Spatial and Spatio-temporal Epidemiology 38 (2021) 100436

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Are spatial models advantageous for predicting county-level HIV epidemiology across the United States?

Danielle Sass[a,*], Bita Fayaz Farkhad[a], Bo Li[a], Man-pui Sally Chan[a], Dolores Albarracín[a]

University of Illinois at Urbana-Champaign, USA

ARTICLE INFO

Article history:
Received 19 November 2020
Revised 10 June 2021
Accepted 14 June 2021
Available online 16 June 2021

Keywords:
Dynamic bayesian network
Generalized estimating equation
HIV Prediction
Quantile regression
Spatial autoregressive model
Two-part model

ABSTRACT

Predicting human immunodeficiency virus (HIV) epidemiology is vital for achieving public health milestones. Incorporating spatial dependence when data varies by region can often provide better prediction results, at the cost of computational efficiency. However, with the growing number of covariates available that capture the data variability, the benefit of a spatial model could be less crucial. We investigate this conjecture by considering both non-spatial and spatial models for county-level HIV prediction over the US. Due to many counties with zero HIV incidences, we utilize a two-part model, with one part estimating the probability of positive HIV rates and the other estimating HIV rates of counties not classified as zero. Based on our data, the compound of logistic regression and a generalized estimating equation outperforms the candidate models in making predictions. The results suggest that considering spatial correlation for our data is not necessarily advantageous when the purpose is making predictions.

1. Introduction

Prevalence and incidence of infection with the human immunodeficiency virus (HIV) continues to be a major health crisis in the United States, despite the availability of successful biomedical interventions. The Centers for Disease Control (CDC) estimated that 1.2 million people are living with HIV in the US, including about 1 in 7 who are unaware of being infected and need testing (CDC, 2020). Refining predictive modeling techniques of the HIV epidemic continues to be a priority for health care professionals, public health officials, epidemiologists, and statisticians. The ability to predict future regional patterns of HIV epidemiology is key to allocating resources and implementing effective interventions, such as HIV testing and PrEP (PreExposure Prophylaxis). HIV prevalence varies spatially over different counties and demographic information provides key predictors of the county-level infection rate (Jones et al., 2018; Pellowski et al., 2013; Rosenberg et al., 2016; Sanchez et al., 2014). Other geospatial determinants related to HIV prevalence are low income, poverty, and lack of health resources (Douthit et al., 2015; Goswami et al., 2016; Vaughan et al., 2014). Public health decisions must be made with awareness of expected patterns of epidemiology within each county.

Predicting HIV epidemiology in the following year or time period usually relies on the HIV rates in the previous year or time period and the current demographic, risk, or health service predictors. Difficulties developing an appropriate and efficient prediction model at the county-level stems, in part, from data sparsity, due to the large amount of zero values, as well as data suppression for privacy. Despite past efforts to develop prediction models about HIV, to the best of our knowledge, most prior models ignored counties with zero cases or suppressed rates. Both of these limitations are not negligible, because zero and suppressed rates are common. Over the years 2012–2018, 40% of US counties have zero rates and 37% of counties contain suppressed rates, due to the county having either very few new HIV cases per year (1, 2, 3, or 4) or a population smaller than 100. A county may have a zero value or suppressed rate in the previous year, and a positive rate in the prediction year. Therefore, simply excluding counties with zeros and suppressed data would prohibit predicting a diagnosis rate for those counties that could newly be at risk, resulting in a loss of information for public health officials. Moreover, ignoring zeros and suppressed data may introduce bias to the statistical inferences made based only on the observed data (Little and Rubin, 2014; Rotnitzky and Wypij, 1994).

Most prior studies of HIV epidemiology and social determinants of health have used aggregated data at the state or regional level (Aral et al., 2006; Hanna et al., 2012; Zeglin and Stein, 2015). Due to increased accessibility to data at a finer scale, conducting HIV analyses at the county or zip code level have become more common, offering more relevant information to local health care officials (Chan et al., 2018; Harrison et al., 2008; Trepka et al., 2013). However, in order to avoid considering zeros and suppressed rates,

* Corresponding author.: Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820, USA.
E-mail address: dsass2@illinois.edu (D. Sass).

https://doi.org/10.1016/j.sste.2021.100436
1877-5845/

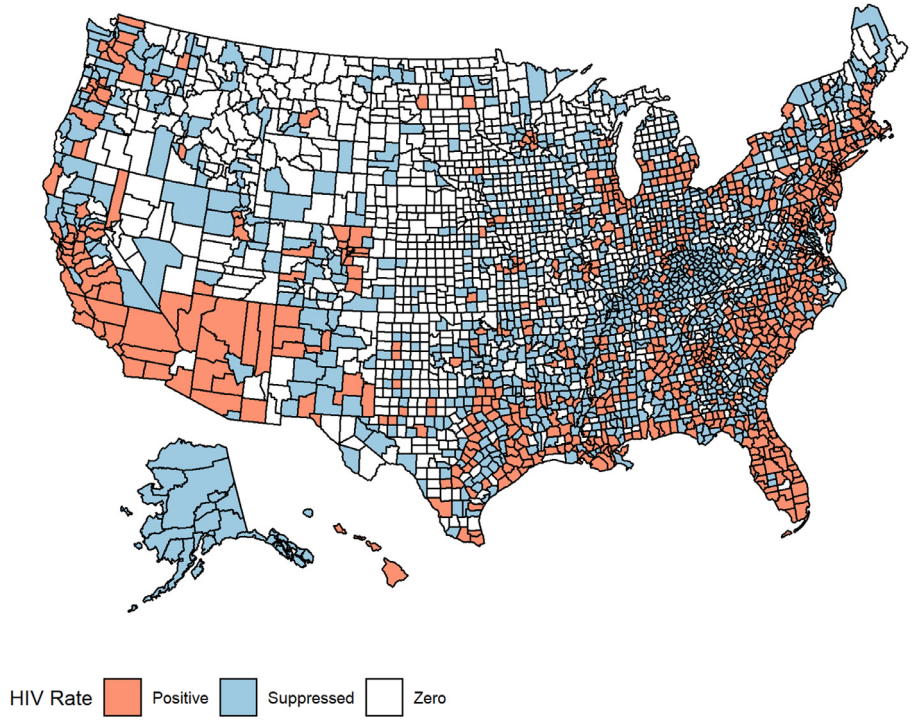


Fig. 1. Counties with zero, positive, or suppressed HIV diagnosis rates in 2015.

previous predictive models often assess only a small number of US counties, leaving a large group of US counties untouched. For example, Gray et al. (2016) conducted a county-level analysis of persons living with HIV in the southern United States, and Gonsalves and Crawford (2018) studied the HIV outbreak and response in a single county: Scott County, Indiana. Shand et al. (2018) developed a spatially varying auto-regressive model in a Bayesian framework to predict new HIV diagnosis rates at the county-level. Their analyses took spatial correlation into account but focused only on Florida, California, and the Northeast, where HIV cases are highly concentrated. Furthermore, their analyses only included counties with new diagnosis rates greater than zero and not suppressed.

The present research aims to identify a simple and efficient approach for HIV prediction over all counties in the US. Due to the large spatial domain, the method should also be computationally efficient for practical purposes. We find modern methods, such as Shand et al. (2018), may not be scalable to such a large dataset. Because of the large number of zeros in the data, we adopt a two-part framework for modeling HIV diagnosis. Two-part models have been widely employed to model short-duration rainfall, health care resources, and health care costs (Belotti et al., 2015; Cole and Sherriff, 1972; Li et al., 2008; Mihaylova et al., 2011), all of which contain a large body of zeros.

We consider both non-spatial models and cutting edge spatial models in each step of the analysis to identify the most efficient and accurate method. Following the two-part modeling framework, we first fit a binary model to estimate the probability of observing a positive-versus-zero rate, where positive includes both observed and suppressed data. This step used both logistic regression and spatial autologistic regression, which additionally incorporates the dependence of neighboring counties (Hughes et al., 2011). Then, for counties with positive rates, we employ five different prediction models ranging from a non-spatial generalized estimating equation, quantile regression, and dynamic Bayesian network, to two spatial models, including the spatial autoregressive lag model (Bivand et al., 2008; Srinivasan, 2015) and Bayesian spatially varying auto-regressive model (Shand et al., 2018). To evaluate the predictions for suppressed data, we calculate the empirical identification rate of how many predictions correctly fall into that category. As for predictions about the observed diagnosis rates, we compute the mean squared prediction errors. We also consider computation time when comparing the performance of all five models. Usually, complex models, such as spatial models, demand much higher computational power than the non-spatial methods, but this disadvantage can be acceptable if the model provides significantly more precise results. However, with more and more covariates to describe the spatial variability of HIV diagnosis rates, we speculate whether a spatial model is still advantageous for making predictions over the entire US.

This paper is organized as follows: Section 2 describes the data, Section 3 introduces the two-part modeling framework and describes the competing models considered, Section 4 provides the model evaluations, prediction results, sensitivity analysis, as well as an investigation of prediction with insufficient covariates. Finally, Section 5 concludes with a brief discussion.

The data that supports the findings of this study are available at https://www.cdc.gov/nchhstp/atlas/ and https://data.hrsa.gov/topics/health-workforce/ahrf.

2. Data

Annual new HIV diagnosis data from 2009 to 2018 at the county-level across the US are available from the CDC NCHHSTP Atlas (https://www.cdc.gov/nchhstp/atlas). We collect data from all 3,142 counties and county-equivalents in the US. County equivalents refer to places comparable to counties but called by different names, such as the Alaska census areas, Louisiana parishes, independent cities, and the District of Columbia. HIV rates are reported as the number of cases per 100,000 people. Approximately 40% of counties have zero HIV diagnoses reported, and only 23% have a non-suppressed positive number of HIV diagnoses. The remaining 37% of counties have suppressed HIV diagnosis rates because the county has very few cases (1, 2, 3, or 4) or a small population size (less than 100). Fig. 1 shows the distribution of the three groups

of HIV diagnosis rates for 2015, including zero, positive, and suppressed HIV diagnosis rates.

We make a simple imputation for counties with suppressed data before our analysis using the *imputeLCMD* package in R modified to use the Poisson distribution for count data. If those counties have a population greater than 100, the number of HIV cases can take on a value of 1, 2, 3, or 4. This method performs the imputation of interval-censored missing data for data missing not at random by using random draws from a truncated Poisson distribution with parameters estimated using quantile regression. We restrict the truncated Poisson distribution such that the resulting values are counts between 1 and 4. Quantile imputation is a flexible method that can be applied to impute dependent, bounded, censored, and count data and does not require the specification of a likelihood (Bottai and Zhen, 2013; Chen, 2014; Geraci and McLain, 2018). We sample from the distribution 100 times and for simplicity choose the mean of each county to be the imputed value of that county, rounded to the nearest whole number. We later perform a sensitivity analysis to evaluate the prediction results of the 100 imputations, the details of which are in Section 4.3. In our data set, only two counties have suppressed rates due to a population of less than 100. There is, in principle, no information about the number of cases in these counties, as the suppression is due to the population size. However, small population size corresponds to a large variability in possible HIV rates (rate $= 100{,}000 \cdot$ cases / population). To minimize the random noise that these counties add to the model, we assume the HIV rate in these two counties is zero. Given their neighboring counties have either zero cases or a very low number of cases, this assumption is reasonable and will have a minimal effect on our analysis.

To model new HIV diagnosis rates, we consider a variety of explanatory variables in the Area Health Resource Files, including data on population characteristics, economic, and environment at the county-level from over 50 data sources (AHRF, 2019). In particular, we include county demographics such as population, age, gender, race, unemployment rate, poverty rate, median income, education attainment, household occupancy information, primary mode of transportation, commute time, and occupation industry. Geographic related variables such as region and urbanization are also considered. We further collect prevalence rates of other sexually transmitted infections from the CDC NCHHSTP Atlas website, including chlamydia, gonorrhea, and syphilis, as well as HIV diagnosis rates from the prior years. In total, our analysis considers 96 variables.

We next compute the missing values of all explanatory variables, if any. Approximately 0.36% of the explanatory data are missing and imputed. Among the most popular methods for imputation is the k-Nearest Neighbour (KNN) algorithm (Andridge and Little, 2010). The nearest neighbor imputation identifies the k-counties that are most similar to the county with missing data with respect to observed characteristics. The missing value is then estimated using the average of the k-counties. Troyanskaya et al. (2001) showed that KNN is relatively insensitive to the exact value of k within the range of 10–20 neighbors. While Beretta and Santaniello (2015) did not provide an optimal number of neighbors, they advised to limit the number of k neighbors, because of the risk to severely impair the original variability of the data. Therefore, we choose to use the 10 nearest neighbors to impute any missing values, using the *bnstruct* R package (Franzin et al., 2017).

3. Prediction models

Our goal is to predict county-level new HIV diagnosis rates across the US. A major challenge that comes with the prediction lies in the rarity of the disease, leading to no incidents in many regions. Fig. 2 is a density plot of the logarithmic diagnosis rates

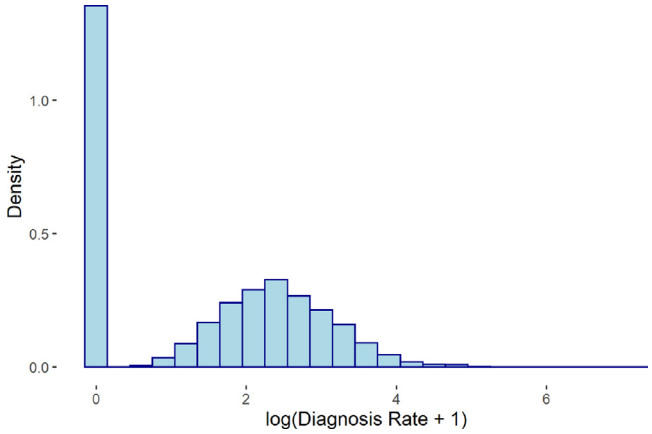


Fig. 2. Histogram of the diagnosis rates used in the training dataset years 2012–2014 including counties with suppressed data with diagnosis count predicted to be 1, 2, 3, or 4 using the interval-censored model.

in 2012–2014, which also includes counties with suppressed data with HIV cases predicted to be 1, 2, 3, or 4 using the interval-censored model described in Section 2. This figure shows a mass point at zero, which has been generally ignored in the literature. To account for the mass of zeros, we propose to construct a two-part model that permits the zeros and non-zeros to be generated by different densities. Let $Y_{i,t}$ be the new diagnosis rate in county $i = 1, \ldots, n$ and year $t = 1, \ldots, T$, and let $\mathbf{X}_{i,t-1}$ denote the vector of covariates for modeling $Y_{i,t}$. Further, let $Z_{i,t}$ denote the binary indicator of whether $Y_{i,t}$ is zero (0) or non-zero (1). If a county is suppressed, $Z_{i,t} = 1$ since we know at least one HIV diagnosis occurred. Our two-part model is

$$Y_{i,t} = \{\mathbb{E}(Y_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1}) + \epsilon_{i,t}\}Z_{i,t}, \tag{1}$$

where $\epsilon_{i,t}$ is an *i.i.d.* error term with $\mathbb{E}(\epsilon_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1}) = 0$. The distribution of $\epsilon_{i,t}$ depends on the distribution of $Y_{i,t}|Y_{i,t} > 0$. For example, if $Y_{i,t}|Y_{i,t} > 0$ follows a Poisson distribution, the errors follow a distribution behaving as a centered Poisson. Under this model,

$$\mathbb{E}(Y_{i,t}) = \mathbb{E}(Y_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1})\mathbb{P}(Y_{i,t} > 0|\mathbf{X}_{i,t-1}).$$

We predict new HIV diagnosis rates at $t + 1$ as

$$\hat{Y}_{i,t+1} = \hat{\mathbb{E}}(Y_{i,t+1}|Y_{i,t+1} > 0, \mathbf{X}_{i,t})\hat{\mathbb{P}}(Y_{i,t+1} > 0|\mathbf{X}_{i,t}). \tag{2}$$

Specifically, we first estimate the probability of a county observing a non-zero HIV diagnosis rate and classify each county into one of the two categories - zero or positive rates, then we estimate the new HIV diagnosis rate of the counties given they have positive rates. The binary classification results and the estimates of positive HIV diagnosis rates are then combined using Equation (2).

We use the previous year's explanatory variables as well as the previous three year's HIV diagnosis rates as covariates to make one year ahead predictions. When choosing an optimal number of variables, it is important to avoid both over-fitting, i.e., low predictive power due to the inclusion of too many variables without statistical justification at the model identification stage, and under-fitting, i.e., poor performance in modeling both the training and test data due to the inclusion of too few variables (Dietterich, 1995). We fit the two model parts separately. Both model parts use stepwise regression, selecting from the 96 explanatory variables described in Section 2. The binary part of Equation (1), classifying positive-versus-zero outcomes, selects 40 significant variables and the positive part of the model, predicting diagnosis rates for counties classified as positive, selects 37 significant variables. While the two model parts are allowed to select different variables, a few of the variables are found significant for both model parts. For example,

population, other STI prevalence, race, age, high school education rate, and prior year HIV diagnosis rate, which are known to be related to the epidemic nature of the HIV disease, are significant in classifying both positive-versus-zero outcomes and predicting the positive diagnosis rates.

In reality, the current year's covariates and HIV data may not be available in time to make prediction of next year, so we also demonstrate model performance using two year ahead predictions in the Appendix. When using data collected over time it is often important to consider a time trend. We incorporate a yearly time variable and evaluate model performance in the Appendix as well.

3.1. Probability for positive HIV diagnosis rates

To estimate the probability of a positive new diagnosis rate for a given county, $\mathbb{P}(Y_{i,t} > 0)$ which is equivalent to $\mathbb{P}(Z_{i,t} = 1)$, we consider both non-spatial logistic regression and centered spatial autologistic regression. The spatial autologistic model (Besag, 1972) can be advantageous over a non-spatial logistic regression model for spatial data because it accounts for effects of covariates as well as spatial dependence among the data simultaneously. We consider a centered spatial autologistic model because it weights the effects of both zero and non-zero neighboring observations, whereas a non-centered model could bias the estimated probability of a non-zero rate toward one (Caragea and Kaiser, 2009; Hughes et al., 2011; Wang, 2012). Two counties are considered neighbors if they have at least one shared boundary point.

The non-spatial logistic regression model is given by:

$$\frac{\mathbb{P}(Z_{i,t} = 1|\mathbf{X}_{i,t-1})}{1 - \mathbb{P}(Z_{i,t} = 1|\mathbf{X}_{i,t-1})} = \exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\theta}), \tag{3}$$

where $\mathbf{X}_{i,t}^T\boldsymbol{\theta}$ denotes the linear effects of the chosen covariates. The centered spatial autologistic regression is modeled as:

$$\frac{\mathbb{P}(Z_{i,t} = 1|Z_{j,t} : j \neq i, \mathbf{X}_{i,t-1})}{1 - \mathbb{P}(Z_{i,t} = 1|Z_{j,t} : j \neq i, \mathbf{X}_{i,t-1})} = \exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\theta} + \lambda \sum_{j \in N_{i,t}} Z_{j,t}^*), \tag{4}$$

where λ represents the spatial autoregressive coefficient, $N_{i,t}$ is the set of neighbors of county i in year t, and $Z_{j,t}^* = Z_{j,t} - p_{j,t}$ represents the centered response, where $p_{j,t} = \mathbb{P}(Z_{j,t} = 1)$ under non-spatial logistic regression. Inference for the centered spatial autologistic model is performed using maximum pseudolikelihood estimation.

3.2. Estimation for positive HIV diagnosis rates

In this section, we consider five different models for estimating $\mathbb{E}(Y_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1})$ for the HIV diagnosis rates. These five models represent both the fundamental statistical models and the cutting edge spatial models. For the spatial models, two counties are considered neighbors if they have at least one shared boundary point. The estimation for positive HIV diagnosis rates focuses on the counties identified as non-zero using the classifiers described in Section 3.1. It is not of interest to predict a positive diagnosis rate for all counties, because the mass point at zero indicates many counties will have a zero HIV diagnosis.

3.2.1. Generalized estimating equation with log-link (GEE)

The generalized estimating equation (GEE) is an extension of the generalized linear model, which is perhaps the most popular statistical technique for modeling non-Gaussian data. When count data follows a Poisson distribution, the generalized linear model with a log-link is the typical choice. However, after converting count data into rates, the Poisson assumption that the mean and variance are equal becomes violated. To overcome this potential pitfall, we use the Poisson generalized estimating equation with a log-link using the R package *geepack*, which is a semiparametric method using the Huber-White sandwich estimator for robust variance estimation (Huber, 1967). The GEE provides consistent estimates, even if the correlation structure is misspecified in modeling the population average effects:

$$\log(\mathbb{E}(Y_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1})) = \mathbf{X}_{i,t-1}^T\boldsymbol{\beta} + \log(100,000/\eta_{i,t-1}),$$

where $\mathbf{X}_{i,t-1}^T\boldsymbol{\beta}$ denotes the linear effects of the chosen covariates, $\log(100,000/\eta_{i,t-1})$ is the offset from converting count data into rates, and $\eta_{i,t-1}$ is the population of county i in year $t-1$. Then we have

$$\mathbb{E}(Y_{i,t}|Y_{i,t} > 0, \mathbf{X}_{i,t-1}) = e^{\mathbf{X}_{i,t-1}^T\boldsymbol{\beta}} \times 100,000/\eta_{i,t-1}.$$

3.2.2. Quantile regression (QUANT)

An alternative to modeling the conditional mean is to model the conditional τth quantile using quantile regression, given a set of explanatory variables. Quantile regression makes no assumption about the data distribution, but to regularize the variance we take a log transformation of the new HIV diagnosis rate to make its distribution approximately Gaussian (Shand et al., 2018). Quantile regression takes the form:

$$\mathbb{Q}_\tau(\log(Y_{i,t})|Y_{i,t} > 0, \mathbf{X}_{i,t-1}) = \mathbf{X}_{i,t-1}^T\boldsymbol{\beta}(\tau),$$

where $\mathbb{Q}_\tau(Y)$ is the τth quantile of Y, and $\mathbf{X}_{i,t-1}^T\boldsymbol{\beta}(\tau)$ denotes the linear effects of the chosen covariates dependent on the value of τ. The function $\boldsymbol{\beta}(\tau)$ is optimized by solving

$$\underset{\boldsymbol{\beta}(\tau)}{\operatorname{argmin}} \sum \rho_\tau(\log(Y_{i,t}) - \mathbf{X}_{i,t-1}^T\boldsymbol{\beta}(\tau)),$$

where $\rho_\tau = \tau - \mathbb{1}\{\log(Y_{i,t}) < \mathbf{X}_{i,t-1}^T\boldsymbol{\beta}(\tau)\}$ is the tilted absolute value function. The model was fit using the R package *quantreg*. Unlike least squares regression, quantile regression is invariant to monotone transformations and avoids assumptions about the parametric distribution of the error process. Quantile regression can be advantageous when the data contains outliers because quantile regression is more robust to outliers than least squares regression, and outliers may be present in our data due to HIV outbreaks.

3.2.3. Spatial simultaneous autoregressive lag model (SSAL)

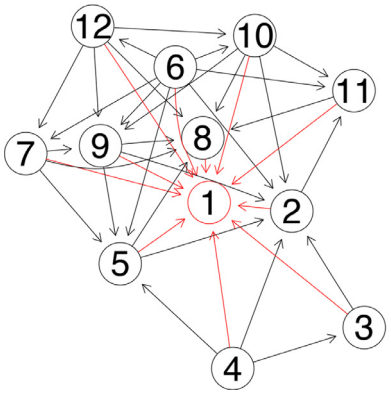
When data comes from a spatial process, nearby observations often tend to be similar. Therefore, a model that can incorporate spatial dependence is often more efficient. To determine the appropriate spatial regression model we use the Lagrange Multipliers diagnosis test for error and lag dependence of the GEE model. The resulting test statistic and p-value for error dependence are approximately 0 and 1, respectively, indicating there is no spatial dependence in the error term. The resulting test statistic and p-value for lag dependence are 5.8 and 0.02, respectively, indicating there is dependence among diagnosis rates in nearby counties. Therefore, we consider a spatial simultaneous autoregressive lag model (SSAL) to incorporate spatial dependence when making predictions.

The SSAL model is an extension of our generalized estimating equation, that uses the spatial variation in nearby observations to predict $Y_{i,t}$. The model was fit by extending the R package *spatialreg* to use the generalized estimating equation. Similar to Section 3.2.1, we use a Poisson distribution with a log-link and model new HIV diagnosis rates as

$$\log(\mathbb{E}(Y_{\cdot,t} - \rho W Y_{\cdot,t}|Y_{\cdot,t} > 0, \mathbf{X}_{\cdot,t-1})) = \mathbf{X}_{\cdot,t-1}^T\boldsymbol{\beta} + \log(100,000/\eta_{\cdot,t-1}),$$

where $Y_{\cdot,t}$ is the vector of new HIV diagnosis rates at time t, $\mathbf{X}_{\cdot,t-1}^T\boldsymbol{\beta}$ denotes the linear effects of the chosen covariates, $\log(100,000/\eta_{\cdot,t-1})$ is the offset, ρ is the autoregressive parameter, and W denotes a row-standardized neighborhood matrix where the ith diagonal element is 1 and the (i,j)th off-diagonal element is -1/total number of neighbours for county i if counties i and j share a border and 0 otherwise. Then we have

$$\mathbb{E}(Y_{\cdot,t}|Y_{\cdot,t} > 0, \mathbf{X}_{\cdot,t-1}) = (I - \rho W)^{-1} + exp(\mathbf{X}_{\cdot,t-1}^T\boldsymbol{\beta}) \times 100,000/\eta_{\cdot,t-1}.$$



1 HIV diagnosis rate (Y_t)
2 HIV diagnosis rate (Y_{t-1})
3 HIV diagnosis rate (Y_{t-2})
4 HIV diagnosis rate (Y_{t-3})
5 HIV binary indicator (Z_{t-2})
6 population
7 gonorrhea rate
8 did not work (age 16-64)
9 white race
10 other race
11 females (age 25-44)
12 walk to work

Fig. 3. A subset of the dynamic Bayesian network structure. The subset shows the 11 variables that are the direct parents, as indicated by the red arrows, of the new HIV diagnosis rate out of the 37 total variables chosen. All variables are taken from the prior year, $t-1$, unless otherwise specified. HIV diagnosis rate (Y_t) is the response variable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.4. Dynamic bayesian network (DBN)

Dynamic Bayesian networks (DBN) are popular for the construction of disease models in the biomedical and health-care field (Weins and Wallace, 2016). A DBN is a directed acyclic graph in which nodes represent the variables and arrows represent the temporal dependencies that are quantified by probability distributions (Margaritas, 2003). A DBN consists of two parts: (a) the construction of the directed network structure and (b) the inference of local probability distributions for each variable conditional on the parent nodes. Many methods exist for learning the network structure, among the most popular methods is the Hill-climbing score-based approach, which searches for the structure that maximizes the networks Bayesian information criterion (Margaritas, 2003). We adapt the Hill climbing approach to build our network structure, shown in Fig. 3. From the network, we observe that the prior year diagnosis rates, population, and other STI rate remain significant to the prediction of new HIV diagnosis rates.

While many methods exist for constructing the network structure, inference for parameter nodes using R packages is restrictive. Different packages in R are available for making inferences for DBN. The package *bnstruct* (Franzin et al., 2017) only allows for inference with discrete variables and continuous variables are quantized, which could result in loss of information. The package *bnlearn* (Scutari, 2010) allows for continuous variables, though parameter node inference for continuous variables is restricted to maximum likelihood estimation using only the parent nodes which could result in suboptimal inferences.

We choose to use the *bnlearn* package in R for constructing our network structure and parameter inference. In general, the nodes of continuous variables are assumed to follow a Gaussian distribution and discrete variables follow a multinomial distribution. We use a Gaussian distribution with variance inversely proportional to $\eta_{i,t}Y_{i,t}$, where $\eta_{i,t}$ is the population of county i in year t (Shand et al., 2018). The parameters for each node of the network structure are optimized by maximum likelihood estimation. Specifically, we use the generalized linear model with a log link function based on the value of parent nodes:

$$\log(\mathbb{E}(Y_{i,t}|Y_{i,t}>0,\boldsymbol{\pi}_{i,t-1})) = \boldsymbol{\pi}_{i,t-1}^T\boldsymbol{\beta},$$

where $\boldsymbol{\pi}_{i,t-1}$ are the observed parent nodes of county i in year t and $\boldsymbol{\beta}$ are the corresponding coefficients. Then the expected value of our new diagnosis rate is

$$\mathbb{E}(Y_{i,t}|Y_{i,t}>0,\boldsymbol{\pi}_{i,t-1})) = e^{\boldsymbol{\pi}_{i,t-1}^T\boldsymbol{\beta}}.$$

3.2.5. Bayesian spatially varying auto-regressive model (SVAR)

Recently, Shand et al. (2018) proposed a spatially varying autoregressive model (SVAR) to predict HIV rates in Florida, California and parts of the Northeastern US. The model assumes an autoregressive model with order 1 (AR(1)) for each county and further assumes that the autoregressive coefficients over all counties follow a spatially correlated process. Their model allows for spatial and temporal correlation as well as space-time interactions. Let $U_{i,t}$ be a centered process of $\log(Y_{i,t})$, where $Y_{i,t}$ are the rates predicted to be greater than zero. The model is designed in a Bayesian hierarchical context:

$$\begin{aligned}
&\text{Level 1:} && U_{i,t} = \mathbf{X}_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t}\rho_i(U_{i,t-1} - \mathbf{X}_{i,t-2}^T\boldsymbol{\beta}) + \epsilon_{i,t}\\
&\text{Level 2:} && \rho_i = \Phi(V_i) \quad \text{where} \quad \mathbf{V} \sim MVN(0, \tau_\rho^2[(1-\lambda_\rho)I + \lambda_\rho W]^{-1}),\\
&\text{Hyperpriors:} && \tau_\rho^2 \sim IG(a_1,b_1), \lambda_\rho \sim Uniform[0,1], \sigma^2 \sim IG(a_2,b_2)
\end{aligned}$$

where $\mathbf{X}_{i,t-1}^T\boldsymbol{\beta}$ denotes linear effects of the previous year's covariates, ρ_i are spatially varying AR(1) coefficient, $\psi_{i,t}$ ensures correlation is measured, $\epsilon_{i,t} \overset{\text{iid}}{\sim} N(0,\sigma^2 q_{i,t})$ for $q_{i,t} = 100,000/(\eta_{i,t}Y_{i,t})$ where $\eta_{i,t}$ is the population of county i in year t, Φ represents the Gaussian cumulative distribution function, *MVN* represents a multivariate normal distribution for the vector $\mathbf{V} = (V_1,\ldots,V_n)'$, and *IG* represents a semiconjugate inverse gamma hyperprior. Lastly, W denotes a neighbourhood matrix with the ith diagonal element as the total number of neighbours for county i whereas the (i,j)th off-diagonal element is -1 if counties i and j share a border and 0 otherwise. Basically, ρ_i is modeled as a probit transformation of a correlated Gaussian process. The estimation is performed through Markov chain Monte Carlo by alternating Gibbs and Metropolis-Hastings sampling. Predictions are obtained by taking the posterior median of the $Z_{i,t+1}^*$ sampling chain. Further details can be found in Shand et al. (2018).

We adapt the SVAR model in Shand et al. (2018) to our data. However, all counties in the US cannot be modeled at once due to the computation bottleneck in their method. Computational burden is a typical issue for spatial models that involve a spatial covariance matrix. To implement their methods, we group the counties into nine divisions (New England, Mid-Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific) as specified by the US census bureau.

3.3. Model assessment measures

To evaluate the performance of predictions, we calculate the mean squared prediction error (MSPE) for counties with observed new diagnosis rates and the empirical identification rate (EIR) for counties with suppressed data. Let $Y_{k,t}$ be the observed HIV diagnosis rate for $k = 1,\ldots,(n-n_s)$, where $n = 3,142$ is the total number of counties and n_s is the total number of counties with sup-

pressed data. Then we have

$$MSPE = \frac{1}{n - n_s} \sum_{k=1}^{n-n_s} (\hat{Y}_{k,t} - Y_{k,t})^2.$$

To evaluate the prediction of counties with suppressed data, we examine the empirical identification rate (EIR) of the predictions falling into 1, 2, 3, 4 cases. Define

$$S_k = \mathbb{1}(0 < \hat{Y}_{k,t}\eta_{k,t}/100,000 < 5)$$

for $k = 1, \ldots, n_s$, where $\eta_{k,t}$ follows the earlier notation in Section 3.2.5 representing the population of county k in year t. The indicator function equals one if the prediction falls in the range of being suppressed. We have

$$EIR = \frac{1}{n_s} \sum_{k=1}^{n_s} S_k.$$

4. Evaluation of models

We use data from the years 2012–2014 as training data and evaluate the prediction skills of different models using data from the years 2015–2018. The reason we employ the same trained model for predictions of different years is because we find no advantage of iteratively training the model, perhaps due to the stationarity of the data in time. The prediction results with iteratively trained models are reported in the Appendix, for example, the 2016 year prediction is obtained based on the models trained using data from years 2013–2015. The reason that we choose only three years of training data is because three years of data is the most data that the spatial autologistic regression model and SSAL model could handle due to the size of the neighborhood matrix.

Below we will first compare the classification performance between the logistic regression model in (3) and the spatial autologistic regression model in (4), then based on the preferred classification model we compare the five prediction models described in Section 3.2 by predicting new HIV diagnosis rates for the years 2015 to 2018. Finally, we assess the sensitivity of the suppressed HIV data imputation to the prediction skills using the best performing models. We close this section by investigating the effect of considering spatial correlation when covariates are insufficient.

4.1. Classification assessment

The classification models predict the probability that at least one new HIV diagnosis occurs in a county. Since it is unlikely $P(Z_{i,t} = 1)$ will be exactly zero, we select a cutoff to classify which counties are likely to be zero, allowing us to distinguish the mass point at zero in our data. This will also allow us to compare the accuracy of the non-spatial logistic regression and centered spatial autologistic regression models. We choose a classification cutoff of 0.3, i.e., if the probability of observing a positive new diagnosis rate is less than 0.3, then $Z_{i,t} = 0$, otherwise $Z_{i,t} = 1$. When choosing the cutoff for our binary prediction, weighing the impact that false negatives (sensitivity) and false positives (1-specificity) have on the results is critical. A lower cutoff decreases the number of false negatives while increasing the number of false positives and a higher cutoff does the opposite. When considering different cutoffs, we want a low enough cutoff that maintains a high true positive rate. This way we are only classifying counties as zero if they are highly likely to be zero. We resort to the receiver operating characteristic (ROC) to help us choose the cutoff value. Fig. 4 shows the ROC curve of the training dataset. We set the false negatives to be at least twice as impactful as false positives resulting in a cutoff of 0.3.

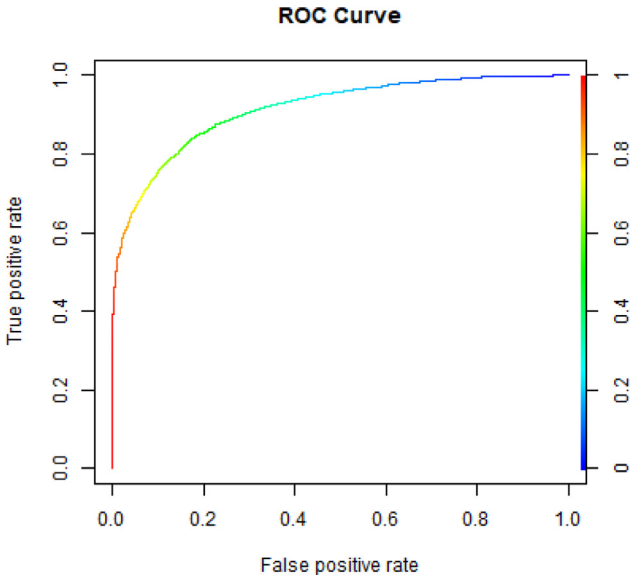


Fig. 4. ROC curve of the training dataset. The ROC curve is color coded based on the cutoff value with the cutoff scale shown on the right vertical axis.

To determine if the centered autologistic regression model is more appropriate than the non-spatial logistic regression, we evaluate the amount of spatial autocorrelation in the residuals of logistic regression using the Moran's I test (Moran, 1950). Table 1 shows the resulting test statistic and p-values for both the non-spatial logistic regression and the centered autologistic regression model based on the testing data. The p-value of the non-spatial logistic regression model is slightly smaller than the 0.05 significance level, indicating the model inadequacy of capturing the spatial correlation in the data. However, the two models produce very similar classification results, which is most important for the purpose of prediction. Given these considerations, we choose the non-spatial logistic regression for its parsimony to estimate the probability of whether a county will have a zero or positive HIV diagnosis rate, even though the computation time advantage is not much compared to the autologistic model.

4.2. Prediction assessment

Having chosen the non-spatial logistic regression as the preferred method for estimating $\mathbb{P}(Y_{i,t} > 0|\mathbf{X}_{i,t-1})$, the HIV rates for counties classified as non-zero are then estimated using the five models described in Section 3.2. The results from the two model

Table 1
Comparing the non-spatial logistic regression model to the centered spatial autologistic model. Test statistics and p-values for Moran's I on testing spatial correlation among residuals. Model accuracy and sensitivity of the testing dataset. Computation time in seconds.

	Moran's I		Model Performance		
	Statistic	p-value	Accuracy	Sensitivity	Time
Non-spatial logistic regression	1.7723	0.0382	80.23%	92.64%	1,397
Centered spatial autologistic	0.7008	0.2417	80.28%	92.54%	1,504

Table 2
One year ahead prediction results of new HIV diagnosis rates. Coefficient estimates are assumed stationary in time and trained using prediction years 2012–2014. Computation time in seconds.

	2015		2016		2017		2018		
Model	MSPE	EIR	MSPE	EIR	MSPE	EIR	MSPE	EIR	Time
GEE	326.94	80.1%	46.37	80.5%	40.26	79.6%	35.81	84.4%	173.7
QUANT	328.61	69.2%	43.06	71.0%	39.94	69.2%	34.25	73.8%	156.3
SSAL	325.26	77.9%	46.60	77.8%	42.97	76.1%	35.49	81.8%	749.8
DBN	328.38	78.4%	42.61	78.7%	30.52	77.9%	35.24	83.0%	200.8
SVAR	342.01	83.3%	68.26	82.4%	34.69	81.5%	38.46	85.7%	162,916

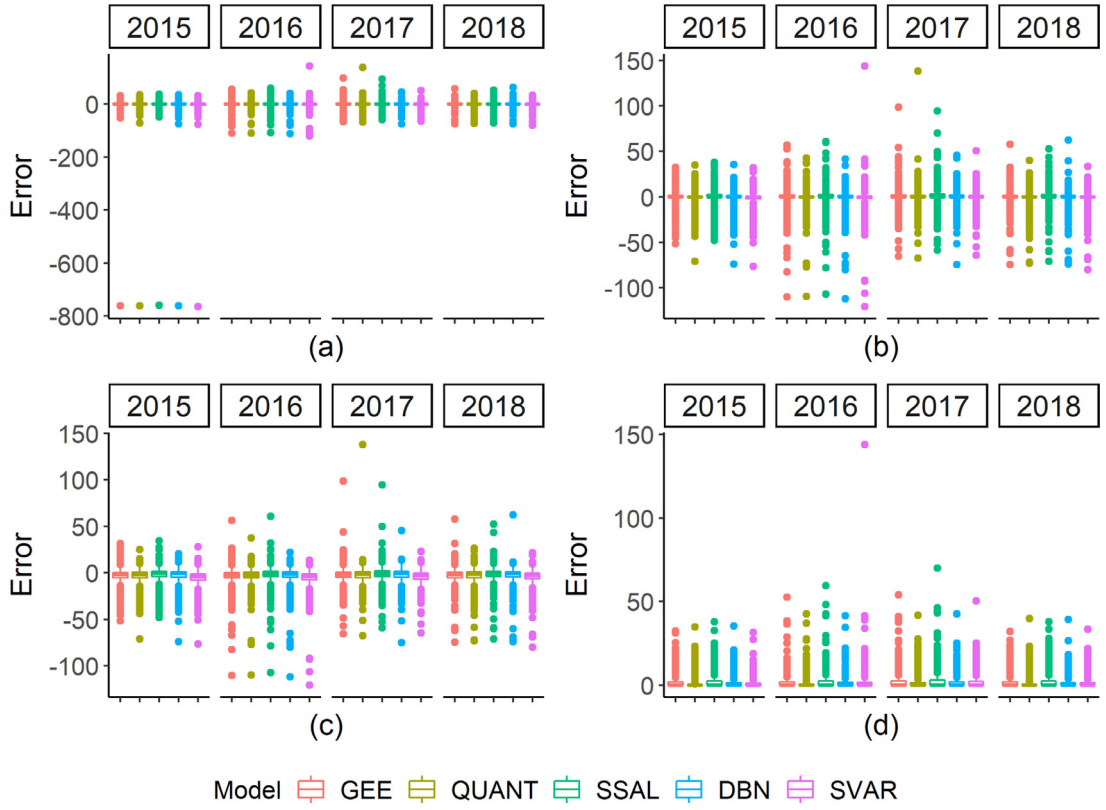


Fig. 5. Boxplot of the prediction error for our 5 models. (a) All counties. (b) All counties except the extreme outlier in 2015. (c) Only counties with a true HIV diagnosis greater than 0. (d) Only counties with a true HIV diagnosis equal to 0.

parts then jointly predict new HIV diagnosis rates, $\hat{Y}_{i,t}$, for all counties using Eq. 2, i.e., $\hat{Y}_{i,t}$ will be 0 if the logistic regression classifies the county as having a rate of zero and otherwise estimated using the five models. We will refer to the joint predictions simply by the model names in Section 3.2 since all models use the same non-spatial logistic regression to classify positive rates. Table 2 summarizes the prediction results of the two-part model for the testing data in terms of MSPE and EIR. The MSPE is to evaluate the prediction for the observed HIV diagnosis rates and the EIR is to evaluate whether the prediction can correctly identify the counties with suppressed data. A larger EIR indicates a more accurate identification for the suppressed data.

The GEE and SSAL model have very similar results, and we also find that the optimized λ in the SSAL model is nearly 0.01. When $\lambda = 0$, the SSAL model simplifies to the GEE model. So a very small λ indicates little spatial influence, resulting in similar parameter estimates and thus, similar predictions of the two models. Between the GEE and SSAL model, the GEE model is preferred for its parsimony and slightly better results.

The DBN model also has similar results as the GEE model. Recall that the DBN maximum likelihood estimation is only making use of the parent nodes, indicating that only a subset of the relevant variables may be necessary for prediction. The DBN model uses the Gaussian distribution and requires a variance specification, which the GEE model avoids. While the DBN can improve the prediction in this instance, it is more volatile to initial covariates considered and suppressed value chosen. The DBN can be most helpful with visualizing dependencies among variables. The QUANT model performs comparable in terms of MSPE, however performs the worst in terms of EIR. That is not very surprising because as a robust regression method, QUANT weakens the influence of the data points that are far from their center and thus the fitted model may not do well on the lower end of the data.

The SVAR model uses 50,000 iterations with a 10,000 burn-in, and surprisingly is not the most competitive model in terms of MSPE for our data given its complexity. The SVAR was developed for regional data and performed well at capturing the correlation between observed positive HIV diagnosis rates at the smaller scale in Shand et al. 2018. However, its advantage does not seem to carry over to a large scale data with more abundant covariates and suppressed data. The short training time series considered could also be a contributing factor. Moreover, the computation time for the

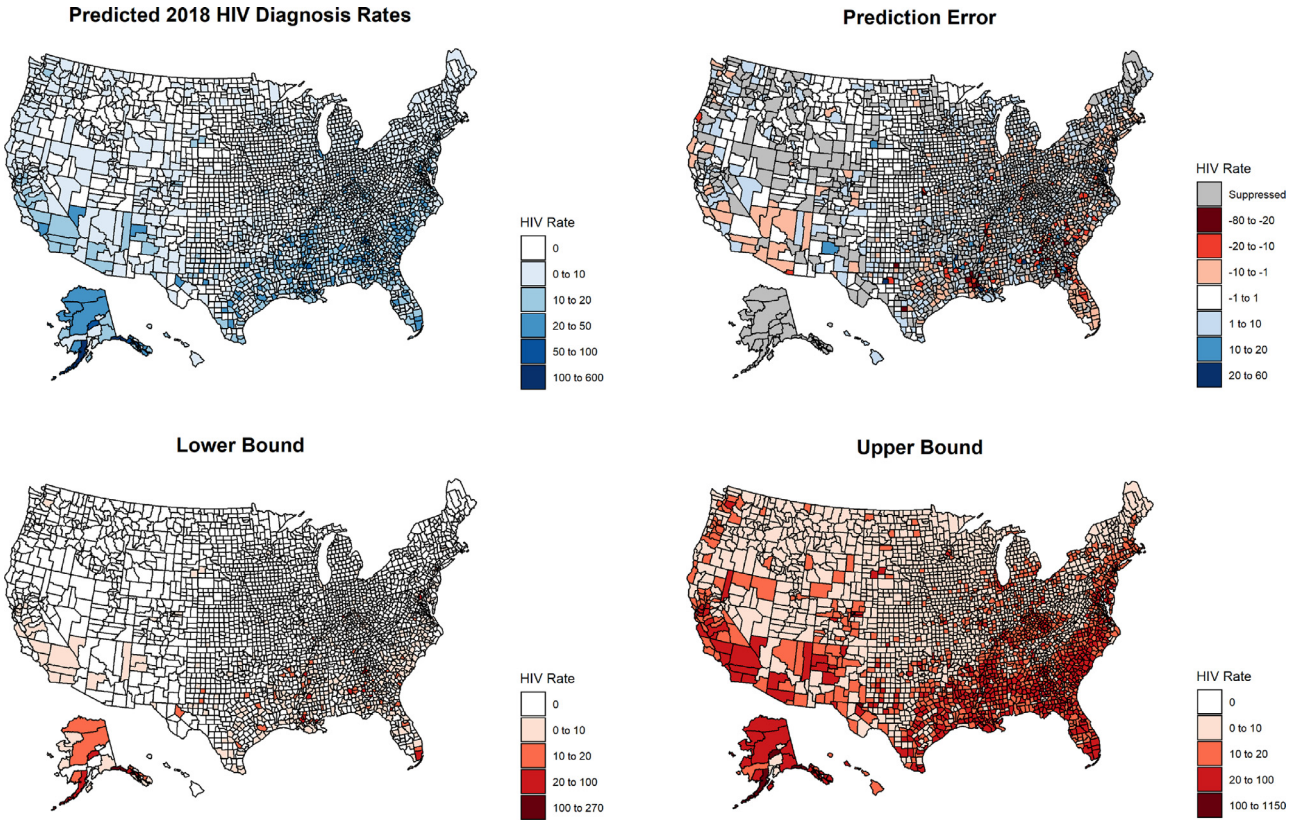


Fig. 6. Predicted HIV diagnosis rates for 2018 across the United States and the corresponding errors (top). 95% confidence interval for the predicted HIV diagnosis rates (bottom).

SVAR model is overwhelming compared to other models, due to its handling of a large covariance matrix and the Bayesian sampling.

Fig. 5 shows a boxplot of the prediction error for each model. Plot (a) of the boxplot makes it clear that the large MSPE in 2015 is due to a single extreme outlier, the HIV outbreak from Scott County, IN, as represented by the small dots at the bottom of year 2015. None of the models are able to capture this because they are all formulated around the mean or median prediction and more importantly the predictor variables are unable to reflect the outbreak. Plot (b) in Fig. 5, excludes the outlier in 2015 to provide a closer visualization of the main results. Plot (c) and Plot (d) further breakdown the error results into two categories: counties with a true HIV diagnosis greater than 0 and those with a true HIV diagnosis of 0, to demonstrate where the largest errors are originating from. The bottom and top of each "box" represents the first and third quartile, respectively. The QUANT model has the smallest interquartile range (third quartile minus first quartile), indicating that it has the most number of counties with very small error. GEE, SSAL, and DBN have very similar errors, as is also seen by the MSPE in Table 2. The SVAR model performs similar as well, except for a few more extreme outliers in 2015 and 2016, as shown by the dots in the top panel, explaining the slightly larger MSPE.

Given the above assessment between the five models, the GEE with log-link function appears to be the most parsimonious and sufficient model for making predictions based on our data set. Fig. 6 shows the predicted HIV diagnosis rates for all counties using the two-part GEE model for the year 2018 along with the corresponding 95% confidence interval. The confidence interval for the two-part model is calculated using the delta method approximation to ensure uncertainty in both parts are taken into account. The prediction error of a suppressed county is labeled as suppressed, because the true diagnosis rate is unknown. From this Figure, we observe that our model had a tendency to under-predict diagnosis rates in Florida and other parts of the south. The model was more accurate in the Midwest and West, with no large prediction errors.

4.3. Sensitivity to suppressed data imputation

The above analysis is based on assuming the suppressed observation is imputed using the county average of 100 simulated samples from a truncated Poisson interval-censored model. In order to assess the sensitivity of the analysis to different imputed values, we calculate predictions for each of the 100 simulated samples and evaluate the performance using the assessment measures as described in Section 3.3. For the classification step, non-spatial logistic regression is used. The SSAL and SVAR models are excluded from this study, due to the large computation time.

Fig. 7 shows the MSPE and EIR of 100 simulation for the years 2015–2018. The '×' in each boxplot represents the results, from Table 2. Based on the plots, the GEE, QUANT, and DBN models perform similar to the results in Table 2 with only a small variability in calculated MSPE for each year. While the GEE and QUANT model have fairly consistent EIR performance, the DBN model shows a lot more variability in its ability to predict suppressed counties. Overall, this implies that the model comparisons are insensitive to the suppressed value chosen.

4.4. Spatial correlation and modeling with insufficient covariates

While HIV diagnosis rates are geographically indexed, we expect spatial autocorrelation exists between neighbors. However, based on our results from Table 2, it appears the covariates employed are sufficient at capturing the spatial variability, rendering the more complex SVAR model obsolete. However, the importance of accounting for spatial correlation has been realized in

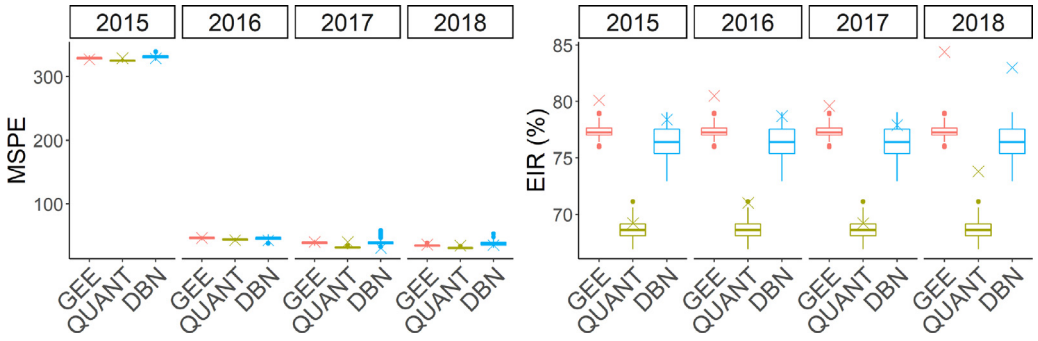


Fig. 7. Boxplots of the MSPE (left) and EIR (right) of 100 simulation samples for the GEE, QUANT, and DBN model. The '×' represents the results, from Table 2, under the assumption that the count of a suppressed county is the average of the 100 samples for that county.

Table 3
One year ahead prediction results of new HIV diagnosis rates with models trained and tested using only three covariates as explanatory variables. Coefficient estimates are assumed stationary in time and trained using prediction years 2012–2014. Computation time in seconds.

	2015		2016		2017		2018		
Model	MSPE	EIR	MSPE	EIR	MSPE	EIR	MSPE	EIR	Time
GEE	348.31	75.6%	74.24	74.7%	49.10	75.7%	56.03	76.9%	7.7
QUANT	357.37	76.2%	88.25	75.0%	63.28	76.5%	72.07	77.7%	6.5
SSAL	350.32	67.6%	81.82	66.9%	58.73	67.4%	64.93	69.7%	713.0
DBN	351.53	65.4%	78.96	65.9%	58.06	65.8%	65.08	67.6%	0.2
SVAR	344.28	80.3%	55.23	77.1%	45.80	75.0%	46.36	77.7%	170,136

many disciplines such as ecology and epidemiology (Auchincloss et al., 2012; Bahn et al., 2006). There are numerous examples that demonstrate the advantages of accounting for spatial correlation in Bayesian hierarchical models for ecological problems (Gelfand et al., 2006; Hoeting, 2009; Hooten and Wikle, 2008; Waller et al., 2007). There are also many known disadvantages of ignoring spatial correlation, such as, the analysis may lead to erroneous conclusions (Kühn, 2007; Shand et al., 2018), standard errors may be underestimated (Hoeting, 2009; Schabenberg and Gotway, 2005), and relevant covariates may be excluded in regression model selection (Hoeting et al., 2006).

To demonstrate the advantage of spatial models when a lack of covariates are available, we consider a case when only three explanatory variables (population, race, and education) are used to predict one year ahead HIV diagnosis rates. We assume the data are stationary in time and the prediction model is trained using the years 2012–2014.

Table 3 shows that when limited or insufficient covariates are employed, the more complex SVAR model is indeed the better performing model. The SSAL model nevertheless still had no advantage over the GEE model. This could be because the cutting edge SVAR model captures the spatial dependency of the HIV data better than the SSAL model. In any case, this example implies that if insufficient covariates are selected in the modeling procedure, the advantages of a spatial model could outweigh the computational burden and model complexity. However, based on our model results in Table 1 and 2, considering spatial correlation in our data set is not advantageous due to the covariates ability to capture the data variability.

5. Discussion

Predicting new HIV diagnosis rates at the county-level across the United States is important for the health department to effectively allocate the intervention resources. Due to the rareness of HIV and the confidentiality concern of health data, the county-level new HIV diagnosis rates contain a high percentage of zeros and suppressed data. We proposed to treat the data with a two-part model, one part for classifying zeros and the other part for making predictions given the county has a positive HIV diagnosis rate. For each part of the model, we explored multiple methods, some of them take into account the spatial correlation and some do not. We compared both the classification and prediction performance between different methods, and found that making predictions based on our data does not benefit from models that consider spatial correlation. In particular, we found that the logistic regression for estimating the probability of positive rates in conjunction with a GEE with a log-link produced the best prediction results. Both logistic regression and the GEE model are easy to implement and computationally inexpensive, which allows for making predictions across the entire US at once.

Because counties are geographically indexed, we expect spatial autocorrelation exists between neighbors. However, spatial models did not show advantages, demonstrating that the covariates we employed alone might be sufficient to capture the spatial variability. Nevertheless, it is never our intention to generalize this conclusion to a new data set. We simply provided a case study to show the possibility that with abundant and informative covariates, spatial models could be unnecessary. However, given a new data set, we recommend a careful analysis to be performed to identify the most appropriate model.

In our analysis, the suppressed data was imputed and the sensitivity to the imputation was examined. We found our models to be insensitive to the suppressed value chosen, allowing for robust prediction estimates. While our model was effective at predicting HIV diagnosis rates for a majority of the counties, it was limited in its ability to capture outbreaks. If the primary interest is in outbreaks, data might need to be collected more frequently than annually. Alternatively, other measures with finer temporal resolution (e.g., Google's search engine data) may be more appropriate to anticipate sudden changes in HIV epidemiology.

There are many challenges with HIV data modeling, Table 4 summarizes these challenges and our associated findings. We assumed an AR structure for the HIV data and stationarity of the regression residuals against the covariates. It is challenging to verify this structure with our data due to its small sample size in time. However, the validity of the assumption should be revisited when more years of data are collected.

Table 4
A summary of the different challenges encountered when modeling HIV data, how we approached the problem, and our findings.

Challenge	Approach	Findings
Excess zeroes	Two-part model	The non-spatial logistic regression model seems more efficient than its spatial counter model for the classification. We chose to use a cutoff of 0.3 to classify rates as 0.
Suppressed data	Simple quantile imputation	This imputation could be improved with more sophisticated methods. However, simulation studies showed the prediction results are insensitive to the suppressed value chosen.
Spatial modeling	SSAL and SVAR model	No advantage is shown for using a spatial model despite the spatial autocorrelation likely exists between neighboring counties. We found that the autoregressive HIV rates along with the other covariates considered are sufficient at capturing the spatial variability of the data.
Temporal modeling	AR covariates	The prior year's HIV rates are significant for making HIV rate predictions. The prior year HIV data can be used directly as a covariate or in an AR model.
One year ahead prediction	GEE	We compared five different classical and modern spatial models for one year ahead prediction and found the non-spatial GEE model to be the best.

Acknowledgments

The authors thank the editor and the two anonymous reviewers for their constructive suggestions that have greatly improved the content and presentation of this article. Research reported in this publication was partially supported by the National Institute of Mental Health under Award Number R01MH114847, the National Institute on Drug Abuse under Award Number DP1 DA048570, the National Institute Of Allergy And Infectious Diseases under Award Number R01AI147487, and NSF grant DMS-1830312. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A

Two year ahead predictions

In reality, the current year's covariates and HIV data may not be available in time to make prediction of next year, so we also demonstrate model performance using two year ahead predictions. We predict new HIV diagnosis rates at $t+2$ as

$$\hat{Y}_{i,t+2} = \hat{\mathbb{E}}(Y_{i,t+2}|Y_{i,t+2} > 0, \mathbf{X}_{i,t})\hat{\mathbb{P}}(Y_{i,t+2} > 0|\mathbf{X}_{i,t}).$$

While before the covariates from year 2011 were used to predict diagnosis rates in 2012, the covariates are now predicting diagnosis rates in 2013. Therefore, our training data are now prediction years 2013–2015 and the testing data are years 2016–2018. Compared to the one year ahead predictions in Table 2, the prediction skills of all methods in Table A1 deteriorates except for the QUANT and SSAL model in year 2017. The EIR of the QUANT model also increases significantly for the two year ahead predictions compared to the one year ahead predictions. This seems to indicate the QUANT model is most robust when the covariates are less informative. Overall, the best performing model of the two year ahead predictions still appears to be a model that does not consider spatial correlation.

Table A1
Two year ahead prediction results of new HIV diagnosis rates. Coefficient estimates are assumed stationary in time and trained using prediction years 2013–2015.

	2016		2017		2018	
Model	MSPE	EIR	MSPE	EIR	MSPE	EIR
GEE	54.40	80.6%	40.75	81.4%	44.01	83.0%
QUANT	46.58	77.3%	35.68	78.8%	36.97	81.7%
SSAL	49.81	77.2%	40.56	79.0%	49.94	79.9%
DBN	47.48	79.6%	34.41	81.9%	36.55	83.2%
SVAR	68.73	80.2%	49.36	80.0%	50.89	81.9%

Time trend

When using data collected over time it is often important to consider a time trend. We incorporate a yearly time variable and evaluate model performance using the methods described in Section 3.3. The SVAR model was excluded from this study due to computation time, however we expect similar results.

Based on Table A2, in general the MSPE of the models had a slight decrease, while the EIR also decreased compared to Table 2. Overall, the impact on the prediction results is minimal. Shand et al. (2018) also found that the temporal random effects have little effect to improve the data modelling and thus prediction. This is most likely due to having such a short time series or the covariates already accounting for the time trend.

Non-stationary coefficients

We assumed the coefficient estimates for the models in Section 4 were stationary in time using years 2012–2014 as training data. To test this assumption we evaluate the model performance assuming non-stationary coefficient estimates and retraining the model each year. This means the models for prediction year 2016 were now trained using data from years 2013–2015, models for prediction year 2017 were trained using years 2014–2016, and models for prediction year 2018 were trained using years 2015–2017. Table A3 shows the prediction results in terms of MSPE and EIR.

Compared to the one year ahead stationary prediction results in Table 2, the predictive performance of the GEE model deteriorated while the QUANT and DBN models improved. The SSAL results are similar to the GEE model. Overall, the non-stationary QUANT model is comparable to the stationary GEE model in terms of MSPE, however, the stationary GEE model has superior EIR. We

Table A2
One year ahead prediction results of new HIV diagnosis rates with time trend. Coefficient estimates are assumed stationary in time and trained using prediction years 2012–2014.

	2015		2016		2017		2018	
Model	MSPE	EIR	MSPE	EIR	MSPE	EIR	MSPE	EIR
GEE	326.50	78.8%	45.83	77.3%	41.00	75.6%	34.35	79.3%
QUANT	328.37	69.1%	43.25	69.7%	37.18	67.4%	33.34	70.7%
SSAL	325.63	76.8%	44.53	76.1%	41.01	73.8%	33.35	78.0%
DBN	328.32	77.7%	42.30	76.9%	30.01	75.2%	34.64	79.1%

Table A3
One year ahead prediction results of new HIV diagnosis rates. Coefficient estimates are assumed non-stationary in time and re-trained each year using the prior 3 years of data as training data. Computation time is the average time for each year in seconds.

	2015		2016		2017		2018		
Model	MSPE	EIR	MSPE	EIR	MSPE	EIR	MSPE	EIR	Time
GEE	326.94	80.1%	46.84	81.6%	48.41	79.9%	43.63	86.1%	166.8
QUANT	328.61	69.2%	41.34	72.6%	34.63	71.6%	33.90	81.0%	165.2
SSAL	325.26	77.9%	46.63	78.7%	63.54	76.9%	39.56	83.4%	592.0
DBN	328.38	78.4%	40.78	78.8%	31.31	77.4%	33.59	78.4%	194.1

Table A4
One year ahead prediction results of new HIV diagnosis rates using different autoregressive lags as covariates. Coefficient estimates are assumed stationary in time and trained using prediction years 2012–2014.

		2015		2016		2017		2018	
Model	AR	MSPE	EIR	MSPE	EIR	MSPE	EIR	MSPE	EIR
GEE	AR1	332.32	79.7%	55.18	75.5%	39.53	74.6%	42.41	76.2%
QUANT	AR1	336.07	62.2%	151.48	62.4%	38.61	59.6%	37.67	62.2%
SSAL	AR1	333.85	77.5%	57.27	73.5%	44.97	73.0%	44.61	74.1%
DBN	AR1	331.51	78.5%	62.86	74.8%	34.41	73.7%	37.59	75.4%
GEE	AR2	328.61	79.2%	46.43	77.5%	56.97	75.8%	33.90	77.2%
QUANT	AR2	329.49	64.7%	44.14	65.2%	44.55	63.3%	33.86	65.0%
SSAL	AR2	325.33	77.1%	43.74	76.4%	40.42	73.5%	33.11	76.6%
DBN	AR2	328.67	79.2%	42.55	77.3%	43.44	75.8%	34.18	77.5%

choose our best performing model to be the stationary GEE due to the similar MSPE and better EIR.

Autoregressive HIV rate lag

The prior year's HIV rates are significant for making one year ahead HIV predictions. We consider different autoregressive lags as covariates in the model to determine the most appropriate number of lags to include in the variable selection. It would be difficult to use a partial autocorrelation function to determine the number of lags due to the short time series and suppressed data in many counties. Table A4 shows that if AR(1) is used as the only autoregressive HIV rate covariate there will be too much dependency on the prior year's rate. That means when there is an outlier, such as in 2015, the prediction in 2016 assumes the rate will be similar, resulting in overestimation. In general, including AR(2) HIV rate covariates is an improvement to AR(1) in all years except 2017. However, we found an AR(3) HIV rate covariate process to be best, allowing for smoothing and providing more robust estimates toward any potential outliers. The results of this chosen model are in Table 2. The SVAR model was excluded from this study due to computation time.

References

AHRF, 2019. Ahrf [internet] area health resources files. Cited 2020 September 7.

Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for surveynonresponse. Int. Stat. Rev. 78 (1), 40–64. doi:10.1111/j.1751-5823.2010.00103.x.

Aral, S.O., O'Leary, A., Baker, C., 2006. Sexually transmitted infections and hiv in the southern united states: an overview. Sex. Transm. Dis. 33 (7), S1–S5. doi:10.1097/01.olq.0000223249.04456.76.

Auchincloss, A.H., Gebreab, S.Y., Mair, C., Roux, A.V.D., 2012. A review of spatial methods in epidemiology 2000–2010. Annu. Rev. Public Health 33, 107–122. doi:10.1146/annurev-publhealth-031811-124655.

Bahn, V., O'Connor, R.J., Krohn, W.B., 2006. Importance of spatial autocorrelation in modeling bird distributions at a continental scale. Ecography 29, 835–844. doi:10.1111/j.2006.0906-7590.04621.x.

Belotti, F., Deb, P., Manning, W.G., Norton, E.C., 2015. Twopm: two-part models. Stata J. 15 (1), 3–20. doi:10.1177/1536867X1501500102.

Beretta, L., Santaniello, A., 2015. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Mak. 74 (16), 197–208. doi:10.1186/s12911-016-0318-z.

Besag, J., 1972. Nearest-neighbour systems and the auto-logistic model for binary data. J. R. Stat. Soc. Ser. B (Methodol.) 34 (1), 75–83. doi:10.1111/j.2517-6161.1972.tb00889.x.

Bivand, R., Pebesma, E., Gómez-Rubio, V., 2008. Applied spatial data analysis with r. Springer, New York, NY doi:10.1007/978-0-387-78171-6.

Bottai, M., Zhen, H., 2013. Multiple imputation based on conditional quantile estimation. Epidemiol. Biostat. Public Health 10 (1), e8758.

Caragea, P.C., Kaiser, M.P., 2009. Autologistic models with interpretable parameters. J. Agric. Biol. Environ. Stat. 14 (3), 281–300. doi:10.1198/jabes.2009.07032.

CDC, 2020. Cdc [internet] | basic statistics | hiv basics | hiv/aids. Cited 2020 Apr 28.

Chan, M.S., Lohmann, S., Morale, A., Zhai, C., Ungar, L.H., Holtgrave, D.R., Albarracn, D., 2018. An online risk index for the cross-sectional prediction of new hiv, chlamydia, and gonorrhea diagnoses across U.S. counties and across years. AIDS Behav. 22 (7), 2322–2333. doi:10.1007/s10461-018-2046-0.

Chen, S., 2014. Imputation of missing values using quantile regression. Graduate Theses Dissertat. 13924.

Cole, J.A., Sherriff, J.D.F., 1972. Some single- and multi-site models of rainfall within discrete time increments. J. Hydrol. (Amst.) 17, 97–113. doi:10.1016/0022-1694(72)90068-6.

Dietterich, T., 1995. Overfitting and undercomputing in machine learning. ACM Comput. Surv. 27 (3), 326–327. doi:10.1145/212094.212114.

Douthit, N., Kiv, S., Dwolatzky, T., Biswas, S., 2015. Exposing some important barriers to health care access in the rural usa. Public Health 129 (6), 611–620. doi:10.1016/j.puhe.2015.04.001.

Franzin, A., Sambo, F., Camillo, B.D., 2017. Bnstruct: an r package for bayesian network structure learning in the presence of missing data. Bioinformatics 33 (8), 1250–1252. doi:10.1093/bioinformatics/btw807.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander Jr., J.A., Latimer, A., Rebelo, A.G., 2006. Modelling species diversity through species level hierarchical modeling. J. R. Stat. Soc. Ser. C (Appl. Stat.) 54 (1), 1–20. doi:10.1111/j.1467-9876.2005.00466.x.

Geraci, M., McLain, A., 2018. Multiple imputation for bounded variables. Psychometrika 83 (4), 919–940.

Gonsalves, G.S., Crawford, F.W., 2018. Dynamics of the hiv outbreak and response in scott county, indiana, 2011–2015: a modeling study. Lancet HIV 5 (10), 569–577. doi:10.1016/S2352-3018(18)30176-0.

Goswami, N.D., Schmitz, M.M., Sanchez, T., et al., 2016. Understanding local spatial variation along the care continuum: the potential impact of transportation vulnerability on hiv linkage to care and viral suppression in high-poverty areas, atlanta, georgia. J. Acquir. Immune Defic. Syndr. 72 (1), 65–72. doi:10.1097/QAI.0000000000000914.

Gray, S.C., Massaro, T., Chen, I., Edholm, C.J., Grotheer, R., Zheng, Y., Chang, H.H., 2016. A county-level analysis of persons living with hiv in the southern united states. AIDS Care 28 (2), 266–272. doi:10.1080/09540121.2015.1080793.

Hanna, D.B., Selik, R.M., Tang, T., Gange, S.J., 2012. Disparities among states in hiv-related mortality in persons with hiv infection, 37 u.s. states, 2001–2007. AIDS 26 (1), 95–103. doi:10.1097/QAD.0b013e32834dcf87.

Harrison, K.M., Ling, Q., Song, R., Hall, H.I., 2008. County-level socioeconomic status and survival after hiv diagnosis, United States. Ann. Epidemiol. 18 (12), 919–927. doi:10.1016/j.annepidem.2008.09.003.

Hoeting, J.A., 2009. The importance of accounting for spatial and temporal correlation in analyses of ecological data. Ecol. Appl. 19 (3), 574–577. doi:10.1890/08-0836.1.

Hoeting, J.A., Davis, R.A., Merton, A.A., Thompson, S.E., 2006. Model selection for geostatistical models. Ecol. Appl. 16 (1), 87–98. doi:10.1890/04-0576.

Hooten, M.B., Wikle, C.K., 2008. A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. Environ. Ecol. Stat. 15, 59–70. doi:10.1007/s10651-007-0040-1.

Huber, P.J., 1967. The behavior of maximum likelihood estimation under nonstandard conditions. Proc. Fifth Berkeley Sympos. Math. Stat. Probab. 1, 221–233.
Hughes, J., Haran, M., Caragea, P., 2011. Autologistic models for binary data on a lattice. Environmetrics 22 (7), 857–871. doi:10.1002/env.1102.
Jones, J., Grey, J.A., Purcell, D.W., Bernstein, K.T., Sullivan, P.S., Rosenberg, E.S., 2018. Estimating prevalent diagnoses and rates of new diagnoses of hiv at the state level by age group among men who have sex with men in the united states. Open Forum Infect. Dis. 5 (6), 1–8. doi:10.1093/ofid/ofy124.
Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Divers. Distrib. 13 (1), 66–69. doi:10.1111/j.1472-4642.2006.00293.x.
Li, B., Eriksson, M., Srinivasan, R., Sherman, M., 2008. A geostatistical method for texas nexrad data calibration. Environmetrics 19 (1), 1–19. doi:10.1002/env.848.
Little, R.J.A., Rubin, D.B., 2014. Statistical analysis with missing data. Hoboken: Wiley doi:10.1002/9781119013563.
Margaritas, D., 2003. Learning bayesian network model structure from data. Theses Dissertat.–School Comput. Sci. at CMU.
Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G., 2011. Review of statistical methods for analysing healthcare resources and costs. Health Econ. 20 (8), 897–916. doi:10.1002/hec.1653.
Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1), 17–23. doi:10.2307/2332142.
Pellowski, J.A., Kalichman, S.C., Matthews, K.A., Adler, N., 2013. A pandemic of the poor: social disadvantage and the u.s. hiv epidemic. Am. Psychol. 68 (4), 197–209. doi:10.1037/a0032694.
Rosenberg, E.S., Grey, J.A., Sanchez, T.H., Sullivan, P.S., 2016. Rates of prevalent hiv infection, prevalent diagnoses, and new diagnoses among men who have sex with men in us states, metropolitan statistical areas, and counties, 2012–2013. JMIR Public Health Surveillanc.e 2 (1), e22. doi:10.2196/publichealth.5684.
Rotnitzky, A., Wypij, D., 1994. A note on the bias of estimators with missing data. Biometrics 50 (4), 1163–1170. doi:10.2307/2533454.
Sanchez, T.H., Kelley, C.F., Rosenberg, E., et al., 2014. Lack of awareness of human immunodeficiency virus (hiv) infection: problems and solutions with self-reported hiv serostatus of men who have sex with men. Open Forum Infect. Dis. 1 (2). doi:10.1093/ofid/ofu084.
Schabenberg, O., Gotway, C. A., 2005. Statistical methods for spatial data analysis.
Scutari, M., 2010. Learning bayesian networks with the bnlearn r package. J. Stat. Softw. 35 (3), 1–22. doi:10.18637/jss.v035.i03.
Shand, L., Li, B., Park, T., Albarracin, D., 2018. Spatially varying auto-regressive models for prediction of new human immunodeficiency virus diagnoses. J. R. Stat. Soc. Ser. C(Appl. Stat.) 67 (4), 1003–1022. doi:10.1111/rssc.12269.
Srinivasan, S., 2015. Spatial regression models. Springer, Cham. doi:10.1007/978-3-319-23519-6_1294-2.
Trepka, M.J., Niyonsenga, T., Maddox, L., Lieb, S., Lutfi, K., Pavlova-McCalla, E., 2013. Community poverty and trends in racial/ethnic survival disparities among people diagnosed with aids in florida, 1993–2004. Am. J. Public Health 103 (4), 717–726. doi:10.2105/AJPH.2012.300930.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P.O., et al., 2001. Missing value estimation methods for dna microarrays. Bioinformatics 17 (6), 520–525. doi:10.1093/bioinformatics/17.6.520.
Vaughan, A.S., Rosenberg, E., Shouse, R.L., Sullivan, P.S., 2014. Connecting race and place: a county-level analysis of white, black, and hispanic hiv prevalence, poverty, and level of urbanization. Am. J. Public Health 104 (7), 77–84. doi:10.2105/AJPH.2014.301997.
Waller, L.A., Goodwin, B.J., Wilson, M.L., Ostfeld, R.S., Marshall, S., Hayes, E.B., 2007. Spatio-temporal patterns in county-level incidence and reporting of lyme disease in the northeastern united states, 1990–2000. Environ. Ecol. Stat. 14, 83–100. doi:10.1007/s10651-006-0002-z.
Wang, Z., 2012. Analysis of binary data via spatial-temporal autologistic regression models. Theses Dissertat.–Stat. 3.
Weins, J., Wallace, B., 2016. Editorial: special issue on machine learning for health and medicine. Mach. Learn. 102 (3), 305–307. doi:10.1007/s10994-015-5533-9.
Zeglin, R.J., Stein, J.P., 2015. Social determinants of health predict state incidence of hiv and aids: a short report. AIDS Care 27 (2), 255–259. doi:10.1080/09540121.2014.954983.