# Modeling HIV Mortality with a Spatiotemporal Bayesian Approach

**- explaining and extending the approach of Aktekin and Musal (2015) -**

by
Jia Yan Ng (3035599),
Lena Schwertmann (3035605),
Thummaporn Nimpanomprasert (3035607)

**Module:** Probabilistic Modeling (# 82005000)

**Examiner:** Prof. Burkhardt Funk

**Date of Submission:** July 31, 2019

Management & Data Science
Master of Science
Leuphana University of Lüneburg

# Contents

# 1 Introduction

Bayesian modeling allows for the modeling of real-world phenomena including quantitative statements about the (un)certainty of results. By using data, prior knowledge and plausible assumptions, it is attempted to find the influencing factors that mainly determine the model outcome.
This report takes a look at death cases due to the human immunodeficiency virus (HIV) and the related acquired immunodeficiency syndrome (AIDS). The approach is adopted from Aktekin and Musal (2015) who first introduced their idea in 2012 (Musal and Aktekin, 2012). The more recent paper extends the previous approach, considering the occurring deaths in New York State (United States of America) in a 5 year time period (2000-2004). Their work is motivated by enabling informed decision-making, e.g. in regard to allocating limited resources to high-risk areas to prevent deaths by HIV and acquired immune deficiency syndrome (AIDS).
It is investigated how the model covariates poverty and income inequality and spatiotemporal effects influence the model outcome and model fit. The poverty level and the degree of income inequality in a given area provide valuable insight into the economic situation and have been shown to be relevant explanatory variables of HIV mortality (Harrison et al., 2008; LaMontagne and Stockemer, 2010). Aktekin and Musal (2015) use different versions of the same basic inequality measure (Theil index) to show that there are different ways to define inequality that also influence model fit.
Spatial and temporal effects are included as random effects, but also using a conditional autoregressive (CAR) model structure. CAR models can also be used to model contagion of diseases, but this is not the reason why it is used here, as HIV is a non-contagious disease. Aktekin and Musal (2015) explain the use of the CAR model structure with effects of neighbouring counties on each other that go beyond the covariates that are used explicitly. Following this logic, this enables the model to capture more of the variance present in the data.
The HIV mortality data itself also has a particularity that influences the choice of model structure. The data is available on county level, which are the administrative subregions of a US state. NY State consists of 62 counties. Many counties do not have any occurrences of HIV-related deaths and thus no existing HIV risk, which results in zero-inflated count data. To model this appropriately, the zero-inflated Poisson (ZIP) model is used which is well-documented in the literature (Agarwal et al., 2002; Zuur et al., 2012).

In what follows, the approach presented above is reproduced using data obtained from public sources and Markov Chain Monte Carlo (MCMC) sampling is used to estimate the model parameters. Our focus in this report lies on explaining the modeling approach in an understandable way, where Aktekin and Musal (2015) did not provide comprehensive explanations assuming prior knowledge of the subjects.
This also includes discussing the quality of the MCMC samples. The convergence of

MCMC chains and the representativeness of the samples of the posterior distribution are discussed on the basis of diagnostic plots and measures provided by the sampling software. Then, a variety of candidate models of different complexity and with different covariate combinations are evaluated. To extend this approach, we introduce an additional covariate, the unemployment rate. This measure, just like the poverty level, has been identified as a social determinant of health (Wilkinson et al., 2003), and might thus also be an informative explanatory variable of HIV mortality. Accordingly, models including the unemployment rate are also estimated and discussed in our approach. Aktekin and Musal (2015) were able to identify a specific inequality measure, $T^4$, to consistently deliver the best fit models and discussed which covariates influenced HIV mortality in what way.
This report will follow the basic structure of Aktekin and Musal (2015) accordingly, comparing the reproduced results with the original work.

# 2 Material and Methods

## 2.1 Data

The data was retrieved from sources similar to the ones mentioned in Aktekin and Musal (2015) and is generally slightly different than the one used in the original data. Specific details of these differences are only mentioned here partly, a detailed list can be found in the appendix. The CSV file that contains all necessary data can be found in our repository (`https://github.com/Giant316/bayesianinference_HIVmortality/tree/master/input_files`).
The multiple-cause mortality data set is one of the main data sources as it provides the observations for the target variable that we want to predict - the number of HIV deaths. We consider data for New York state between 2000 and 2004. The data is given by the National Bureau of Economic Research (`https://www.nber.org/data/vital-statistics-mortality-data-multiple-cause-of-death.html`) which contains all causes of death for the deaths occurring within the United States. HIV as a cause of death is defined as B20-B24 in the International Classification of Diseases (ICD-10) system (Centers for Disease Control and Prevention, October 2002). Each record is based on information from death certificates of each state and county with The Federal Information Processing Standard Publication (FIPS) code and Vital Statistics code. The FIPS code for NY state is 36 while the Vital Statistics code for NY state is 33. That means that the 5-digit FIPS code for the counties belonging to NY State have 36 as their first two digits. We only consider the deaths of people occurring and residing in NY State. The zero count counties are not included in this list. Over five years, we determine 2,190, 2,028, 1,920, 1,844 and 1,675 death counts for NY state.

The number of people living with HIV and AIDS (NPLWHA) at the end of 2003 are defined as the population of interest. From the modelling perspective, this is used to calculate the expected number of deaths $E_i$, as people living with HIV and AIDS con-

tribute directly to positive HIV death counts. This data is collected annually for each county, but is at present only available starting from 2007 (`https://www.health.ny.gov/diseases/aids/general/statistics/annual/index.htm`). Thus, we obtained the values from figure 5 (c) in Aktekin and Musal (2015) for all counties except the 5 New York City counties. For these counties the data was available for the end of 2003 from the New York City health department (NYC Health, 2019).

For calculation of the covariates, data from the Census 2000 in the United States of America was retrieved using the advanced search within the American FactFinder page (`https://factfinder.census.gov`) which is maintained by the US Census Bureau. The database was filtered for files providing information on zip code tabulation area (ZCTA) level for New York State. The ZCTAs are sub-units of the 62 counties (the 5 boroughs of New York city are also considered as counties). The ZCTAs are similar to zip code areas used by the US Postal Service based on street addresses. By contrast, the ZCTAs are created by the US Census Bureau in polygonal shape and can also cross county borders, but have a unique, 5-digit identifier (US Census Bureau, 2019).
Three different files were used to obtain all the necessary variables:

1. **DP-3** "Profile of Selected Economic Characteristics" (Dataset: SF3 Sample Data)
   - poverty (number of individuals below poverty level)
   - per-capita income (in dollars)
   - number of individuals in labor force (16 years and older)
   - number of unemployed individuals (in labor force, 16 years and older)
2. **P001** "Total Population" (Dataset: SF1 100% Data)
   - population (number of individuals)
3. **P002** "Urban and Rural" (Dataset: SF1 100% Data)
   - number of urban population (number of individuals in urban area or cluster)

In total, 1677 ZCTAs were associated with New York State for Census 2000. As it is assumed that HIV mortality is most prevalent in urbanized areas, only inhabited ZCTAs with a minimum urban population of 10 are considered as relevant for covariate calculation. Thus, we excluded 49 "only water", 9 "only land" ZCTAs and 686 ZCTAs with $< 10$ people living in an urban area or cluster. This definition is different than the one used in Aktekin and Musal (2015), where urban ZCTAs are classified as having "1000 and 500 people per square mile" resulting in 957 ZCTAs considered. As this data was not available to us, we arrived at a final number of 933 urban ZCTAs using the alternative definition described above. This leads to Hamilton County being excluded from the entire analysis as it does not contain any urban ZCTAs.

To associate the ZCTAs with their respective counties, the Missouri Census Data Center (MCDC) Data Applications provided the connecting data source to access the MABLE09 geographic with Census2000 to get the correlation between ZCTAs and counties (`http://`

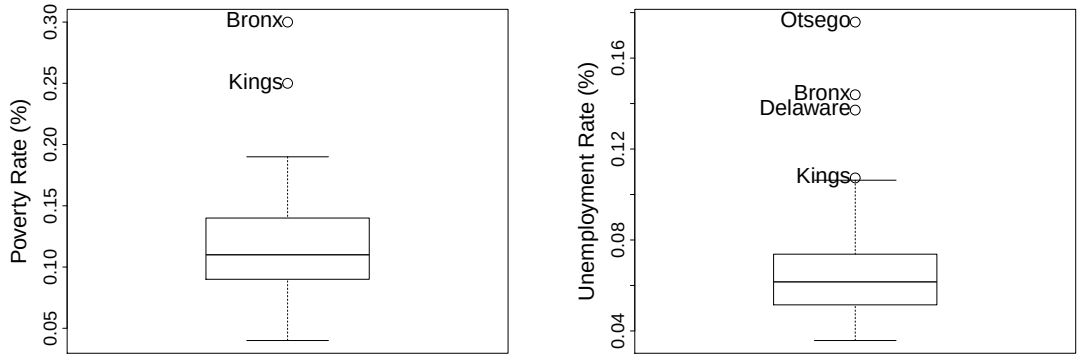
mcdc.missouri.edu/applications/geocorr2000.html). The application has a feature to select NY State as a criteria and mapping county as a geographies source to ZIP/ZCTA as the target geographies on an interface to proceed to requested data set. The data set contains 2,070 rows in total. When filtering for the 933 urban ZCTAs considered, 1,056 rows remain, showing that some ZCTAs are associated with more than one county.
This issue was not addressed by Aktekin and Musal (2015), but these ZCTAs have a high potential to distort the data as they can not be treated as individual ZCTAs. To give an example, for a ZCTA that is located at the intersection of three counties, all of its variable values would be counted thrice. For the population variable, this would result in an artificial population increase of $\approx 2$ million people. To avoid this, all values from the variables "population", "individuals in poverty" and "per-capita income" were divided by the number of counties the respective ZCTA belongs to. This ensures that the overall sum of the variables does not change. This slightly distorts the data as no data regarding the population/income distribution within ZCTAs is available.

The Consumer Price Index (CPI) is obtained from Olsen et al. (2012) to calculate the CPI-adjusted inequality measure $T^4$ (for details, see chapter 2.2.3). This source covers the period 1982 through 2012 and was used instead of the older estimate mentioned in Aktekin and Musal (2015). The revised data contains price indices for metropolitan and non-metropolitan area that existed between 1971 and 2012 for all areas in the United State. A few minor errors from the previous estimate was corrected by the authors. As a result, a small variation the CPI value from the original paper is expected. Similar to the methodology in Aktekin and Musal (2015), the 5 boroughs in New York state, namely, Richmond, Queens, New York, Kings and Bronx are treated as separate counties. With the other 57 counties, only Hamilton County's CPI is assigned to be zero since it has been excluded from our analysis. Since the CPI measurement in Olsen et al. (2012) was only done for metropolitan and non-metropolitan areas, we have to link their respective CPIs to the counties they are located in. There are 20 metropolitan areas for New York state found in this CPI dataset. The remaining counties in New York are regarded as non-metropolitan areas and hence we set their CPI to be 1.042. As for the 5 boroughs of NY City, their CPIs are set to be 1.223 since they are categorized as the NY primary metropolitan statistical areas.

## 2.2 Calculation of Covariates

### 2.2.1 Poverty

The poverty covariate used is defined as the population share of individuals that are living in poverty (Aktekin and Musal, 2015). To calculate this value for each county, the population and poverty data that was obtained from the Census 2000 data for each zip code tabulation area (ZCTA) is aggregated on county level. Figure 1 (a) shows a boxplot of the poverty level for all counties. The counties Bronx and Kings (Brooklyn), both




**(a) Poverty Rate** **(b) Unemployment Rate**

**Figure 1:** Boxplots showing the distribution of (a) the poverty rate and (b) the unemployment rate for all counties of NY State.

located in NY city are the outliers having the highest poverty rate with 30% and 25%, respectively. The mean poverty is ca. 10%.

### 2.2.2 Unemployment Rate

The unemployment rate is defined as the share of unemployed individuals which are in the labor force. Similar to the calculation of the poverty covariate, the number of individuals in labor force and the total number individuals in labor force obtained from Census 2000 are aggregated on county level. Figure 1 (b) shows the unemployment rate for all counties. Bronx and Kings (Brooklyn) county are also among the highest outliers, like for the poverty level, but Otsego county has the highest unemployment rate of ca. 18%.

### 2.2.3 Inequality Measures

Income inequality has been shown to be an informative measure regarding the modeling of HIV mortality (LaMontagne and Stockemer, 2010). While the Gini coefficient is the best-known and most-used measure, the Theil index offers the advantage of its decomposability (Conceição and Ferreira, 2000).
In a simplified way, this means that the Theil index of a region of interest (here: county) is composed of the sum of the Theil indices of all its subcomponents (here: ZCTAs). The variables needed for the calculation are the population and per-capita income obtained from Census 2000 data. Aktekin and Musal (2015) created four different variations of the Theil index, named $T^1 - T^4$, in order to find out which one results in the best model fit (see equations (1) - (4)). Despite their differences, $T^1 - T^3$ have the same basic structure: The income share of the respective ZCTA multiplied with the logarithm of the ratio of the income and the population share. Due to the logarithm, the Theil index of a ZCTA can be either negative, positive or equal to zero, with each of the possibilities showing a different kind of inequality. For instance, a positive value occurs when the income share

is larger than the population share, indicating that there are relatively few people with a high income. If the Theil index is zero, this implies equality. Thus, its quantified notion of equality is that the income share and population share should be equally large.
The four inequality measures mainly differ in regard to how their income and population shares are calculated. Before the calculations are presented, it needs to be stressed that the measures are not the same ones that Aktekin and Musal (2015) used. While trying to reproduce the boxplots of inequality measures they presented in figure 2, it became apparent that the formulas for $T^1$, and $T^2$ need to be switched to achieve the same results. This results in the following income and population shares being used in a slightly different way.
The term $s.inc_i^j$ describes the relative income share of a ZCTA $i : \{1, ..., Z_j\}$ located in county $j : \{1, ..., J\}$ that has $Z_j$ ZCTAs. As $pci_i^j$ is the per-capita income and $pci_i^j$ is the population, $pci_i^j pop_i^j$ describes the income located in a ZCTA. The important note here is that the income share $s.inc_i^j$ and the population share $s.we_i^j$ which are used for the calculation of $T^1$, are seen in respect to the accumulated income or population of the entire state of NY. To rephrase, they provide the answer to the question: "Which percentage of NY state's population/income is located in this ZCTA?"

$$s.inc_i^j = \frac{pci_i^j pop_i^j}{\sum_{j=1}^{i=J} \sum_{i=1}^{i=Z_j} pci_i^j pop_i^j} \qquad s.we_i^j = \frac{pop_i^j}{\sum_{j=1}^{i=J} \sum_{i=1}^{i=Z_j} pop_i^j}$$

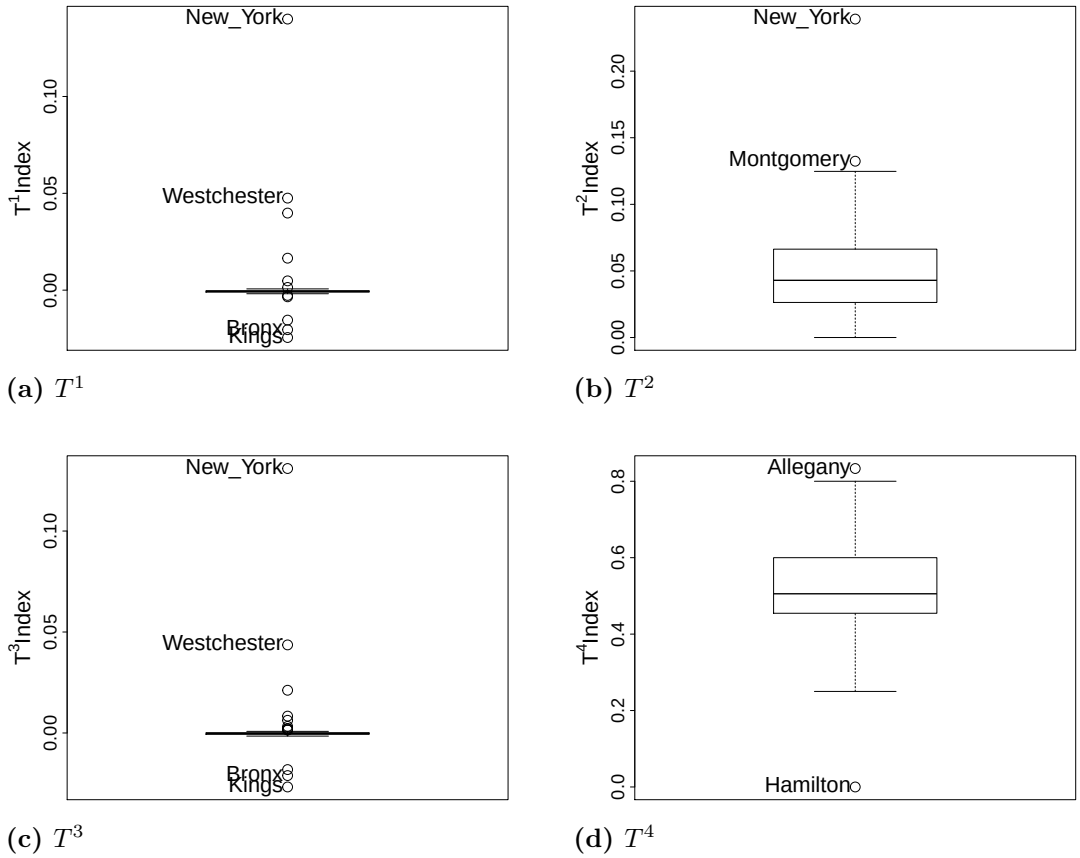
$T^2$, is calculated in a very similar way, except that here the denominator of the income share $c.inc_i^j$ and the population share $c.we_i^j$ only accumulates the variables for the respective county. Accordingly, the shares for $T^2$ refer to the respective county $j$ and not the entire state of NY as it is the case for $T^1$.

$$c.inc_i^j = \frac{pci_i^j pop_i^j}{\sum_{i=1}^{i=Z_j} pci_i^j pop_i^j} \qquad c.we_i^j = \frac{pop_i^j}{\sum_{i=1}^{i=Z_j} pop_i^j}$$

$T^3$ represents the county price index (CPI)-adjusted $T^1$ inequality measure, where the income share $s.inc_i^{j,CPI}$ is used instead. To that end, each $pci_i$ is divided by the respective county's CPI.

$$s.inc_i^{j,CPI} = \frac{\frac{pci_i^j}{CPI_i^j} pop_i^j}{\sum_{j=1}^{i=J} \sum_{i=1}^{i=Z_j} \frac{pci_i^j}{CPI_i^j} pop_i^j}$$

Finally, $T^4$ is defined as the percentage of ZCTAs that have a negative value for $T^2$ within county $j$. As this measure is a ratio, its values range between 0 and 1.
Below, the equations 1-4 specify the formulas for the inequality measures $T^1$-$T^4$:




**(a)** $T^1$ **(b)** $T^2$

**(c)** $T^3$ **(d)** $T^4$

**Figure 2:** Boxplots of $T^1$ - $T^4$

$$T_j^1 = \sum_{i=1}^{Z_j} s.inc_i^j \log \frac{s.inc_i^j}{s.we_i^j} \tag{1}$$

$$T_j^2 = \sum_{i=1}^{Z_j} c.inc_i^j \log \frac{c.inc_i^j}{c.we_i^j} \tag{2}$$

$$T_j^3 = \sum_{i=1}^{Z_j} s.inc_i^{j,CPI} \log \frac{s.inc_i^{j,CPI}}{s.we_i^j} \tag{3}$$

$$T_j^4 = \sum_{i=1}^{Z_j} I(T_{i,j}^2 < 0)/Z_j \tag{4}$$

Where $I(.)$ is the indicator function that is equal to 1 when the $T^2$ value of the respective ZCTA is negative and otherwise equal to zero.

Figure 2, together with table 1 show how the four variations of the Theil index result in a different assessment of the income inequality on NY State's counties. The near-perfect correlation of 0.99 between $T^1$ and $T^3$ is also evident from figure 2, where the boxplots look quite identical, with New York State (Manhattan) having the highest positive and

**Table 1:** Correlation between the four different inequality measures.

| | **T1** | **T2** | **T3** | **T4** |
|---|---|---|---|---|
| **T1** | 1 | 0.62 | 0.99 | -0.06 |
| **T2** | | 1 | 0.63 | 0.14 |
| **T3** | | | 1 | -0.06 |
| **T4** | | | | 1 |

Kings County (Brooklyn) having the highest negative income inequality. According to the $T^2$ measure, again New York county (Manhattan) has the highest income inequality. Generally, $T^2$ is different from $T^1$ and $T^3$ because its values are either zero or positive. The correlation with $T^1$ (0.62) and $T^3$ (0.63) also shows that $T^2$ works differently. As $T^4$ is a ratio, thus its correlation with the other three inequality measures is very low, also emphasized by Allegany county being the highest outlier instead of New York (Manhattan).

## 2.3 Model Structure

The existing HIV-mortality data have been shown to be zero-inflated by Musal and Aktekin (2012), meaning that the number of zeros occurring in the data is higher than how it is predicted by the random variable of the Poisson model. To account for this, a binomial point mass at zero is introduced, such that with a probability of $1 - p_{it}$, zeros are sampled from a binomial distribution and with a probability $p_{it}$, zeros are sampled from a Poisson distribution. This mixture model is called the zero-inflated Poisson (ZIP) model that can distinguish between two types of 0s: structural and non-structural. Structural zero counts are considered as the incidences where there are no HIV deaths in regions with no prevailing HIV/AIDS occurrences. They are denoted by $1 - p_{it}$ in the model, where $p_{it}$ is the probability of a non-zero HIV death count. In contrast to that, non-structural zero counts are those that occur by pure chance. They are modeled as $p_{it} \exp -\lambda_{it}$. This leads to the following model structure of the ZIP model:

$$P\left(N_{it} = n_{it} | p_{it}, \lambda_{it}\right) = 1 - p_{it} + p_{it} \exp\left(-\lambda_{it}\right) \quad \text{for } n_{it} = 0 \text{ (zero count)} \tag{5}$$

$$P\left(N_{it} = n_{it} | p_{it}, \lambda_{it}\right) = p_{it} \frac{\exp\left(-\lambda_{it}\right) \lambda_{it}^{n_{it}}}{n_{it}!} \quad \text{for } n_{it} > 0 \text{ (non-zero count)} \tag{6}$$

where $N_{it}$ is the number of HIV deaths in county $i : \{1, \ldots, M\}$ at time $t : \{1, \ldots, T\}$.
The main model parameters $p_{it}$ and $\lambda_{it}$ come from different distributions, highlighting that the ZIP model is a mixture model. The binomial process is the component modeling a point mass at zero determined by $p_{it}$ while the remaining zero and non-zero counts are modeled by a Poisson distribution with the expected Poisson count $\lambda_{it}$.
Using these parameters, the expected value of $N_{it}$ and its variance $V$ can be calculated in the following way:

$$E\left(N_{it}|p_{it},\lambda_{it}\right)=p_{it}\lambda_{it} \tag{7}$$

$$V\left(N_{it}|p_{it},\lambda_{it}\right)=p_{it}\lambda_{it}+p_{it}\left(1-p_{it}\right)\lambda_{it}^{2} \tag{8}$$

Aktekin and Musal (2015) showed and emphasized that using the expected value $E_i$ alone does not explain the observed HIV mortality sufficiently. This is why the covariates, the conditional autoregressive (CAR) model, as well as regional and temporal random effects are used to capture more heterogeneity of the data. The covariates are introduced with a matrix $X_{ij}$ and a respective regression coefficient $\alpha_j$. The heterogeneity between counties based on the CAR model is included as the parameter $\beta_i^{(t)}$ in the Poisson model structure represented by $\lambda_{it}$. As it is apparent from the indices, $\beta_i^{(t)}$ is nested in time, as it was proposed by Waller et al. (1997):

$$\log\left(\lambda_{it}\right)=\log\left(E_i\right)+\sum_{j=1}^{J}\alpha_j X_{ij}+\beta_i^{(t)} \tag{9}$$

The log-link function which is used for $\lambda_{it}$ is the most common one for regression models using the Poisson distribution. It is necessary as it matches the continuous linear regression space with the range $[-\infty,+\infty]$ on the RHS of the equation with the positive, discrete range on the LHS of the equation that is characteristic for count data.
To recall, the CAR model structure is used to capture the effects of unobserved covariates. This is only included in the model structure for $\lambda_{it}$ so it does not affect $p_{it}$. What the CAR model does is that it assigns neighbouring counties a higher correlation to each other than more distant ones. This includes the prior assumption in the model that neighbouring regions affect each other's HIV mortality.
The specification of the CAR model structure contains the following components:

1. the adjacency matrix containing the values $w_{ij}$ (if the entry is non-zero, the two respective counties are neighbours),
2. the precision parameter $\tau^{(t)}$ which is time-dependent and therefore (potentially) different for each year,
3. the $\beta_i$ values of adjacent counties,
4. the constraint that $\sum_{i=1}^{M}\beta_i^{(t)}=1$ for all t to achieve a more reliable MCMC implementation.

In that way, each $\beta_i$ - meaning each county - receives its own conditional distribution of the CAR spatial effects. Following the notation in Aktekin and Musal (2015), the CAR prior is specified in this way:

$$\left(\beta_1^{(t)},\ldots,\beta_N^{(t)}|\tau^{(t)}\right)\propto\left(\tau^{(t)}\right)^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}\sum_{j<i}w_{ij}\left(\beta_i^{(t)}-\beta_j^{(t)}\right)^2\right\} \tag{10}$$

where $w_{ij} = 0$ for counties that are not neighbours and $w_{i.} = \sum_{j=1}^{N} w_{ij}$ for neighbouring counties. As equation 10 shows, there is a relation between the $\beta_i$s for each $t$ which according to Besag and Kooperberg (1995) follows as:

$$\left(\beta_i^{(t)} | \beta_{-i}^{(t)}, \tau^{(t)}\right) \sim N\left(\overline{\beta}_i^{(t)}, 1/\tau^{(t)} w_{i.}\right) \tag{11}$$

where $\beta_{-i}^{(t)} = \{\beta_j^{(t)} : j \neq i\}$ and $\overline{\beta}_i^{(t)}$ is the mean of the values of $\beta_i$ for the neighbours of county $i$ at time $t$.

The probability of a non-zero HIV death count defined as $p_{it}$ is also explained by a more complex structure than just the expected mortality $E_i$. As Agarwal et al. (2002) pointed out that using the CAR prior here causes an unstable model fit, temporal random effects $c_t$ and regional random effects $d_i$ are used instead. The covariates are included in the same way as for $\lambda_{it}$: using the covariate matrix $X_{ij}$ and a respective regression coefficient $a_j$. This results in the following model structure:

$$\text{logit}\left(p_{it}\right) = \log\left(E_i\right) + c_t + d_i + \sum_{j=1}^{J} a_j X_{ij} \tag{12}$$

where $\text{logit}(p_{it}) = \log(\frac{p_{it}}{1-p_{it}})$. The logit structure is used because it maps the probability values $p_{it}$ having a $[0, 1]$ range on the LHS of the equation to the continuous linear regression space with the range $[-\infty, +\infty]$ on the RHS of the equation.

## 2.4 Parameter Estimation with MCMC Sampling

### 2.4.1 Prior Specification

Markov Chain Monte Carlo (MCMC) methods are used to estimate the model parameters. The prior distributions of the model parameters are away to introduce prior knowledge about the model parameters, e.g. which values are most likely. The covariates' regression coefficients $\alpha_j$ and $a_j$ contained in the model structure for $\lambda_{it}$ and $p_{it}$, respectively, are assumed to be normally distributed using the prior $N(0, 0.001)$ for all $j$. For the random temporal and regional effects on the structure of $p_{it}$ the same prior was assumed, so $c_t, d_i \sim N(0, 0.001)$. These priors represent very informed priors centered at zero, which implies that the respective parameters are not assumed to influence the HIV mortality much before the inference process. For the precision parameter $\tau^{(t)}$ of the CAR model, the gamma-distributed prior $G(0.001, 0.001)$ is assumed which also represents an informed prior with a high frequency of zero values.
Prior sensitivity was not explored, as Aktekin and Musal (2015) concluded that the priors mentioned above do not show sensitive behaviour, as the inference results did not change significantly with stronger priors.

### 2.4.2 Candidate Models

To recall, the aim of the modeling approach is to find the most informative covariates for modeling HIV mortality. To that end, all possible combinations of the covariates used by Aktekin and Musal (2015) are used by various candidate models, which are specified below. The models increase in complexity, ranging from a model with no covariates to the four models using both covariates, plus an interaction term:

(a) **MNoCov:** the benchmark model with no covariates
(b) **Mp:** model with poverty
(c) **M1i:** model with $T^1$
(d) **M2i:** model with $T^2$
(e) **M3i:** model with $T^3$
(f) **M4i:** model with $T^4$
(g) **M1:** model with poverty and $T^1$
(h) **M2:** model with poverty and $T^2$
(i) **M3**: model with poverty and $T^3$
(j) **M4:** model with poverty and $T^4$
(k) **M1inter:** like M1, plus an interaction term
(l) **M2inter:** like M2, plus an interaction term
(m) **M3inter:** like M3, plus an interaction term
(n) **M4inter:** like M4, plus an interaction term

Using the unemployment rate as a third covariates results in many more potential combinations of covariates, as well as different possibilities for the interaction term. As modeling of all possible models was not feasible due to time constraints, the best fit inequality measure is chosen as the one that consistently results in the lowest DIC value for the candidate models of Aktekin and Musal (2015) listed above. Thus, the extension models using the unemployment rate are only run using this best covariate ($\mathrm{T}^{\mathrm{best}}$):

(a) **ME:** model with unemployment rate
(b) **MpE:** model with poverty and unemployment rate
(c) $\mathbf{MT_{best}iE}$**:** model with $\mathrm{T}^{\mathrm{best}}$ and unemployment rate
(d) $\mathbf{MT^{best}E}$**:** model with $\mathrm{T}^{\mathrm{best}}$, poverty and unemployment rate
(e) $\mathbf{MT^{best}EinterpE}$**:** like $\mathrm{MT}^{\mathrm{best}}\mathrm{E}$, plus an interaction term between poverty and unemployment rate
(f) $\mathbf{MT^{best}EinterT^{best}E}$**:** like $\mathrm{MT}^{\mathrm{best}}\mathrm{E}$, plus an interaction term between $\mathrm{T}^{\mathrm{best}}$ and unemployment rate
(g) $\mathbf{MT^{best}EinterT^{best}pE}$**:** like $\mathrm{MT}^{\mathrm{best}}\mathrm{E}$, plus an interaction term between $\mathrm{T}^{\mathrm{best}}$ poverty and unemployment rate

### 2.4.3 Implementation in WinBUGS

In order to estimate the coefficients of the covariates of interest, we use samples generated from posterior distribution by doing posterior simulation in WinBUGS, the Windows version of BUGS (Bayesian Inference Using Gibbs Sampling). It is relatively straightforward to implement the required distributions of our model using WinBUGS as compared to another general-purpose modelling package JAGS ("Just Another Gibbs Sampler"), since the later require a standalone module implementation which demand knowledge of C++ in writing the custom extensions. (Wabersich and Vandekerckhove, 2013; Spiegelhalter et al., 2002). Therefore, the posterior simulation was done in WinBUGS just as the original paper.

The model structure described in chapter 2.3 results in the following likelihood function:

$$L(\mathbf{p}, \boldsymbol{\lambda}; \mathbf{N}) \prod_{t=1}^{T} \left[ \prod_{i=1}^{M} \{1 - p_{it} + p_{it} \exp(-\lambda_{it})\}^{\mathbf{I}(n_{it}=0)} \left\{ p_{it} \frac{\exp(-\lambda_{it}) \lambda_{it}^{n_{it}}}{n_{it}!} \right\}^{1-\mathbf{I}(n_{it}=0)} \right] \quad (13)$$

where **I()** is the indicator function and the $\lambda_{it}$s and $p_{it}$s are specified further by equations (9) and (12).

The likelihood function, the model structure for $p_{it}s$ and $\lambda_{it}$ are specified in the WinBUS code that is also available here: (`https://github.com/Giant316/bayesianinference_HIVmortality`) The CAR structure for the spatial effects between the counties for each year as expressed in equation 11 is implemented using the `car.normal` function. The adjacency matrix specifying the neighbouring counties for each county is taken form Aktekin and Musal (2015). Other priors distribution as described in section 2.4.1 are implemented using built-in functions in WinBUGS.

On the other hand, we can use the "zero trick" to implement a sampling distribution that is not within the standard distribution in WinBUGS as mentioned in Lunn et al. (2012). Hence, the model likelihood can be formulated as the product of the densities of some invented random variables, $z_i$ as a set of observations. These variables are assumed to follow the Poisson distribution with the expected mean,$(\phi_i)$ equal to the minus the log-likelihood function and all of the values are initialized to be zero. As a Poisson mean, the $(\phi_i)$ should always be positive, therefore we should add an appropriate constant, $K$ where it is selected in such way that -log$(L_i + K)$ is always greater than zero. The likelihood function will not be affected by this action because it is equivalent to multiplying the unnormalized posterior distribution by a constant term equal to $e^{-nK}$ (Ntzoufras, 2011).

Finally, for each models we ran three chains with a thinning interval of 3 and a burn-in period of 10,000 samples. The first chain is initalized with zero values for all parameters,

while the initial values for the remaining two chains are generated by WinBUGS. After the burn-in period, 100,000 samples were collected. For each model, the coda files were saved and imported to R for MCMC diagnostics. The diagnostic plots including the autocorrelation, Gelman-Rubin, trace plots and density plots are created using the `CODA` package for assessing chain convergence and sample representativeness.

## 2.5 Model Evaluation with Deviance Information Criterion (DIC)

To evaluate the candidate models, Aktekin and Musal (2015) use the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). There are several approaches to evaluate Bayesian models but the freedom of the model parameters is restricted by the prior distributions when informative ones are used. In this case, Spiegelhalter et al. (2002) made a proposal for considering the effective number of parameters, $p_D$ which is defined as the difference between the posterior mean deviance and the deviance of posterior means of the parameters of interest. In general, adding parameters to a model results in a better model fit, but can also result in overfitting. In order to account for this effect when comparing models, the effective number of parameters $p_D$ is introduced, that serves as the complexity penalty term of the DIC that is defined in the following way:

$$\text{DIC} = \overline{D}(\theta) + p_D = D(\overline{\theta}) + 2\,p_D \tag{14}$$

This equation shows that the DIC consists of two parts: a measure of fit and a measure of complexity, which explains why a lower value of DIC reflects a better model quality. As the best fit model may achieve a low DIC value with either one of its two components, an adequate model is always a trade-off between fitness and complexity. According to Lunn et al. (2012), it is only important to compare the relative differences between models in DIC but not the absolute values. We do our model evaluation (see chapter 3.3) based on this notion. It is also worth mentioning that for a model implemented with the "zero trick" in WinBUGS (see chapter 2.4.3), a simple adjustment must be made to calculate the actual DIC value:

$$\text{DIC} = \text{DIC}_{\text{zero}} - 2\,n\,K \tag{15}$$

where $n$ is the number of observations, in our case the number of counties, $K$ is the constant term mentioned in chapter 2.4.3 and $\text{DIC}_{\text{zero}}$ is the DIC value computed in WinBUGS with zero trick implementation.

# 3 Results

## 3.1 Convergence and Sample Representativeness

In approximating distribution for sampling from posterior distributions with a Markov chain, we want to ensure that the values produced in the chain are representative of the posterior distribution and a sufficient sample size is retained after the burn-in period, so that the estimates are accurate. We perform these examinations through graphical methods and a numerical check where the results are presented in the following subsections respectively.

### 3.1.1 Graphical Examination

The simplest way to detect unrepresentativeness or lack of convergence is by eyeballing the density plot, chain trajectory, trace plot as shown in figure 3, which is the result taken from the diagnostics of the interaction term in M4EinterT4E model. This is an exemplary that the Markov chains have reached stationarity. We can see that the trace plot at the bottom has a "fat hairy caterpillar" appearance, randomly scattering about a stable mean value with all the chains overlapping each other. This is a good sign that suggests convergence. From the middle right plot, the 3 chains start at different initial values eventually converge together and start to behave similarly, implying that the sample is representative after the few thousands iterations. This preliminary duration that precedes the representative region of the posterior distribution is the burn-in period. This is the reason the burn-in period of 10,000 was chosen so that these unrepresentative samples are excluded. Another visual representation is the density plot that appears at the top left in figure 3 where the three chains are superimposing each other very well which suggests that the chains are generating representative values from the posterior distribution. For the accuracy of the estimates, it can be gauged with the help of auto-correlation function (ACF) plot as shown in the top-right plot of figure 3. This plot demonstrates that the auto-correlation is large at short lags but drops off quickly toward zero, indicating a convergence to independent samples. Most of the parameters in the models show similar results as in Figure 3. However as the model complexity (M1inter-M4inter, M4EinterpE, M4EinterT4E, M4EinterT4pE) increases, the coefficient of the interaction term, for $\lambda$ in particular, shows signs of non-convergence. The coefficient of the respective inequality covariate and the temporal parameter in those models also exhibit a mediocre convergence. The spatial and CAR prior, on the other hand, demonstrate good convergence in general. All other diagnostics plots can be found in the repository. (`https://github.com/Giant316/bayesianinference_HIVmortality/tree/master/HTML_Diagnostics`)

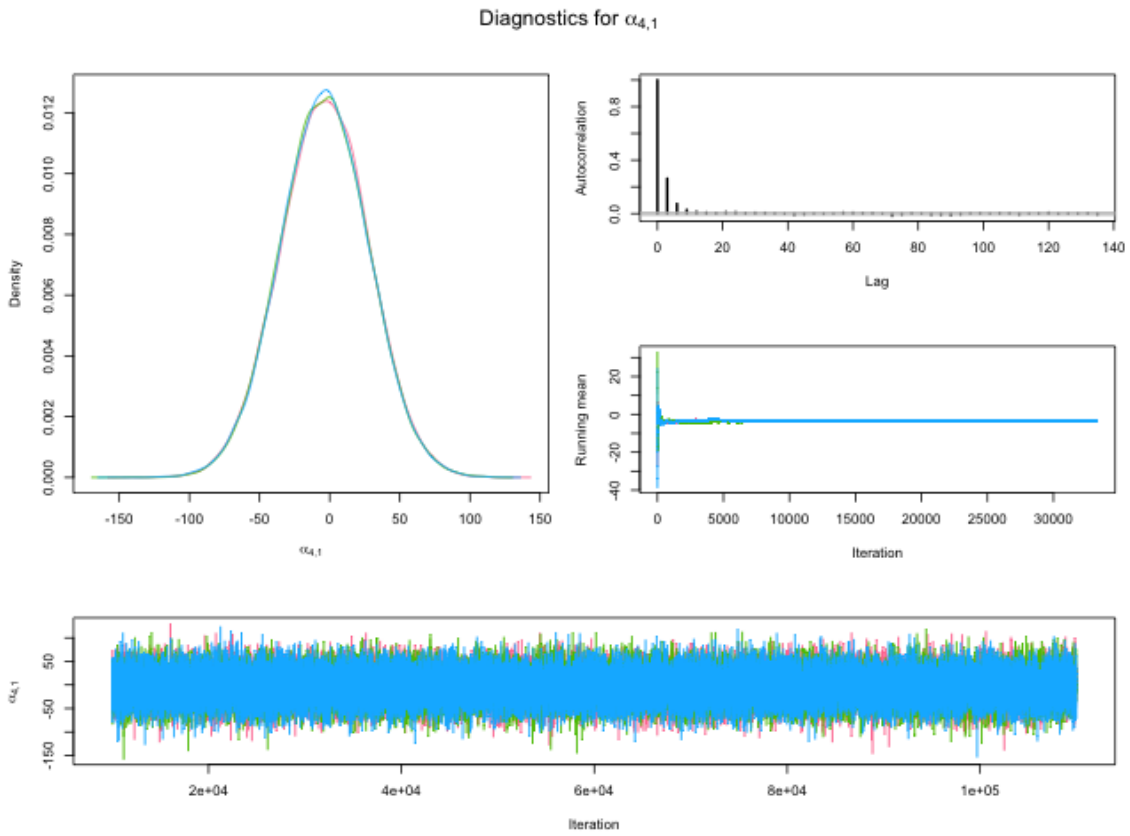



**Figure 3:** Exemplary Graphical diagnostics

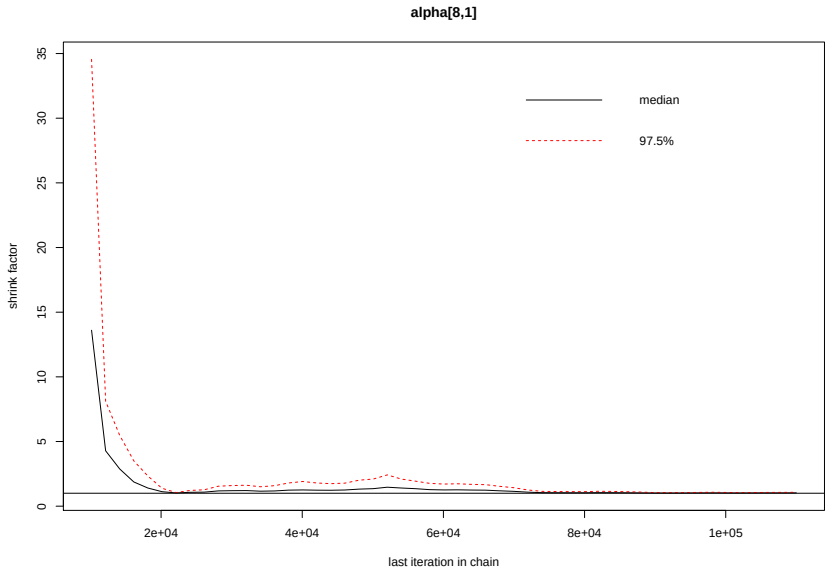



**Figure 4:** Exemplary Gelman & Rubin diagnostics

### 3.1.2 Numerical Check

Other than visual inspection, there are also numerical checks. Using the Gelman-Rubin diagnostic which is an ANOVA-type test, comparison of within-chain variance and between-chain variance can be made. The average difference between the chains and the average difference within the chain will be the same when all the chains have converged into a representative sampling. This is reflected by the potential scale reduction factor (PSRF) where its value equals to one if the chains are fully converged which means that the samples for each parameter are representative. The evolution of the PSRF as the number of iterations increases can be plotted in R as shown in Figure 4 where the PSRF has converged to approximately 1 over iterations. As a rule of thumb, when PSRF is greater than 1.1, the chains should be run longer in order to improve convergence. As tabulated in Table 2, M4inter model is the only model with PSRF value greater than 1.1. Noticeably, the models with less complexity (MnoCov, Mp, M1i-M4i, M1-M4, ME, M4E, M4iE) have PSRF value closer to 1. However, M1inter-M3inter models also have relatively low PSRF despite these models have higher complexity. Meanwhile M4inter model has the highest PSRF value although models like M4EinterpE, M4EinterT4E and M4EinterT4pE are more complex but they share a lower PSRF value.

**Table 2:** PSRF in Gelman & Rubin Test

| **Model Name** | **PSRF** | **Model Name** | **PSRF** | **Model Name** | **PSRF** |
|---|---|---|---|---|---|
| MnoCov | 1.00 | M1 | 1.01 | ME | 1.00 |
| Mp | 1.00 | M2 | 1.02 | MpE | 1.08 |
| | | M3 | 1.00 | M4E | 1.02 |
| | | M4 | 1.02 | M4iE | 1.02 |
| M1i | 1.00 | M1inter | 1.03 | M4EinterpE | 1.10 |
| M2i | 1.00 | M2inter | 1.02 | M4EinterT4E | 1.05 |
| M3i | 1.01 | M3inter | 1.01 | M4EinterT4pE | 1.05 |
| M4i | 1.00 | M4inter | 1.13 | | |

## 3.2 Posterior Regression Coefficients

Table 3 shows the posterior statistics for all candidate models which include the models' respective covariate effect of poverty, the four inequality measures and the interaction term. Firstly, considering the effects on the Ps, the posterior distributions of the coefficients for all models share similar intervals, ranging roughly from -60 to 60 with comparable standard deviations of around 30 except for the coefficient of $T_4$ which is approximately 20 with a slightly different interval, ranging roughly from -55 to 20. The mean value for coefficients of the same covariate for the Ps across different models does not differ very much. For instance, the mean value of the coefficients of $T_4$ is -16.40, -16.96, -16.75 and -17.89 in M4i, M4, M4E and M4iE models respectively. Other covariates also show a similar pattern with the exception of poverty covariate. The coefficient of poverty has marginal negative mean values in Mp, M1, M2, M3 and MpE models but it has positive values which are close to 5 in M4 and M4E models.

For the effects on $\lambda$s, there is no prominent range of distribution across different models unlike the ps. In fact, the ranges differ significantly. For example, the coefficient of the interaction term in M3inter has 95% of the distribution covers the most negative range, from -133.70 to -1.41, on contrary, the coefficient of the $T_1$ inequality measure in M1i has its 95% of distribution covers the most positive range, from 33.44 to 56.13. The standard deviations also vary greatly, having the highest value of 34.48 in M3inter model and the lowest value of 0.10 in M4i model. Unlike the Ps, the posterior mean changes substantially across different models. Only $T_4$ maintains similar range with value of 1.43, 1.49, 1.24 and 1.38 in M4i, M4, M4E and M4iE respectively.

Next, when comparison is done among models that have interaction term, the posterior means for the effect on Ps are close to zero in overall. As mentioned earlier, the standard deviations are all around 30. However for the effect on $\lambda$s, there is a huge variation in the means as well as the standard deviations. For the means, the largest value is 6.67 in M4EinterpE model and the smallest value is -73.76 in M1inter model. As for the standard deviations, M3inter model has the highest value of 34.48 and M4inter model has the lowest value of 6.90. Lastly, for the additional covariate, the unemployment rate that was introduced in our study, it has similar effect on Ps like other covariates as described earlier but it has negative posterior mean values across different models. As for the effect on $\lambda$s, the posterior distribution is shifting toward negative range as more other covariates were added in the model.

**Table 3:** Results from the posterior regression coefficients for $p$ and $\lambda$. The models are the same ones that were also considered by Aktekin and Musal (2015).

| **Model** | **Covariate** | **2.5%** | **mean** | **97.5%** | **sd** | **MC error** |
|---|---|---|---|---|---|---|
| Mp | poverty (p) | -58.63 | -0.37 | 58.59 | 29.92 | 0.29 |
| | poverty ($\lambda$) | 4.68 | 5.56 | 6.41 | 0.44 | 0.00 |
| M1i | $T^1$ (p) | -56.44 | 4.63 | 65.73 | 31.28 | 0.11 |
| | $T^1$ ($\lambda$) | 33.44 | 43.84 | 56.13 | 5.82 | 0.15 |
| M2i | $T^2$ (p) | -62.81 | -2.57 | 57.42 | 30.63 | 0.17 |
| | $T^2$ ($\lambda$) | 10.04 | 11.63 | 13.25 | 0.82 | 0.01 |
| M3i | $T^3$ (p) | -57.24 | 4.03 | 65.50 | 31.42 | 0.11 |
| | $T^3$ ($\lambda$) | 33.01 | 42.71 | 53.65 | 5.29 | 0.12 |
| M4i | $T^4$ (p) | -53.01 | -16.40 | 21.18 | 18.94 | 0.49 |
| | $T^4$ ($\lambda$) | 1.23 | 1.43 | 1.62 | 0.10 | 0.00 |
| M1 | poverty (p) | -58.83 | -0.19 | 58.18 | 29.67 | 0.31 |
| | $T^1$ (p) | -56.77 | 4.50 | 66.09 | 31.46 | 0.10 |
| | poverty ($\lambda$) | 4.57 | 5.45 | 6.30 | 0.44 | 0.01 |
| | $T^1$ ($\lambda$) | 1.35 | 5.56 | 10.19 | 2.24 | 0.06 |
| M2 | poverty (p) | -58.74 | -0.07 | 58.19 | 29.81 | 0.32 |
| | $T^2$ (p) | -61.00 | -0.99 | 58.95 | 30.66 | 0.17 |
| | poverty ($\lambda$) | 2.12 | 3.88 | 5.41 | 0.82 | 0.02 |
| | $T^2$ ($\lambda$) | 1.02 | 4.22 | 7.63 | 1.63 | 0.05 |
| M3 | poverty (p) | -58.50 | -0.32 | 58.33 | 29.73 | 0.32 |
| | $T^3$ (p) | -56.56 | 4.53 | 66.11 | 31.28 | 0.11 |
| | poverty ($\lambda$) | 4.55 | 5.45 | 6.30 | 0.44 | 0.01 |
| | $T^3$ ($\lambda$) | 1.52 | 5.56 | 10.01 | 2.14 | 0.06 |
| M4 | poverty (p) | -54.37 | 5.20 | 64.19 | 30.26 | 0.31 |
| | $T^4$ (p) | -54.31 | -16.96 | 20.45 | 19.28 | 0.49 |
| | poverty ($\lambda$) | -2.64 | -0.28 | 2.00 | 1.17 | 0.04 |
| | $T^4$ ($\lambda$) | 0.93 | 1.49 | 2.02 | 0.28 | 0.01 |
| M1inter | interaction (p) | -61.29 | 0.43 | 62.44 | 31.55 | 0.10 |
| | interaction ($\lambda$) | -127.80 | -73.76 | -14.82 | 28.28 | 1.15 |
| M2inter | interaction (p) | -61.11 | 0.63 | 62.97 | 31.59 | 0.11 |
| | interaction ($\lambda$) | -80.57 | -59.73 | -37.43 | 11.03 | 0.42 |
| M3inter | interaction (p) | -61.21 | 0.42 | 62.86 | 31.58 | 0.11 |
| | interaction ($\lambda$) | -133.70 | -68.18 | -1.41 | 34.48 | 1.42 |
| M4inter | interaction (p) | -60.03 | 0.98 | 61.97 | 31.13 | 0.19 |
| | interaction ($\lambda$) | -20.10 | -7.32 | 5.29 | 6.90 | 0.29 |

**Table 4:** Results from the posterior regression coefficients for $p$ and $\lambda$. All of the models use the additional covariate E, the unemployment rate. The models using an inequality measure were only calculated for the best fit model which uses $T^4$.

| **Model** | **Covariate** | **2.5%** | **mean** | **97.5%** | **sd** | **MC error** |
|---|---|---|---|---|---|---|
| ME | unemployment (p) | -67.41 | -6.65 | 54.13 | 30.98 | 0.21 |
| | unemployment ($\lambda$) | 9.25 | 10.88 | 12.48 | 0.83 | 0.01 |
| MpE | poverty (p) | -59.06 | -0.15 | 58.40 | 30.00 | 0.34 |
| | unemployment (p) | -67.27 | -6.70 | 54.05 | 30.93 | 0.22 |
| | poverty ($\lambda$) | -12.34 | -4.65 | 4.05 | 4.30 | 0.18 |
| | unemployment ($\lambda$) | 3.12 | 19.76 | 34.40 | 8.22 | 0.34 |
| M4E | poverty (p) | -54.18 | 4.91 | 64.39 | 30.10 | 0.32 |
| | $T^4$ (p) | -55.38 | -16.75 | 22.01 | 19.76 | 0.54 |
| | unemployment (p) | -64.02 | -3.65 | 57.24 | 31.05 | 0.20 |
| | poverty ($\lambda$) | -10.10 | -3.85 | 3.03 | 3.22 | 0.13 |
| | $T^4$ ($\lambda$) | 0.58 | 1.24 | 1.93 | 0.34 | 0.01 |
| | unemployment ($\lambda$) | -7.43 | 8.73 | 21.93 | 7.39 | 0.31 |
| M4iE | $T^4$ (p) | -55.78 | -17.89 | 19.70 | 19.24 | 0.49 |
| | unemployment (p) | -63.71 | -3.12 | 57.60 | 30.99 | 0.20 |
| | $T^4$ ($\lambda$) | 0.71 | 1.38 | 2.04 | 0.34 | 0.01 |
| | unemployment ($\lambda$) | -5.01 | 0.34 | 5.52 | 2.66 | 0.09 |
| M4EinterpE | interact. p*E (p) | -61.90 | 0.13 | 62.47 | 31.77 | 0.11 |
| | interact. p*E ($\lambda$) | -24.49 | 6.67 | 38.84 | 15.73 | 0.64 |
| M4EinterT4E | interact. $T^4$*E (p) | -65.48 | -3.63 | 57.76 | 31.46 | 0.14 |
| | interact. $T^4$*E ($\lambda$) | -33.42 | -15.37 | 3.30 | 9.58 | 0.39 |
| M4EinterT4pE | interact. $T^4$*p*E (p) | -62.03 | -0.05 | 61.88 | 31.64 | 0.10 |
| | interact. $T^4$*p*E ($\lambda$) | -35.19 | 1.15 | 37.98 | 18.67 | 0.72 |

## 3.3 Model Evaluation with DIC

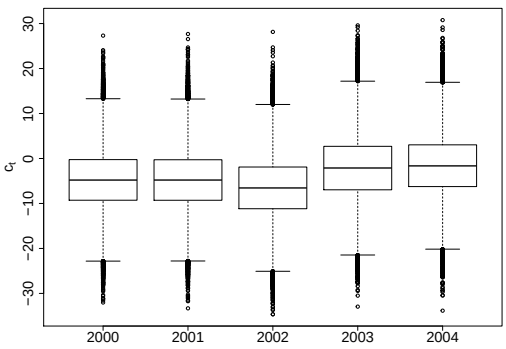
We access the fit of the model by comparing their DICs as mentioned in equation 15 where generally a lower DIC value indicates a better fit but the comparison in the value differences rather than the absolute value brings far more meaningful interpretation. Table 4 shows the fit measures of all models mentioned in chapter 2.4.2. On first glance, M2inter and M4EinterT4E give the lowest DIC value but on a closer comparison between all the models, M4i, M4, M4E, M4iE, M4EinterpE, and M4EinterT4pE give a consistent low DIC value with marginal differences. In the other spectrum, MnoCov which is the benchmark model, counter-intuitively, does not give the highest DIC value but the M1i and M3i models give the highest two DIC values. Although the DIC values reflected by our models differ from the original paper however our result shows a similar trend.

**Table 5:** DIC results

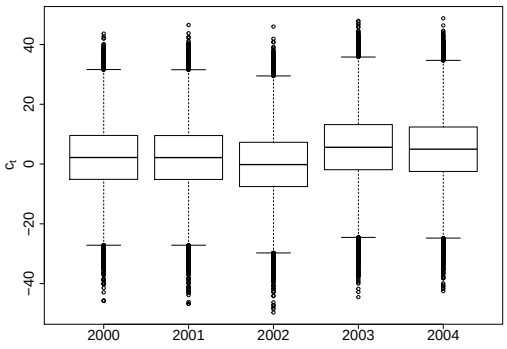
| **Model Name** | **DIC** | **Model Name** | **DIC** | **Model Name** | **DIC** |
|---|---|---|---|---|---|
| MnoCov | 1353 | M1 | 927 | ME | 910 |
| Mp | 937 | M2 | 917 | MpE | 893 |
| | | M3 | 928 | M4E | 854 |
| | | M4 | 852 | M4iE | 853 |
| M1i | 2497 | M1inter | 901 | M4EinterpE | 852 |
| M2i | 1077 | M2inter | 841 | M4EinterT4E | 846 |
| M3i | 2270 | M3inter | 910 | M4EinterT4pE | 852 |
| M4i | 852 | M4inter | 851 | | |

## 3.4 Spatial and Temporal Effects for Best Fit Model

Another feature of the ZIP model proposed by Aktekin and Musal (2015) is to take temporal variation into account via time-dependent random and spatial effects. We can observe the posterior statistics of the two best fit models M2inter and M4EinterT4E to interpret how temporal components influence HIV mortality. As shown in the boxplots of the temporal effects $c^{(t)}$ in figure 5, a similar trend was observed in both models where the effect was about the same for 2000, 2001, 2003 and 2004. Only in 2002, the posterior distribution was shifted negatively in value. The boxplots of spatial precisions $\tau^{(t)}$ for the best two models are shown in figure 6, indicate that spatial effects varied slightly over the years.

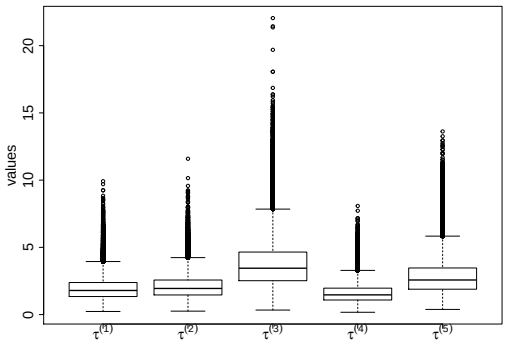



**(a) M2inter**

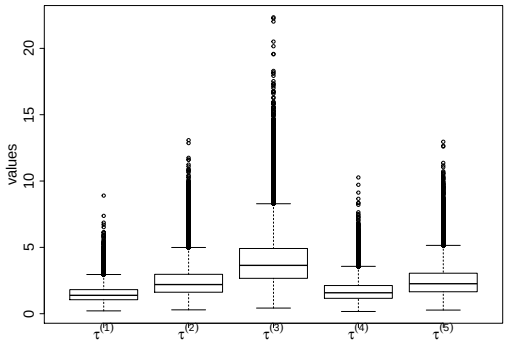



**(b) M4EinterT4E**

**Figure 5:** Temporal effects for the two best fit models: (a) M2inter and (b) M4EinterT4E




**(a) M2inter**




**(b) M4EinterT4E**

**Figure 6:** Spatial precision, $\tau^{(t)}$ for the two best fit models: (a) M2inter and (b) M4_Einter_T4E

# 4 Discussion

In the following section, further implications for the ZIP model with covariate effects such as inequality measures, poverty, unemployment rate and interaction are discussed. Owing to the logit structure mentioned in equation 12, a negative value for the coefficients indicates a positive effect on HIV mortality on the Ps portion in interpreting the regression coefficients of Ps. In other words, the HIV risk is small (with probability of non-zero death count less than 50%). This argument can only be valid if and only if all the spatiotemporal effects and the covarites of interest lie in the positive range. As observed in the boxplots of inequality measure in Figure 2, the distributions lie mainly in the positive region, this also holds for all other parameters of interest. For the case of $\lambda$s, a positive sign for a covariate coefficients implies a positive effect on non-zero death count. However, to access the overall effect on the HIV mortality we need to take into account of the relationship of Ps and $\lambda$s as defined in equation 7.

To discover inequality measures that help to enhance understanding of HIV mortality, we need to compare the posterior results of M1 model and M3 model to check the effects of CPI inclusion in $T_3$ while having poverty as a control factor. Subsequently, we need to compare the posterior results of M2 model and M4 model in relation to those in M1 model and M3 model. The posterior distributions in M1 model and M3 model are similar for effects on both Ps and $\lambda$s. Their DIC results also reflect that there is no advantage in adding CPI information to the inequality measure's definition (see definitions in 2.2.3). This finding contradicts to the paper in Aktekin and Musal, 2015, this is most likely caused by data disparity. Comparing M2 model in relation to M1 model and M3 model, its posterior mean and DIC result indicates $T_2$ has better performance in explaining HIV mortality. When comparing the posterior results of M4 model to M1-M3 models, M4 model outperforms all the other three models. To recap, $T_1$ and $T_3$ consider how the per-capita income of a ZCTA in relation to the population of the entire NY state while $T_2$ only considers how the per-capita income of a ZCTA in relation to the population in the county the ZCTA located. As for $T_4$, it is simply the percentage of ZCTAs that have lower per-capita income share in relation to their population share which is calculated from $T_2$. Hence, we argue that using county-specific information in calculating inequality measures results in better fit.

As mentioned in chapter 3.2, the posterior mean values for coefficient of $T_4$ show consistency across different models and their distributions lie in the most negative range with the highest confidence interval, suggests that $T_4$ is the most reliable covariate in explaining HIV mortality. Due to the limitation of computing power, we could not run our extension models with all inequality measures other than just $T_4$ so we could not do a model-for-model comparison for the DIC results. Nonetheless, as we compare the DIC results, the models incorporated with $T_4$ as covariate (refer table 5) generally produce low DIC values.

Hence we can infer that $T_4$ inequality measure provides the most information in explaining HIV mortality as compared to the other three inequality measures (see definitions in 2.2.3). It is also important to note that even without controlling poverty in a particular region as modelled in M1i - M4i models, the inequality measures still have influence on HIV death count either via Ps or $\lambda$s as shown in Table 3.

Another point that we can emphasize is that using poverty in conjunction with any inequality measures in a model always improves model fit. We can infer the same argument for unemployment rate. This is supported by the posterior means of their coefficients with negative value for Ps and positive value for $\lambda$s, which indicate that the mortality risk is expected to be higher for regions with higher poverty or with higher unemployment rate. This reasoning can further be justified by comparing the DIC results of the models M1i-M4i (with only inequality) with the models M1-M4 (with poverty and inequality) and M1inter-M4inter (with poverty, inequality and interaction). Although we didn't run the models incorporated $T_1$ - $T_3$ with unemployment rate, our argument can still stand by the comparison of DICs between ME model and MpE model. However, adding the unemployment rate to a model with more than one covariate (e.g. M4E) does not further improve model fit.

Lastly, we want to interpret the effect of including different interaction terms in different models. Based on the posterior summary statistics, we can conclude that all the interaction terms of the four default models and the other three from our extended models have a non-existent effect on $P$s and aside from M4EinterpE model and M4 EinterT4pE model, they have an insignificant effect on $\lambda$s. In term of model fit, it is not surprising that the M4EinterT4E model (with interaction term between $T_4$ and E) has the second lowest DIC value since the respective covariates perform well in explaining HIV mortality. It is unexpected to have the lowest DIC value for M2inter model (with interaction term between $T_2$ and poverty). The reason of this occurrence is yet to be investigated but judging from the posterior summary statistics for all interaction models, the interaction terms convey no information in explaining HIV mortality but only improve the prediction occasionally.

In summary, our results show similar pattern and trend as the results presented in Aktekin and Musal, 2015. By only considering spatiotemporal effects is insufficient to explain HIV mortality risk and by using the county-specific information in formulating inequality measure gives better result. Increasing model complexity up to certain point show no further improvement in model fit which align with the trade-off between adequacy and complexity. The model that incorporates $T_4$ and unemployment rate is the best model in explaining HIV risk.

# 5 Conclusion and Outlook

In addition to the candidate models in Aktekin and Musal, 2015, we extended their work by introducing another covariate, the unemployment rate. By choosing the best inequality measure, $T_4$, the poverty and the unemployment rate combinatorially, we can get 4 new possible candidate models. Via different combination of $T_4$, poverty rate and unemployment rate, three new interaction terms can be formed and added alongside to the model with all these three covariates to obtain another 3 candidate models. Therefore, in total of 7 new possible candidate models are modelled in this paper.

We were facing several challenges in collecting all the data to reproduce the original paper so it took more time than estimated. Due to the unavailability of some referenced data sources, we spent most of the time finding comparable data and integrating data with different data structure which collected from various sources. For instance, the HIV mortality data from the Center of Disease Prevention have some counties data with less than 10 occurences "suppressed" from viewing because of privacy protection. Besides that, only the latest CPI data is available online which is dated from 2007 onwards. Despite that we were very fortunate to be able to get in touch with Professor Musal and gain his earnest help. Eventually a workaround was found and so the data collection issues were resolved. On top of that, we also faced a roadblock in attempting to run WinBUGS model from R by invoking the BUGS software via R2WinBUGS package. In the end, the modelling was executed in WinBUGS and the resulted coda files were saved which is then imported to R for customized plotting and further analysis.

There are several areas for further improvement. First of all, implementation of the model in JAGS is a worthwhile exploration. It could potentially save some time in running the model so more models can be investigated and most importantly, JAGS is still under an on-going development with larger user base hence resources for troubleshooting are more accessible. Besides, since we have reduced the burn-in period to 10,000 samples, the regression coefficients of the covariates on $\lambda$s show sign of unrepresentative sampling in complex models. Therefore, models with more burn-in period as proposed in the original paper could be examined using JAGS as it is more likely to be less time-consuming. In future research, more recent data could be used for modelling as its comparison with the old findings might be insightful to HIV risk study. Moreover, a different way in redefining the inequality measures could possibly produce better result. One possible approach is to use Gini Index instead of Theil Index.

# References


Agarwal, D. K., A. E. Gelfand, and S. Citron-Pousty (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* 9.4, pp. 341–355.

Aktekin, T. and M. Musal (2015). Analysis of income inequality measures on human immunodeficiency virus mortality: A spatiotemporal Bayesian perspective. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 178.2, pp. 383–403.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregressions. *Biometrika* 82.4, pp. 733–746.

Centers for Disease Control and Prevention (October 2002). *Instruction Manual (Part 9) for ICD-10 Cause-of-Death Lists for Tabulating Mortality Statistics*. URL: `https://www.cdc.gov/nchs/data/dvs/im9_2002.pdf.pdf`.

Conceição, P. and P. Ferreira (2000). "The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications".

Harrison, K. M. D., Q. Ling, R. Song, and H. I. Hall (2008). County-Level Socioeconomic Status and Survival After HIV Diagnosis, United States. *Annals of Epidemiology* 18.12, pp. 919–927.

LaMontagne, B. and D. Stockemer (2010). Determinants of HIV prevalence: A global perspective. *International Politics* 47.6, pp. 698–724.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2012). *The BUGS Book - A Practical Introduction to Bayesian Analysis*. Boca Raton: CRC Press.

Musal, M. and T. Aktekin (2012). Bayesian spatial modeling of HIV mortality via zero-inflated Poisson models. *Statistics in Medicine* 32, pp. 267–281.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. Vol. 698. John Wiley & Sons.

NYC Health (2019). *HIV/AIDS Annual Surveillance Statistics*. URL: `https://www1.nyc.gov/site/doh/data/data-sets/hiv-aids-annual-surveillance-statistics.page` (visited on 07/13/2019).

Olsen, E., P. E. Carrillo, and D. W. Early (2012). A Panel of Price Indices for Housing Services, Other Goods, and All Goods for All Areas in the United States 1982-2010. *SSRN Electronic Journal*.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B* 64, pp. 583–639.

US Census Bureau (2019). *Zip Code Tabulation Areas (ZCTAs)*. URL: `https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html` (visited on 07/13/2019).

Wabersich, D. and J. Vandekerckhove (2013). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior research methods* 46.

Waller, L., B. P. Carlin, H. Xia, and A. E. Gelfand (1997). Hierarchical Spatio-Temporal Mapping of Disease Rates. *Journal of the American Statistical Association* 92.438, pp. 607–617.

Wilkinson, R., M. Marmot, and World Health Organization (2003). *The Solid Facts: Social Determinants of Health*. Second. Copenhagen: World Health Organisation.

Zuur, A. F., A. A. Saveliev, and E. N. Ieno (2012). *Zero Inflated Models and Generalized Linear Mixed Models with R*. Newburgh: Highland Statistics.

# A Differences to the Original Paper

| Multiple Cause of Death Data | |
|---|---|
| **Aktekin and Musal (2015)** | **Our Approach** |
| 2101, 1943, 1829, 1476 and 1302 HIV-related deaths in NY state for the years 2000-2004 | 2,190, 2,028, 1,920, 1,844 and 1,675 death counts for NY state between 2000 and 2004. |

| Census 2000 Data | |
|---|---|
| **Aktekin and Musal (2015)** | **Our Approach** |
| urban = "areas that have 1000 and 500 people per square mile" | = ZCTAs with >10 individuals living in an urban area or cluster |
| 957 ZCTAs classified as urban | 937 ZCTAs classified as urban |
| poverty rate of Hamilton county = 10 % | = 0.1050072 which corresponds to the mean poverty |

| Inequality Measures | |
|---|---|
| **Aktekin and Musal (2015)** | **Our Approach** |
| $T_1$ and $T_3$ are equations on county level. | $T_1$ and $T_3$ are equations on state level. |
| $T_2$ is calculated base on the entire state of NY. | $T_2$ is equation of the county. |

| Settings for MCMC sampling | |
|---|---|
| **Aktekin and Musal (2015)** | **Our Approach** |
| burn-in period: 80.000 | burn-in period: 10.000 |

# Author Contributions

- **data collection:** Jia Yan Ng (JYN), Lena Schwertmann (LS), Thummaporn Nimpanomprasert (TN)
- **data cleaning:** JYN, LS, TN
- **calculation of covariates:** JYN, LS, TN
- **running and interpreting the models with WinBUGS:** JYN, LS, TN
- **drafting and reviewing of report:** JYN, LS, TN