, Chronic Stress, Conduct and

Behavior, Dental Diagnoses, Depression, Depressive and Other Psychoses, Developmental Language Disorder, Developmental Speech and Learning, Diabetes2, Epilepsy and Epilepsy Complex, Major Mental Health, Psoriasis, Schizophrenia, Social Problems, Upper Respiratory Infections. These conditions were selected due to their high prevalence among children enrolled in the Medicaid program. According to the data use agreement with CMS, we cannot disclose any information when the cohort population is less than 11 patients, thus lower prevalence conditions cannot be captured in our analysis.

For each condition or EDC, we obtained the population of Medicaid-enrolled children with the condition along with the number of enrollment months of these children within each zip code and county. We derived the prevalences of conditions by dividing the total number of member months of patients treated for a given condition by the total number of member months of all children on Medicaid for each county and zip code area. We further estimated the census tract prevalence using the zip code and county estimates along with geographic information of the boundaries of the different geographic devisions (county, zip code, census tracts) and the information on the population count across the geographic divisions. For cells with less than 11 patients, we used the mean estimation at the county level, along with a generated beta noise term based on zip code level and state level estimations.

Overall, we have a total of 64,873 census tracts for which we have obtained prevalence estimates for the 25 conditions. The census tracts cover the entire United States excluding Colorado and Idaho due to data unavailability. Details on the prevalence of the EDCs and their denomination as provided by the 3M Clinical Risk Grouping software along with details on the derivation of the census tract prevalence estimates using the MAX claims data are provided in the Online Supplemental Material A [Zheng and Serban (2018)].

2.3. *Exploratory analysis*. The number of Medicaid enrolled children in each census tract varies from 0 to 10,319, with an average of 401. Out of the total 1.62 million data cells or prevalence estimates across all the conditions considered in this study, 58 % of the cells have less than 11 patients. Most of instances correspond to rare conditions and rural areas, where the estimates at the census tract level are very similar to those at the county level. On average, patients are enrolled in Medicaid for 10 months within the year, with very low state to state variation. The prevalence across the 25 chronic conditions varies widely, with Epilepsy as the least prevalent condition (ranging from 0.2% to 0.8%) and upper respiratory infections as the most prevalent condition (ranging from 12.6% to 61.3%). Figure 1 shows the histogram and heat map of the prevalence for upper respiratory infections and major mental health in the state of Georgia. The distribution for the upper respiratory infections is approximately uni-modal but for the major mental health condition it is multi-modal. The heat map shows the presence of spatial dependencies, where nearby geographical locations tend to have similar level of

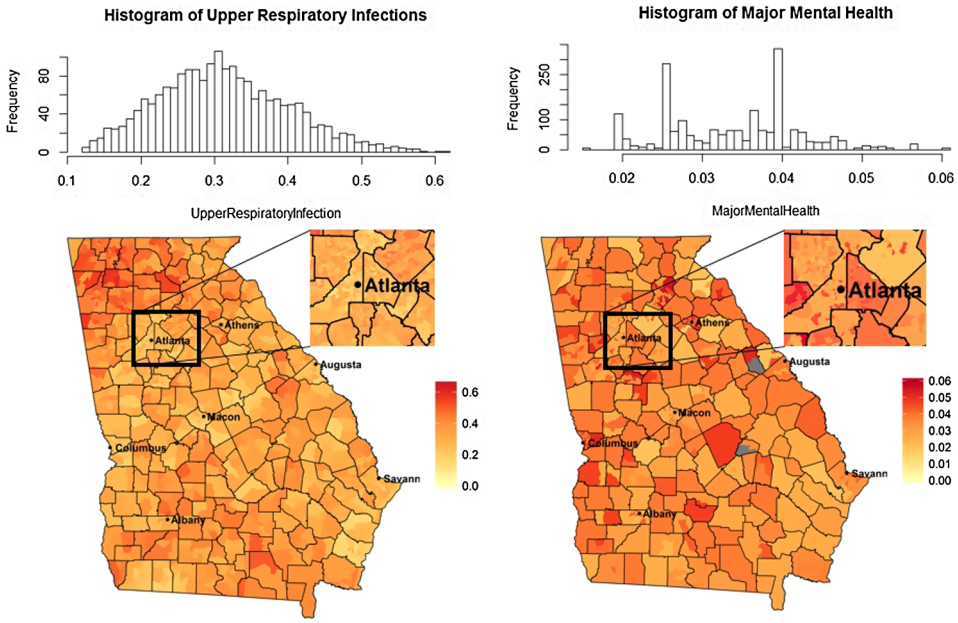


FIG. 1. *Histogram and heat map of prevalence for upper respiratory infections and major mental health in the state of Georgia.*

prevalence. The strength of spatial dependence however differs from region to region and by condition. In urban locations, such as the Atlanta metropolitan area, the census tracts tend to be much smaller and denser. The prevalences are more similar in these areas than in rural locations where census tracts are larger and further away. For more prevalent conditions, such as the upper respiratory infections, the prevalence across the map is smoother than for less prevalent conditions, thus differences between nearby census tracts are relatively small.

3. Statistical modeling using distributed computing.

3.1. *Nominal EM algorithm for Gaussian mixture models.* The Expectation Maximization (EM) Algorithm is a class of iterative methods for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobservable or latent variables [Dempster, Laird and Rubin (1977)]. Each EM iteration alternates between performing an expectation E-step, which updates the expectation of the log-likelihood function evaluated using the current estimates for the parameters, and a maximization M-step, which estimates parameters maximizing the expected log-likelihood given the input from the E-step of the previous iteration. The EM is frequently used for modeling mixtures of distributions, where data are commonly assumed to be generated from mixtures of multivariate

Gaussian distributions (GMM) assuming unknown number of mixtures and unknown mixture weights. Modeling the mixture of Gaussian distributions can also be viewed as a data clustering method [Fraley and Raftery (1998, 2002)].

The observed response data are $Y_1, Y_2, \ldots, Y_N$ where Y_i is a p-dimensional vector of measurements, in this paper, the prevalence estimates for the 25 pediatric chronic conditions. N is the number of responses. We further assume that the distribution of Y_i is a realization from a finite Gaussian mixture model with C components:

$$p(y|\Theta) = \sum_{k=1}^{C} w_k p_k(y|\theta_k).$$

- $p_k(y|\theta_k)$ is the kth mixture component where this mixture is identified by the parameter θ_k. For mixtures of Gaussians, $\theta_k = \{\mu_k, \Sigma_k\}$ are the mean and the covariance specifying the kth Gaussian.
- w_k are the mixture weights, representing the probability that a randomly selected Y was generated by component k.

The unobserved data are the latent variables $Z_1, Z_2, \ldots, Z_N$ where Z_i has a C-dimensional multinomial distribution specifying the cluster membership of Y_i. Thus given $Z_{ik} = 1$ and $Z_{ic} = 0$ for $c \neq k$ where k takes values in $\{1, 2, \ldots, C\}$, Y_i has a distribution with the density function $p_k(y|\theta_k)$.

The EM algorithm is an iterative algorithm that starts from some initial estimates of $\Theta = (\theta_1, \ldots, \theta_C)$ and of $\mathbf{w} = (w_1, \ldots, w_C)$, and then proceeds to iteratively update Θ and $\mathbf{w}$ until convergence. Each iteration consists of an E-step at which we update the mixture weights and impute the cluster memberships and an M-step at which we estimate Θ given the imputed cluster membership.

For classic mixtures of multivariate Gaussian distributions, the responses to be clustered are generally assumed independent and hence the EM algorithm can be distributed easily across multiple computing nodes [Wolfe, Haghighi and Klein (2008)]. However, in this paper, we assume the response data are spatially interdependent.

3.2. *Correlation structure*. The proposed spatial EM algorithm extends the nominal EM algorithm (under independence assumption) by incorporating spatially correlated random errors. In our application, the spatial correlation structure is a function of the proximity between pairs of census tract centroids, assumed to be defined by an exponential correlation function, which is widely used in spatial statistics and geostatistics [Ripley (2005), Cressie (2015)].

The most granular information we have for each patient is the residential zip code, not the exact address. Therefore, we are treating the prevalence estimates as point masses at the centroids of each area unit, instead of a point process across the geographic area. Alternatively, we can use a power law on the order or proximity

of the neighborhoods, such that a small neighboring region would be attributed a stronger link than a large neighbor with its centroid further apart [Meyer and Held (2014)]. However, as census tracts are defined based on settlement density, it is desirable that dense areas, mostly consisting of small neighboring regions, have stronger spatial dependencies than rural areas, where large neighbors' centroids are further apart.

Furthermore, instead of considering spatial correlation between every possible pair of locations, which deems to be intractable, we enforce a neighborhood structure—that is, for each location, we only consider the closest $M-1$ number of neighbors, resulting in a neighborhood of size M. M can be fixed, or can vary for different response i. For example, we can assign a larger M to urban locations than to rural locations, since the spatial effect is expected to be stronger. An alternative is to sparsify the spatial correlation matrix by setting a hard threshold. In addition, the correlation is assumed to decay exponentially as the distance between two locations increases. Other correlation structures can be considered but for simplicity of the interpretation and implementation, we use classic approaches to specify the correlation structure.

The neighborhood criterion is similar to spatial tampering and the Gaussian Markov random fields (GMRF) advocated in Furrer, Genton and Nychka (2006), Rue and Held (2005), Rue, Martino and Chopin (2009). Under GMRF modeling, the conditional distribution of a latent GMRF parameter depends only on the neighbors [Rue and Held (2005), Rue, Martino and Chopin (2009)]. GMRF can efficiently model most of the spatial covariance functions [Rue and Tjelmeland (2002)]. While this method applies effectively to moderate size dataset, an application with a large number of spatial points, for example, 64,873 locations as for the U.S. prevalence data, can be computationally challenging.

We also assume the features are uncorrelated. This can be achieved by assuming independence on the feature space; to achieve independence between the features, the feature set can be preprocessed into uncorrelated orthogonal basis set, for example, using the Principal Component Analysis (PCA), which is common practice [Ding and He (2004)]. In the next section, we will see that the assumption of independence on the feature space significantly reduces the computational complexity and makes the algorithm parallelizable.

3.3. *Expectation step*. In the E-step, we evaluate the expected cluster membership probability for each response based on the parameters estimated in the M-step. In the derivations below, since the parameters specifying the mixtures $\Theta = (\theta_1, \ldots, \theta_C)$ are assumed fixed in the E-step (provided by the estimates derived in the M-step), we drop the conditioning on the set of parameters Θ for ease of illustration.

Conditional model: The model for the ith response or measurement is:

$$Y_i|(Z_{ik}=1) = \mu + \mu_k + s_i + e_i,$$

where Z_i is the latent variable (cluster membership) for the response i, μ is the global mean, μ_k is the cluster mean for cluster k, with $\sum_{k=1}^{C} \mu_k = 0$, where C is the total number of clusters, the spatial random effect s_i, and the independent error term e_i. Y denotes the vector of all responses and Y_i denotes the ith response; the kth membership probability for the response i is denoted as w_{ik}.

Under interdependence among the responses, the estimation of w_{ik} involves more complex computations:

$$\begin{aligned} w_{ik} &= E[Z_{ik} = 1|Y] = P(Z_{ik} = 1|Y) \approx P(Z_{ik} = 1|Y_i, Y_{N(i)}) \\ &= \frac{P(Z_{ik} = 1) f(Y_i, Y_{N(i)}|Z_{ik} = 1)}{\sum_{c=1}^{C} P(Z_{ic} = 1) f(Y_i, Y_{N(i)}|Z_{ic} = 1)}, \end{aligned}$$

where $N(i)$ denotes the set of indexes of the responses that are neighbors of the ith response. In this case, the dependence structure among the responses is encoded in the parameter estimations of the latent classes, which we will discuss in more detail in the M-step. Therefore, the expected membership probability for sample i depends on its neighbors, that is, the probability density function $f(Y_i, Y_{N(i)}|Z_{ik} = 1)$ needs to be calculated jointly. In what follows, we will focus on how to estimate this joint probability efficiently. For ease of presentation, we will use μ to represent $\mu + \mu_k$.

Denote the mth neighbor of response i as $Y_{N(i,m)}$, where $m = 1, 2, \ldots, M - 1$. Calculating the joint probability of this neighborhood can be computationally intense and not scalable, since only $Z_{ik} = 1$ is given, but not its neighbor's cluster memberships. The joint density is calculated as

$$\begin{aligned} &f(Y_i, Y_{N(i)}|Z_{ik} = 1) \\ &\quad = f(Y_i, Y_{N(i,1)}, \ldots, Y_{N(i,M-1)}|Z_{ik} = 1) \\ &\quad = \sum_{k_{N(i,1)}=1}^{C} f(Y_i, Y_{N(i,1)}, \ldots, Y_{N(i,M-1)}|Z_{ik} = 1, Z_{N(i,1)k_{N(i,1)}} = 1) \\ &\qquad \times P(Z_{N(i,1)k_{N(i,1)}} = 1). \end{aligned}$$

For ease of display, we rewrite $Z_{N(i,m)k_{N(i,m)}}$ as $Z_{i,m,k}$ and denote the mixture weight $P(Z_{N(i,m)k_{N(i,m)}} = 1)$ with $w_{i,m,k}$. We then expand the joint density function for all responses in the neighborhood:

$$\begin{aligned} &f(Y_i, Y_{N(i)}|Z_{ik} = 1) \\ &\quad = \sum_{k_{N(i,M-1)=1}}^{C} w_{i,1,k} \cdots \sum_{k_{N(i,M-1)=1}}^{C} w_{i,M-1,k} \\ &\qquad \times f(Y_i, Y_{N(i,1)}, \ldots, Y_{N(i,M-1)}|Z_{ik} = 1, Z_{i,1,k} = 1, \ldots, Z_{i,M-1,k} = 1). \end{aligned}$$

In each summation, the joint density function are conditioned on the membership of the response i and its neighbors. However, the amount of computation doesn't

scale with the increasing size of the neighborhood, as the joint density function needs to be expanded in the neighborhood of each response and in each cluster, which results in C^M joint density estimations. One alternative is to perform a hard clustering on $w_{i,m,k}$ $\forall m = 1, 2, \ldots, M-1$ such that $w_{i,m,k^*} = 1$ for $k^*_{N(i,m)}$ which maximizes over all $k_{N(i,m)}$ and $w_{i,m,k} = 0$ for all other $k_{N(i,m)}$. Thus we have the following approximation:

$$f(Y_i, Y_{N(i)}|Z_{ik} = 1) \approx f(Y_i, Y_{N(i,1)}, \ldots, Y_{N(i,M-1)}|Z_{ik} = 1, Z_{i,1,k^*} = 1, \ldots, Z_{i,M-1,k^*} = 1).$$

An interpretation of the approximation above is as follows: the memberships of the neighboring responses are assumed to be fixed based on the membership matrix calculated in the previous M-step, and only the membership of response i varies. This heuristic is similar to successive methods such as backfitting and Gauss–Seidel. We denote the approximation of $f(Y_i, Y_{N(i)}|Z_{ik} = 1)$ as $f(\mathbb{Y}_i|Z_{ik} = 1)$, where $\mathbb{Y}_i$ is a M-by-p matrix. Denote the M-by-M matrix S_i as the spatial covariance matrix for the neighborhood around i, and the p-by-p matrix Σ_i as the covariance matrix for the random error ε_i for response $i, i = 1, 2, \ldots, M$, where Σ_i is a diagonal matrix with the diagonal provided by $[\sigma^2_{i1}, \ldots, \sigma^2_{ip}]$. The neighborhood $\mathbb{Y}_i$ thus follows a matrix normal distribution, whose variance is the Kronecker product of the S_i and the corresponding Σ_i for each response in the neighborhood.

We further decompose

$$\mathbb{Y}_i|(Z_{ik} = 1) = \mu + S_i^{\frac{1}{2}} A,$$

where each row l of A is independent, $A_{l\cdot} \sim N(\mu_l, \Sigma_l)$. We then have

$$f(\mathbb{Y}_i|Z_{ik} = 1) = \prod_{l=1}^{M} \frac{1}{\sqrt{\det(S_i)}} f(A_{l\cdot}|\mu = 0, \Sigma = \Sigma_l).$$

$f(\mathbb{Y}_i|Z_{ik} = 1)$ $\forall i = 1, \ldots, n, k = 1, \ldots, C$ can be computed using distributed computing, for each i separately or for groups of i's. This can further be used in estimating the cluster weights w_{ik}, concluding the E-Step.

3.4. *Maximization step*. In the Spatial EM algorithm, the parameter set Θ contains of $(\mu_k, \Sigma_k) = \theta_k$, for all $k = 1, \ldots, C$. It is however computationally challenging to obtain the MLEs for these parameters when there is dependence in the sample data. Alternatively, we can use the Maximum Pseudo-likelihood Estimation [Besag (1986), Liu and Ihler (2012)]

$$\max E\big[l(\Theta; Y)\big] = \max \sum_{i=1}^{n} \sum_{k=1}^{C} w_{ik} \log f(Y_i|Y_{N(i)}, Z_{ik} = 1).$$

We use a similar technique as used in the computation in the E-step to account for the dependence structure. For each response i belonging to cluster k, we have

$$X_i = (Y_i - \mu_k)\Sigma_k^{-\frac{1}{2}},$$
$$X_{N(i)} = \sum_{l=1}^{M-1} e_l \sum_{k=1}^{C} w_{ik}(Y_{i_l} - \tilde{\mu}_k)\tilde{\Sigma}_k^{-\frac{1}{2}},$$

where e_l is a vector of length $M-1$, where the lth element is 1 with all other values being zero, $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$ are parameters estimated in the previous iteration of the EM algorithm. We then have the following:

$$\max \sum_{k=1}^{C}\sum_{i=1}^{n} w_{ik} \log f(Y_i|Y_{N(i)}, Z_{ik} = 1)$$
$$= \max \sum_{k=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ik} \log\Big(\frac{1}{\sigma_{kj}} f\big(X_{ij}|[X_{N(i)}]_{\cdot j}, Z_{ik} = 1\big)\Big).$$

Expand the spatial correlation matrix S_i as

$$\begin{bmatrix} S_{i_{11}}, S_{i_{12}} \\ S_{i_{21}}, S_{i_{22}} \end{bmatrix},$$

where $S_{i_{11}}$ is 1, the vector $S_{i_{12}}$ of length $M-1$ is the correlation between response i and its neighbors, $S_{i_{21}}$ is the correlation between response i's neighbors and itself, and the $(M-1)$-by-$(M-1)$ matrix $S_{i_{22}}$ is the correlation matrix among response i's neighbors. We then have $X_{ij}|[X_{N(i)}]_{\cdot j}, Z_{ik} = 1 \sim N(\bar{\mu}_{ij}, \bar{\Sigma}_{ij})$, where

$$\bar{\mu}_{ij} = S_{i_{12}} S_{i_{22}}^{-1} [X_{N(i)}]_{\cdot j},$$
$$\bar{\Sigma}_{ij} = 1 - S_{i_{12}} S_{i_{22}}^{-1} S_{i_{21}}.$$

Therefore, we have

$$\max \sum_{k=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ik} \log\Big(\frac{1}{\sigma_{kj}} f\big(X_{ij}|[X_{N(i)}]_{\cdot j}, Z_{ik} = 1\big)\Big)$$
$$= \max \sum_{k=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ik}\Big(-\log(\sigma_{kj}) - \frac{(\frac{Y_{ij}-\mu_{kj}}{\sigma_{kj}} - \bar{\mu}_{ij})^2}{2\bar{\Sigma}_{ij}}\Big) = G(Y).$$

Setting the first derivatives of the pseudo-likelihood to zero, we get the following estimation:

$$\hat{\mu}_{kj}^{\text{mple}} = \frac{\sum_{i=1}^{n} \frac{w_{ik}}{\bar{\Sigma}_{ij}}(Y_{ij} - \bar{\mu}_{ij}\sigma_{kj})}{\sum_{i=1}^{n} \frac{w_{ik}}{\bar{\Sigma}_{ij}}},$$

and $\sigma_{kj}^{\text{mple}}$ is the positive root of the following quadratic equation:

$$\sum_{i=1}^{n} w_{ik}\sigma_{kj}^2 - \sum_{i=1}^{n} \frac{w_{ik}}{\bar{\Sigma}_{ij}}(Y_{ij} - \mu_{kj})\bar{\mu}_{ij}\sigma_{kj} + \sum_{i=1}^{n} \frac{w_{ik}}{\bar{\Sigma}_{ij}}(Y_{ij} - \mu_{kj})^2 = 0.$$

We initialize $\hat{\mu}_{kj}^{\text{mple}}$ and $\hat{\sigma}_{kj}^{\text{mple}}$ with the estimates from the previous iteration, and solve the equations iteratively.

If the correlation among samples is minimal, the spatial correlation matrix S_i $\forall i = 1, \ldots, n$ becomes diagonal, with $\bar{\mu}_{ij} = 0$ and $\bar{\Sigma}_{ij} = 1$; therefore, we have

$$\mu_{kj} = \frac{\sum_{i=1}^{n} w_{ik} Y_{ij}}{\sum_{i=1}^{n} w_{ik}},$$

$$\sigma_{kj}^2 = \frac{\sum_{i=1}^{n} w_{ik}(Y_{ij} - \mu_{kj})^2}{\sum_{i=1}^{n} w_{ik}},$$

which coincides with the estimation based on the nominal EM algorithm with independent responses. The proposed method is therefore a generalization of the nominal EM algorithm.

3.5. *Model selection.* Similar to most of the model-based clustering algorithms, the number of clusters needs to be finely tuned to obtain a set of meaningful clusters. Common variable selection methods such as the Akaike information criterion (AIC), and Bayesian information criterion (BIC) have been employed for estimating the number of clusters [Fraley and Raftery (2002)]. In our application, we chose to use BIC as a starting point to identify an inflection point (where BIC starts to tip-off) to identify an initial number of clusters, then merge similar clusters in a more empirical way, resulting in the most sensible clustering of the prevalence responses.

3.6. *Distributed implementation.* There are two important challenges of the distributed computing implementation of the clustering algorithm. The first challenge is the storage and retrieval of the data throughout the computation process. The size of the data can be too large to be stored and computed using only one computing node. Thus in our implementation, we partition the data onto multiple storage nodes and execute the algorithm on each subset of data in a Map Reduce fashion. In more complex cases where data are naturally collected and stored in a decentralized approach, communicating all the data onto one centralized location can be very expensive. More sophisticated design of distributed storage topologies are required, as outlined in Wolfe, Haghighi and Klein (2008).

A second challenge is in the distributed computation itself for making the EM algorithm more scalable. However, without the independence assumption, the rows of the data matrix are coupled, thus the likelihood function cannot be decomposed in a way that allows distribution of the computation of its maximization. To address

this challenge, we decompose and transform the correlation structure, allowing for the implementation of both the distributed data storage/retrieval and the parallel computation. In the E Step, the estimation of ith response's membership probability in cluster k only requires information from its immediate neighbors. The expected sufficient statistics for each observed response can be computed independently in blocks given a current estimate of the parameters. In the M Step, the data in each neighborhood are transformed assuming the correlation structure. The parameter estimation can then be written in closed form summation, which can be efficiently implemented in a Map Reduce fashion in parallel.

The algorithm was implemented in Julia, a high-performance dynamic programming language for numerical and distributed computing [Bezanson et al. (2017)]. The implementation will be made available as a supplemental online material.

4. Results. In this section, we present the results for the clustering approach to study the burden of chronic conditions for Medicaid-enrolled children in the United States. We first compare the clustering results under the Nominal EM (under the independence assumption) and the Spatial EM algorithm, to motivate the need of the additional computational effort of modeling the spatial structure in the chronic condition prevalence data. We then show the superior performance in runtime utilizing distributed computing versus sequential computing. Last, we provide results on the overall clustering throughout the United States with inference on differences of the chronic condition burden across states and urbanicity levels.

4.1. *Nominal vs. spatial clustering.* We study and compare the clusters of census tracts under the nominal EM algorithm and the Spatial EM algorithm. Both algorithms use a randomized membership initiation scheme; that is, each census tract was randomly assigned to a cluster and an initial estimation of mean and covariance were calculated thereafter. In this section, for illustration purposes, we choose the number of clusters to be three since it produces the most meaningful division of census tracts among other selections for the number of clusters. Details on the model selection can be found in Section 4.3.

Although most of the health conditions have very weak correlation, between -0.1 and 0.1, there still exists some moderate correlation, especially in the group of mental health conditions. Therefore, the features are first transformed into orthogonal principal components using principal component analysis. By using PCA, we are assuming that the feature correlation structures are approximately the same among different clusters. To test the validity of this assumption, we calculate the sample correlation matrices among the 25 conditions in their original scale for the entire population, and for each of the clusters. The point-wise 95% confidence interval for the difference between the population correlation matrix and the correlation matrices for cluster 1, 2, and 3 are $[-0.01, -0.003]$, $[-0.008, 0.009]$, and $[-0.008, 0.003]$ respectively. The differences are minimal, which justifies using PCA in our study.

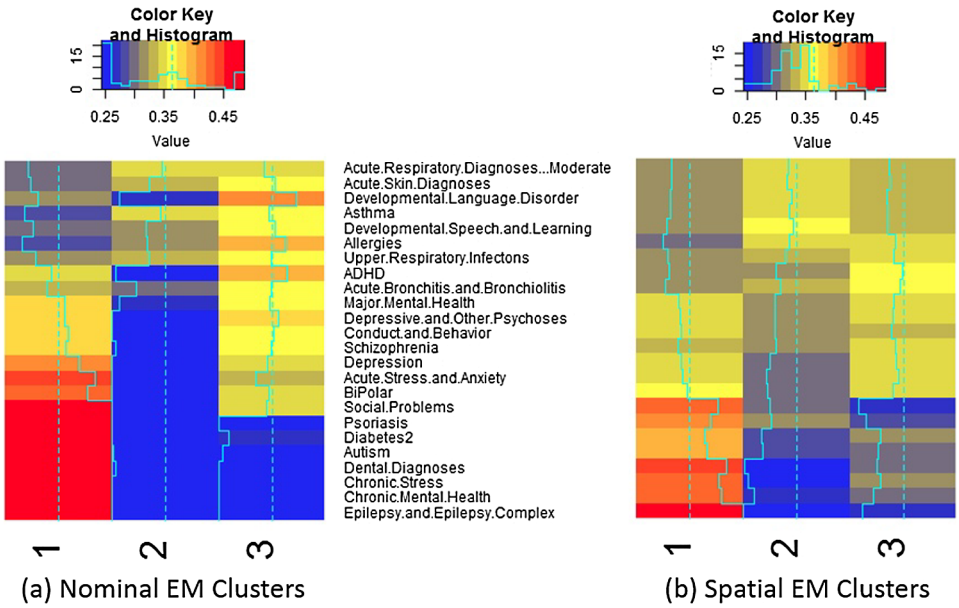


FIG. 2. *Make up of the prevalence in each cluster under* (a) *nominal EM Algorithm and* (b) *spatial EM Algorithm. The values are normalized so that each row sums to* 1.

Figure 2 shows the prevalence for all the chronic conditions for each of the three clusters, contrasting the results based on the two clustering approaches. To better compare the composition of conditions across clusters, each row of the plot was normalized to sum to one. Under the Nominal EM algorithm, we see a clear separation of conditions in each cluster. Cluster 1 consists of 11,512 census tracts (17.7%), predominantly with chronic and moderate mental health diseases, along with some acute and major conditions. Cluster 2 consists of 25,473 census tracts (39.3%), where the prevalences of mental diseases are mostly low, with moderate prevalence in some respiratory and skin related diseases. Cluster 3 consists of 27,888 census tracts (43%), where moderate prevalences for all conditions exist, except for some severe chronic mental conditions, Diabetes, and Dental diseases. The clear separation is as expected, since the nominal EM algorithm clusters the census tracts solely based on the absolute distribution of the prevalence of each condition.

Under the Spatial EM algorithm, the conditions are more blended in each cluster. Cluster 1 is very similar to the first cluster under the nominal EM algorithm in composition, with 23,896 census tracts (36.8%). Cluster 2 consists of 11,964 (18.4%) census tracts, where all of the respiratory related conditions, such as acute/moderate respiratory diagnoses, Asthma, Allergies, upper respiratory infections, Bronchiolitis among others are moderately prevalent. Cluster 3 consists of 29,013 census tracts (44.7%), with less respiratory and skin conditions, but more

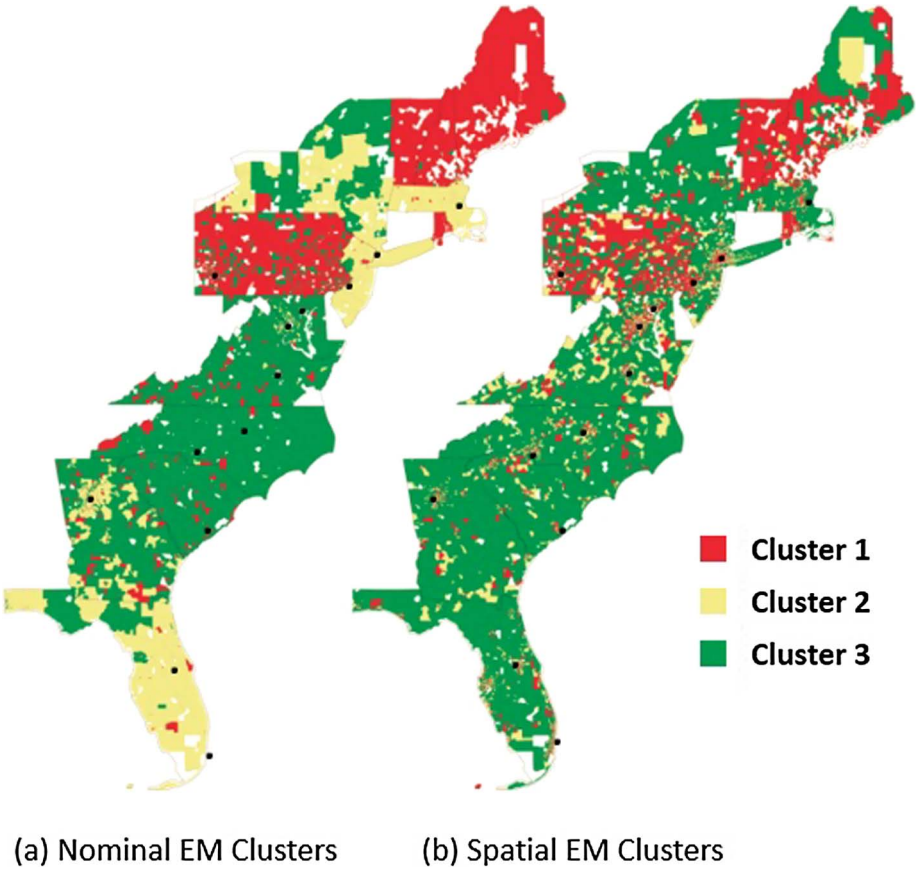


FIG. 3. *Maps of the census tracts located in the east coast states of the United States, color coded by the cluster membership under* (a) *nominal EM Algorithm and* (b) *spatial EM Algorithm. Each black dot represents a major city.*

mental health conditions, in contrast to the Cluster 3 from the nominal EM algorithm.

Figure 3 shows the map of census tracts located in the states near the east coast of the United States. The census tracts are color coded based on the cluster membership under the two EM algorithms. Figure 4 takes a closer look at the areas near major cities, the coast line near Miami, New York City area, and Washington D.C.–Baltimore area. Generally, the locations of different clusters are similar, with rural areas consisting primarily of census tracts in Cluster 3, representing a larger portion of acute and major mental health issues. Pennsylvania, Vermont, New Hampshire, and Maine exhibit considerably less prevalence for acute and major mental conditions.

The biggest difference between the two maps are the areas around major cities, labeled as black dots. The clusters generated from the nominal EM are homo-

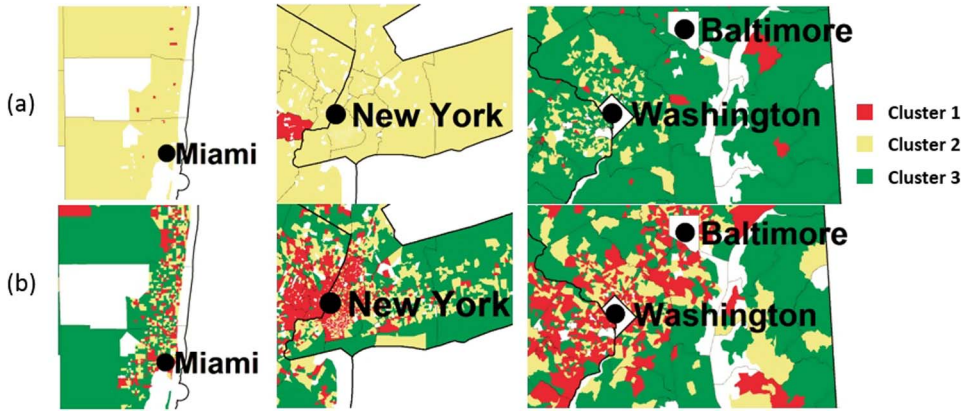


FIG. 4. *Zoomed-in maps of the census tracts close to major cities, color coded by the cluster membership under* (a) *nominal EM Algorithm and* (b) *spatial EM Algorithm.*

geneous across the map. Census tracts in the same area tend to be from the same cluster, such as the coast line near Miami, and around New York City. The nominal EM algorithm failed to capture the heterogeneity in small areas, especially where the population is dense and diverse. On the contrary, in addition to the absolute distribution of the features, the Spatial EM algorithm accounts for the magnitude of prevalence values on the relative scale by modeling the spatial correlation in small areas. Therefore, it discovers relative differences on the spatial domain.

4.2. *Distributed computation.* The computation of the Spatial EM algorithm is significantly more complex than the nominal EM algorithm and requires more time and computing resources to execute. In this section, we illustrate how distributed computation can help alleviate the computational burden. In order to compare the computational results under the sequential and parallel implementations, we fix the number of iterations to be 100. Figure 5 shows the computational results with different number of computing cores (Intel Core Haswell Processors). The algorithm was written in Julia, and executed on a Linux server with X86-64 bit architecture.

The job execution required a total of 36.3 GB in memory allocation, thus infeasible to store and retrieve the data on a single machine/computing node—even the implementation using serial computation (one computing core) had to utilize a distributed storage framework. Running the algorithm using one computing core took more than 11 hours. This number can easily skyrocket to weeks as additional runs are required for sensitivity analysis, parameter tuning (e.g., number of clusters, size of neighborhood), and statistical inference, for example. Running the algorithm in a parallel fashion greatly reduces the computational time. With 10 computing cores, the run time was reduced to 1.8 hours, with a 6.3 times speed

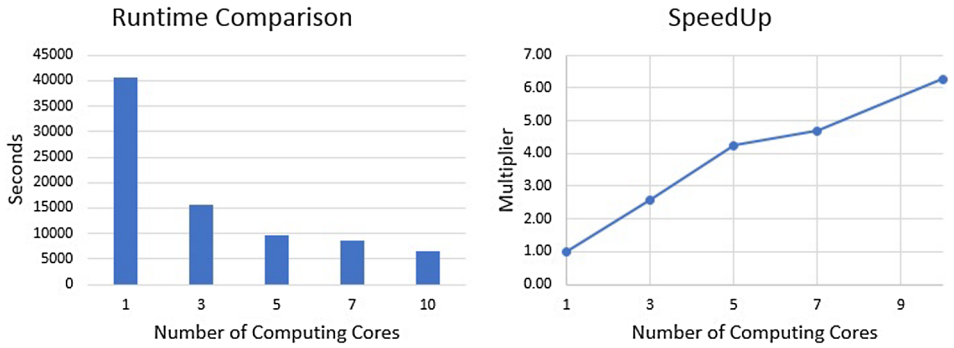
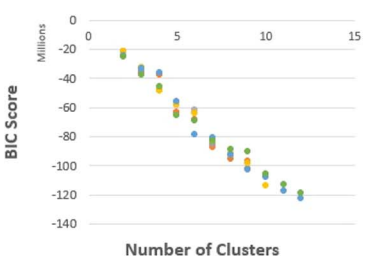


FIG. 5. *Runtime comparison in seconds and speed up with varying number of computing cores.*

up. We note that the speed up is not exactly proportional to the number of computing cores. In fact, the run time improvement is most significant with the first few added cores, and gradually decays as the number of cores further increases. This is commonly known as the Amdahl's Law, where the potential program speedup is defined by the fraction of code that can be parallelized [Amdahl (1967)]. In addition, other architectural and synchronization constraints such as memory-CPU bus bandwidth, communication bandwidth, load balancing and memory locks play key roles in coordinating the distributed execution and become more complex as the number of cores increases.

4.3. *Model selection.* We use the BIC score as the model selection criteria to identify the number of clusters. In addition, since the clustering results tend to vary with different initializations, we run the algorithm five times for each setting to study the sensitivity of the imputed cluster membership and number of clusters to initialization. Part (a) of Figure 6 shows that the BIC curve decreases with the number of clusters ranging from 2 to 12, and starts to flatten after 10 clusters. This suggests that, using the BIC criterion, the number of clusters chosen can be large



(a)

Matching Percentage / Adjusted Rand Index	Size of Neighborhood	0	2	4	6	8	10
	0		0.11	0.03	0.03	0.02	0.01
	2	49%		0.18	0.15	0.1	0.07
	4	43%	63%		0.13	0.2	0.16
	6	44%	62%	71%		0.44	0.3
	8	45%	62%	66%	81%		0.52
	10	45%	62%	63%	75%	85%	

(b)

FIG. 6. (a) *BIC score under different number of clusters.* (b) *The upper triangle shows the adjusted Rand index, and the lower triangle shows the matching percentage under varying neighborhood sizes.*

thus BIC may not provide an upper threshold for the number of clusters. This is a possible indication that there are a few outliers that do not belong to any given cluster.

We visually inspected the clustering results with varying number of clusters. As the number of clusters increases to more than three, additional clusters yield similar patterns, with a few very small clusters in size ($\leq 0.01\%$) that capture mostly the outlying features and big clusters that are not clearly distinguishable. Consequently, we choose to analyze the clustering with three clusters. More details of the analysis on the number of clusters can be found in Appendix B.

4.4. *Sensitivity analysis*. To evaluate how the uncertainty or the sampling error in the prevalence estimates affects the clustering results, we simulated 20 samples of prevalence data from a binomial sampling model. The prevalence of each condition in each simulation was calculated by dividing the simulated total number of member months of patients treated for the given condition by the total number of member months of all patients on Medicaid for each census area. We then measure the change of membership among the census tracts relative to the baseline prevalence in percentages and through the adjusted Rand Index, which measures the similarity between two clusterings, adjusting for the chance of grouping [Rand (1971)]. For the 20 comparisons, the adjusted Rand Index values have an average of 0.9, with a small standard deviation of 0.008, and the percentages of census tracts that changed membership are consistently less than 4%. Thus the standard errors from the prevalence estimates have limited impact on the clustering membership. This is due to the fact that the total number of member months of all Medicaid-enrolled children for most of the census tracts is large and thus the errors are small. Rural areas with low member months exhibit relatively large variation in prevalence estimates comparing to more population dense areas. However, the sets of census tracts that changed memberships are not the same when compared across the 20 simulated samples, and appear randomly on the map, with only slightly lower average member months of 3702 comparing to the nationwide average of 4011 member months.

In order to reduce the computation complexity, we assumed a fixed neighborhood structure. Part (b) of Figure 6 shows two measures for comparing any two clusterings obtained for varying neighborhood sizes. The upper triangle shows the adjusted Rand Index, and the lower triangle shows the matching percentage. The nominal EM algorithm coincides with the spatial EM algorithm when the size of neighborhood is 0. When the neighborhood size is small, a slight change of the neighborhood can have big impact on the clustering result. This is an indication that the results can be sensitive when the spatial effect is not properly incorporated. As the neighborhood size increases, the similarity between clusterings improves drastically. It is therefore not necessary to consider the spatial correlation between every possible pair of locations, since a neighborhood of size 10 can produce a sufficiently stable clustering result.

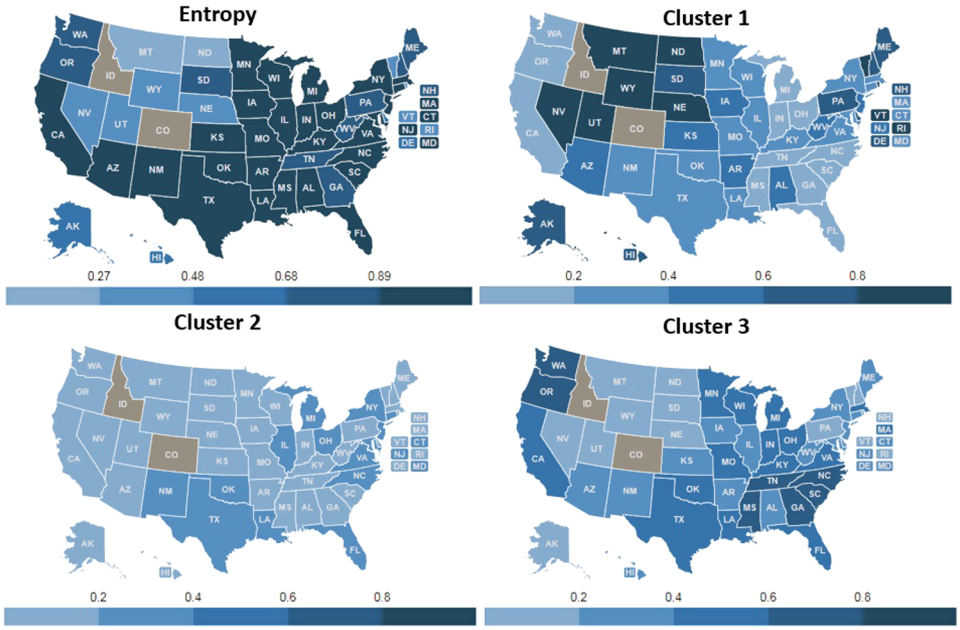


FIG. 7. *The entropy and proportion of census tracts within each state that belongs to each of the clusters.*

4.5. *Clustering results*: *United States*. Figure 7 displays the similarity (or dissimilarity) in clustering at the state level using the entropy measure (upper left map) and using the percentages for the three clusters. The west states have less variability in the clustering (lower entropy) than south west states. West states either predominantly are in cluster 1 (primarily represented by more severe chronic conditions) or cluster 3 (represented by mental health and respiratory chronic conditions). Cluster 2 (primarily represented by respiratory conditions) has low representation in most of the states except for a few southern states (e.g., FL, LA, NM, TX) and northern states (e.g., IL, NY, NJ, VA). These state-level differences point to pediatric chronic conditions the states might need to focus on disease management as well as prevention of severe outcomes.

Figure 8 shows the composition of each cluster by state and urbanicity for a subset of states. Urbanicity is defined using the rural-urban commuting area (RUCA) codes, which classify U.S. counties using measures of population density, urbanization, and daily commuting. The code is a single digit (1–9) classification, grouping counties based on the population of their metro area or their proximity to an urban area [United States Department of Agriculture (2004)]. We further grouped the 1–9 code into 3 major categories. Category 1, with an RUCA index 1, represents urbanized metropolitans areas; Category 2, with an RUCA index 2–6, represents smaller metropolitans and micropolitan areas; Category 3, with an RUCA index 7 and above, represents small towns and rural areas. The distribution of the

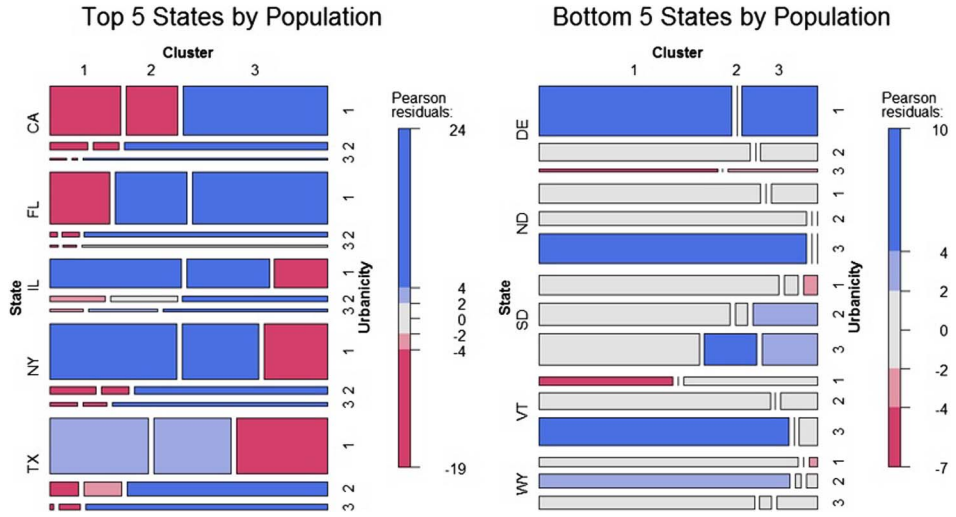


FIG. 8. *Visualization of the composition of each cluster by state and urbanicity for the top five and bottom five states by population under the Spatial EM Algorithm.*

census tracts across the three clusters in these areas varies by state. As the census tracts become more rural, the proportion of clusters 1 and 2 decreases drastically; that is, chronic mental conditions, Diabetes, Autism, and Respiratory conditions are more prevalent in urban areas. For states with the least dense population, Cluster 1 dominates across different urbanicity, and Cluster 2 is mostly nonexistent. These states exhibit much less heterogeneity comparing to states with higher population density and larger metropolitan areas. Further insights across other states are provided in Supplemental Material C [Zheng and Serban (2018)].

Figure 9 shows community-level cluster membership for Georgia. Most rural Georgia is predominantly in Cluster 3, with a mix of both mental health and respiratory chronic conditions, while suburban and urban areas are predominantly in Cluster 1 or 2, pointing to either heavily weighted mental and behavioral conditions or severe chronic conditions. We zoomed in the metropolitan Atlanta area, where several communities are assigned to Cluster 1 or 2. As noted in the heat map of Figure 2, the prevalences of the 25 conditions are differently weighted in Clusters 1 and 2; however, we see that there are many neighboring communities in the Atlanta area which are assigned to different clusters. Overall, this suggests that interventions for managing chronic conditions need to be much more targeted in urban areas.

Similar geographically granular analysis can be performed for other states. The maps for other states will be made available upon request from the authors of this paper.

5. Conclusions. The primary focus of this research paper is on deriving a spatial clustering of pediatric chronic conditions at the community level in the

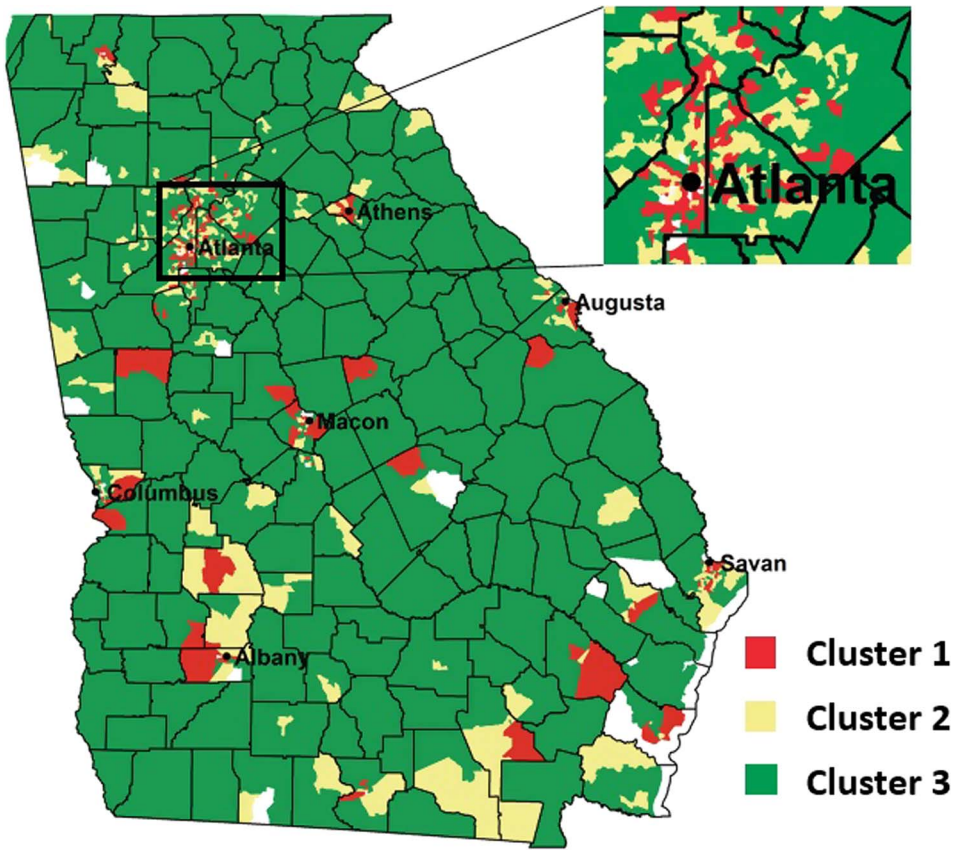


FIG. 9. *Clustering membership for the state of Georgia.*

United States. The data supporting this analysis consists of prevalences for 25 chronic conditions for the Medicaid-enrolled children.

The implementation of the spatial clustering approach relies on distributed computing to overcome the computational effort needed to perform the clustering analysis. While we were able to obtain the clustering after 11 hours of computing time with only the distribution of the data storage and retrieval, for a thorough analysis on the sensitivity of the clustering to the EM initialization and on the selection of the number of clusters, we needed much faster computations. Such large-scale studies can only be tackled by bridging statistical modeling and computational innovations.

This study has several limitations. The approach for estimating the prevalences at the census tract level from the prevalences observed at other geographic divisions, for example, zip code, falls under the modifiable areal unit problem or MAUP. Our approach is one of the simplest MAUP approaches, with noted limitations [Gotway and Young (2002)]. However, obtaining more rigorous prevalence

estimates at the census tract level requires extensive computational effort, which may be infeasible given the large scale of our data.

The correlation structure in the response data was assumed to follow an exponential correlation function using the Euclidean distance between pairs of census tracts centroids. The distance metric can be further improved, such as to use road distance between centroids, or similarity measures in urbanicity, socio-economics factors, or demographics. Follow up analysis based on the clustering results and additional area specific covariates can provide insights in determining the main drivers of the spatial variation and discrepancies in prevalence. Although we limited the implementation of the proposed distributed model-based clustering analysis to spatial correlation, the proposed algorithm can be applied to any type of correlation structures. In addition, we assumed the correlation structure to be fixed for each feature and each of the C components. Alternatively, we can extend the model to concurrently re-evaluate the correlation functions for each feature and cluster component at each EM iteration as the membership changes. Moreover, the neighborhood size was assumed to be fixed across all census tracts, which can be improved by a more granular definition of neighborhood based on urbanicity, for example.

Even though this research has several limitations, it has some important implications for interventions in managing chronic conditions. Many rural communities across the United States do not show a high burden of any particular condition, with similar weighting across respiratory conditions and behavioral & mental health conditions, with the lowest weight on more severe chronic conditions. This similarity in clustering across most of the rural communities points to that generally rural communities are in need of similar interventions, for example, improving access to mental and behavioral health providers. On the other hand, urban communities and some suburban communities present wide heterogeneity in clustering, with many of the urban communities being assigned in either high prevalence of severe chronic conditions or high prevalence of mental & behavioral conditions, which often are more severe for the Medicaid child population, overall pointing to a higher burden of severe conditions in some communities. While we cannot pinpoint the factors triggering such variations, we do recommend more targeted interventions for urban communities, with a focus on managing severe conditions.

Acknowledgments. The authors are thankful to Matt Sanders and Richard Starr in assisting with data safeguards and the information technology infrastructure. The authors are thankful to Dr. Julie Swann for the leadership in the protocol submission of the use of the MAX Medicaid claims data to the Centers of Medicare and Medicaid and in the Internal Review Board (IRB) approval process. The authors are also thankful to Pravara Harati and Preston Devaney for the support in deriving the prevalence estimates. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart,

Lung, and Blood Institute of the National Institutes of Health. The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

SUPPLEMENTARY MATERIAL

Supplement to "Clustering the prevalence of pediatric chronic conditions in the United States using distributed computing" (DOI: 10.1214/18-AOAS1173SUPP; .pdf). Supplementary Materials contain four sections. In Supplementary Material A, we describe the approach for estimating the census tract prevalence for chronic conditions using the Medicaid Analytic eXtract (MAX) claims data. In Supplementary Material B, we provide further details on the selection of the number of clusters. In Supplementary Material C, we present additional mosaic maps showing the composition of each cluster by state and urbanicity for all the states in our analysis. In Supplementary Material D, we share the implementation of the distributed computing approach for spatial clustering along with a read me file for guidance on how to use the software implementation.

REFERENCES

AMDAHL, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the Spring Joint Computer Conference* 483–485. ACM, New York.

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BESAG, J. and NEWELL, J. (1991). The detection of clusters in rare diseases. *J. Roy. Statist. Soc. Ser. A* **154** 143–155.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. MR3605826

BIRANT, D. and KUT, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.* **60** 208–221.

CAMERON, E., BATTLE, K. E., BHATT, S., WEISS, D. J., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., HAY, S. I., SMITH, D. L., GRIFFIN, J. T. et al. (2015). Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nat. Commun.* **6** Art. ID 8170.

CARSON, C., BELONGIE, S., GREENSPAN, H. and MALIK, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1026–1038.

CENTER FOR MEDICARE AND MEDICAID SERVICES (2017a). September 2017 Medicaid and CHIP enrollment data highlights. Available at https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html.

CENTER FOR MEDICARE AND MEDICAID SERVICES (2017b). Quality of care health disparities. Available at https://www.medicaid.gov/medicaid/quality-of-care/improvement-initiatives/health-disparities/index.html.

CHU, C.-T., KIM, S. K., LIN, Y.-A., YU, Y., BRADSKI, G., OLUKOTUN, K. and NG, A. Y. (2007). Map-reduce for machine learning on multicore. In *Advances in Neural Information Processing Systems* 281–288.

COCKERHAM, W. C., HAMBY, B. W. and OATES, G. R. (2017). The social determinants of chronic disease. *Am. J. Prev. Med.* **52** S5–S12.

CRESSIE, N. A. C. (2015). *Statistics for Spatial Data*, revised ed. Wiley, New York. Paperback edition of the 1993 edition [MR1239641]. MR3559472

DAVILA-PAYAN, C., DEGUZMAN, M., JOHNSON, K., SERBAN, N. and SWANN, J. (2015). Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. *Prev. Chronic Dis.* **12**. DOI:10.5888/pcd12.140229.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537

DIGGLE, P. J. and GIORGI, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *J. Amer. Statist. Assoc.* **111** 1096–1120. MR3561931

DING, C. and HE, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the* 21*st International Conference on Machine Learning* 29. ACM, New York.

ELLIOT, P., WAKEFIELD, J. C., BEST, N. G. and BRIGGS, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford Univ. Press, Oxford.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the* 2*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'*96) 226–231.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

GOTWAY, C. A. and YOUNG, L. J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.* **97** 632–648. MR1951636

GREEN, P. J. and RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.* **97** 1055–1070. MR1951259

JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427

KOPEC, J. A., SAYRE, E. C., FLANAGAN, W. M., FINES, P., CIBERE, J., RAHMAN, M. M., BANSBACK, N. J., ANIS, A. H., JORDAN, J. M., SOBOLEV, B. et al. (2010). Development of a population-based microsimulation model of osteoarthritis in Canada. *Osteoarthr. Cartil.* **18** 303–311.

KRIEGEL, H.-P., KRÖGER, P., SANDER, J. and ZIMEK, A. (2011). Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1** 231–240.

LAWSON, A., BIGGERI, A., BOHNING, D., LESAFFRE, E., VIEL, J.-F. and BERTOLLINI, R. (1999). *Disease Mapping and Risk Assessment for Public Health*. Wiley, New York.

LIU, Q. and IHLER, A. (2012). Distributed parameter estimation via pseudo-likelihood. In *International Conference on Machine Learning* (*ICML*) 1487–1494.

MEYER, S. and HELD, L. (2014). Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8** 1612–1639. MR3271346

NEFF, J. M., SHARP, V. L., MULDOON, J., GRAHAM, J., POPALISKY, J. and GAY, J. C. (2002). Identifying and classifying children with chronic conditions using administrative data with the clinical risk group classification system. *Ambul. Pediatr.* **2** 71–79.

OPENSHAW, S., CHARLTON, M., WYMER, C. and CRAFT, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *Int. J. Geogr. Inf. Syst.* **1** 335–358.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RIPLEY, B. D. (2005). *Spatial Statistics*. Wiley, New York. MR0624436

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR2130347

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. Ser. B* **71** 319–392. MR2649602

RUE, H. and TJELMELAND, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Stat.* **29** 31–49. MR1894379

THE WORLD HEALTH ORGANIZATION (2005). Chronic diseases and their common risk factors. Available at http://www.who.int/chp/chronic_disease_report/media/Factsheet1.pdf.

UNITED STATES DEPARTMENT OF AGRICULTURE (2004). Measuring rurality: Rural-urban continuum codes. Available at https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/.

WAKEFIELD, J. C. (2006). Disease mapping and spatial regression with count data. *Biostatistics* **8** 158–183.

WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley, Hoboken, NJ. MR2075123

WANG, M., WANG, A. and LI, A. (2006). Mining spatial-temporal clusters from geo-databases. In *Advanced Data Mining and Applications*. *Lecture Notes in Artificial Intelligence* **4093** 263–270. Springer, Berlin.

WOLFE, J., HAGHIGHI, A. and KLEIN, D. (2008). Fully distributed EM for very large datasets. In *Proceedings of the* 25*th International Conference on Machine Learning* 1184–1191. ACM, New York.

ZHENG, Y. and SERBAN, N. (2018). Supplement to "Clustering the prevalence of pediatric chronic conditions in the United States using distributed computing." DOI:10.1214/18-AOAS1173SUPP.

STEWART SCHOOL OF INDUSTRIAL
AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
755 FERST DR NE
ATLANTA, GEORGIA 30332
USA
E-MAIL: nserban@isye.gatech.edu