Bias Modeling for Integrating Digital Data and Conventional Surveys for Migration Estimation

Yuan Hsiao

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2021

Reading Committee:

Jonathan Wakefield, Chair

Emilio Zagheni

Program Authorized to Offer Degree:

Statistics

University of Washington

Abstract

Bias Modeling for Integrating Digital Data and Conventional Surveys for Migration Estimation

Yuan Hsiao

Chair of the Supervisory Committee:

Jonathan Wakefield

Department of Statistics

Obtaining reliable and timely estimates of migration flows is critical for advancing migration theory and guiding policy decisions, but it remains a challenge. Digital data provide granular information on time and space but do not draw from representative samples of the population, thus leading to biased estimates. The thesis proposes a method for combining digital and survey data by modeling the spatial and temporal dependence structure of the biases of digital data. We use simulations to demonstrate the validity of the model, then empirically illustrate our approach by combining geo-located Twitter data with data from the American Community Survey (ACS) to estimate state-level emigration in the United States. We show that our model that combines unbiased and biased data produces predictions that are more accurate than predictions based solely on unbiased data. Our approach demonstrates how digital data can be a complement, rather than a replacement of, representative surveys.

Acknowledgements

The training from writing this thesis and the Statistics department in general has tremendously transformed how I conduct research. Jonathan Wakefield has been an incredible advisor that demonstrated the spirit of how one should never be satisfied with the current results. He encouraged me to get to the fundamental aspect of the research question, find out the limitations of the current results, then conduct more analyses to complete the picture. It is this type of scientific rigor that makes statisticians respected, and the process of this thesis gave me an excellent opportunity to learn from Jon. Jon pushed me towards cutting-edge research, and working with him has been both exciting and challenging.

Emilio Zagheni showed me how digital data can be used to identify social processes. He first advised me on how to collect and transform digital data into meaningful statistics, leading me into the world of data repurposing that is so critical in the current society. Then, as digital data are not representative samples, Emilio pushed me to think about what estimates from digital data mean, and how to think about them in the context of our current society where digital data will only be ever-increasing.

More generally, the sequence of courses at the Statistics departments gave me a deeper perspective on my research. As a Ph.D. student in Sociology, I had conducted numerous statistical analyses, but the Statistics department training pushed me to think about the strengths and limitations of my past and ongoing research. I also made friends from taking the courses, which was a fun experience as it exposed me to lifeworlds very different from the Sociology circle. I am very grateful for being part of the Statistics department community.

I would also like to thank my wife, Arwen, for encouraging me and supporting me through the processes of applying for and fulfilling the program.

Contents

1 Motivation

Reliable and timely estimates of migration flows are necessary to guide policy decisions and to improve our understanding of migration processes. Migration flows, such as the mobility of people within a country, have implications on a wide range of phenomena such as urban growth and development (Greenwood, 1981), housing dynamics (Clark et al., 2000), and the market for labor (Moretti, 2013; Molloy et al., 2011, 2017). Consequently, the size of migration flow affects the population size and the population constitution of an area, leading policy makers to integrate the estimation of migration into their considerations of policy making. For instance, the size of the migrant population affects how policy makers design distribution of social welfare (Tolstokorova et al., 2009), forecast employment rates (Sweeney and Goldstein, 2005), or design educational curricula (Chantavanich, 2003).

Traditionally, migration estimation has fallen in the realm of drawing from survey data, such as the American Community Survey in the United States or the Integration Barometer Survey in Germany. Nevertheless, as implementing large-scale surveys are often costly, depending on the country, official estimates of migration flows are often provided annually or biannually in developed countries, or in even longer periods in developing countries. For instance, in the US, the American Community Survey is implemented annually and retrospectively asks people where they lived one year ago. Migration estimates are then calculated from the aggregation of the number of people who changed residency (Census Bureau, 2018). In the case of Bangladesh, migration estimates are only available every five years based on survey estimates. Most importantly, regardless of developed or developing countries, it takes time for the official estimates from traditional surveys to be calculated, producing a time-lag in available estimates.

In contrast to the lag in survey estimates, policy-decisions require much shorter time-intervals to react to the changing social conditions as a consequence of migration. For example, as of now (year 2021), the state-to-state migration flow estimates from the American Community Survey are only available up to year 2018. However, for policy makers, it is critical to make decisions based on data that is as recent as possible, and to make projections with better temporal granularity. Put differently, the current state of data poises a significant gap between the available data and the required data to make better policy decisions. A timely estimation of migration flows would benefit the process of not only understanding migration trends but also inform better policy decision.

Digital data may provide assistance to obtaining migration data that is more timely and with more fine-grained temporal detail. With the advent of the technologies of cellphones and computers, people often leave digital traces of mobility, which have been utilized by scholars to estimate migration flows (Blumenstock, 2012; Zagheni et al., 2014; Jurdak et al., 2015; Fiorio et al., 2017; Hughes et al., 2016; Zagheni and Weber, 2012). The logic is that users of these technologies share their location during their usage, and via comparisons of the locations of users over time one can estimate the mobility of individuals, which can then be aggregated into estimates of migration flows. For instance, when some Twitter users post their tweets, they "geo-tag" the tweet, which shows the location in which the tweet was posted. The advantage of digital data is the high granularity in time and space, where information on the user is often in seconds (for time) and coordinates (for space). With the high granularity, scholars are able to obtain much richer information on users that was not possible from survey data. For instance, Blumenstock (2012) draws on cellphone records to study internal migration patterns in Rwanda. As each cellphone call is associated with a cellphone tower, from the tower locations one can infer the

area where the caller was located in. Via the temporal and spatial granularity of cellphone records, Blumenstock (2012) was able to identify patterns of temporary and circular migration that were not easily captured by governmental surveys. Zagheni and Weber (2012) alternatively draw from email data to estimate international migration rates. As each email is associated with an IP address, through the IP addresses one can identify the location and time of the email sender and in turn infer migration flows between countries. Fiorio et al. (2017) use Twitter data to identify migration patterns by using the "geotags" that reveal the location and time of the user, which can be aggregated to investigate migration patterns with different time spans (i.e., short-term and long-term migration).

However, digital data faces the serious challenge that users are not random samples of the population, and there is therefore inherent bias in the estimates. For instance, Twitter users tend to be younger than the general population (Mislove et al., 2011), and younger people tend to be more mobile than the older generation. If one took the estimate from Twitter at face value, it is likely one would overestimate the migration rates. In short, estimates from digital data are often biased and not reliable, and should not be utilized solely to inform policy-making decisions.

The critical challenge would be how to estimate and adjust for the bias from digital data. One possibility would be to develop a method that can combine estimates from digital data and survey data. Although not as timely and geographically granular as digital data, survey data can provide relatively unbiased estimates of migration flows. Consequently, one reasonable approach may be to combine social media data with official statistics to provide timely estimates. By combining different sources of data, we may be able to draw from the advantage of representativeness of official statistics, and the timely

availability of social media data.

Statistically, one can re-posit the question as how to model the relationship between two different sources of data: one which is unbiased and one that is biased. If the relationship between estimates from digital data and estimates from survey data follow statistical patterns that we can model, information from both data sources can be combined. For instance, if estimates from digital data are always higher than the population rates by a constant "bias" factor, then we may estimate this factor, re-scale and obtain reasonable estimates from social media data. In other words, the problem is not *whether digital data are biased*, but rather *how to model the structure of the bias*.

A shortcoming of the growing methodological literature on migration and social media data, however, is a lack of sophistication with respect to the structure of the bias of digital data. Much of the work that has been done so far assumes that the relationship between social media estimates and survey estimates is constant. We take a more flexible approach to determine whether a model that takes into account the spatial and temporal structure of bias in social media estimates improves overall model prediction. As such, we ask three interrelated questions:

- RQ1: How biased are estimates from digital data?
- RQ2: Are there statistical models that can capture the relationship between estimates from digital data and the true population rates?
- RQ3: Can we combine estimates from digital data and official statistics to improve short-term predictions?

we approach the problem by proposing a model that decomposes the spatial and temporal contributions to the bias of digital data. In doing so, we show that biases can be modelled and combined with survey data to improve prediction.

The model is generic and can incorporate multiple sources of data with simple requirements: (1) data from "unbiased" sources, such as representative surveys, that are perhaps not very timely or missing in certain time points; (2) timely data from potentially biased sources, such as from Twitter, geo-located logins into websites, or cellphone records.

We illustrate the method on state-level emigration rates in the US from years 2010-2016. we combine data from geo-located Twitter data for over 2 million users (2010-2016) with data from the American Community Survey (ACS). Given the research questions, we adopt the following analytical strategy. First we use Twitter data in the US to obtain estimates of state-level emigration rates. Second, we assess the bias by comparing, in a well-defined sense, estimates from Twitter and those from the ACS. Third, we select the model the best captures the structure of the bias based on the data. Finally, we use this model to combine Twitter and ACS data and improve prediction.

The structure of the thesis is as follows. In section 2, we present the statistical model as the general framework for the paper. In section 3, we simulate hypothetical data with one unbiased set of data and one biased set that directly measure the true process. We then show that the model that draws from both process has better predictive accuracy than a model that only utilizes the unbiased process. Section 4 applies the statistical model to real data from Twitter and the ACS in the case of state-level emigration rates in the United States. Section 5 offers discussions and conclusions.

2 The joint model that combines survey and digital data

2.1 Intuition of the model

The general setting for applications of the model is that there are two types of data: survey data that is unbiased but nonetheless may have missing data in certain, often recent periods, and digital data which is biased but more complete in the sense that there are less missing periods. For example, in the empirical case of this thesis, we can draw from Twitter data, which has more spatial and temporal information but is potentially biased, and ACS data, which is representative but may have missing years. To combine data from these data sources, we propose a joint-modeling approach to understand the data-generating mechanisms for both ACS and Twitter data.

The intuition of the model is that there is a common process for the "true emigration rates" in the population, and we have two sources of data that measure this process. The ACS data estimates this process with measurement error, while the Twitter data estimates this process with measurement error and bias. In other words, the ACS and the Twitter estimates for a particular state-year would follow a bivariate distribution with a partially shared mean component (the part without the bias). Over the states and years, the ACS and Twitter estimates follow a multivariate distribution with a partially shared mean component.

The key modeling choices are:

- How do we model the true process?
- How do we model the bias?

We draw from a well-established literature on space-time models modeling

(Knorr-Held, 2000; Mercer et al., 2015; Wakefield et al., 2018). The advantage of space-time models is that they incorporate information across space, as well as information across time. For instance, if the demographic composition of Connecticut state is more or less stable over time, we would expect migration rates of Connecticut to be smooth over time. However, we may also wish to incorporate information that Connecticut is adjacent to Massachusetts, and if their demographics are similar there would be spatial dependence between the two states. Similarly, we might expect that observations in year 2015 to be potentially similar in general, but also adjacent years to be associated. The space-time models decompose population processes into spatial and temporal processes with within and between effects, which can aid in estimation and prediction.

With a space-time framework, adjacent model components can address different dependencies in the data. For example, one could specify migration processes as a spatial process, a temporal process, a combination of independent spatial and temporal processes, or a process with space-time interactions. In the thesis, we would test these possibilities and select the appropriate model.

Similarly, one could conceptualize the bias as a spatial process, a temporal process, a combination of independent spatial and temporal processes, or a process with space-time interactions. Again, we would test these possibilities and select the appropriate model.

2.2 Defining the model components

For an observation in area s, time t, following Rue and Held (2005), define a model that specifies the population mean $\mu_{s,t}$ as a process (e.g., a migration process) as:

$$\mu_{st} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}.$$

As one can see, θ_s and ϕ_s are components related to space, such as states in the US, α_t and γ_t are components related to time, such as years, and δ_{st} is a term that specifies space-time interactions.

We explain each of the parameters θ_s, ϕ_s, α_t, γ_t, δ_{st} in the following (see Mercer et al. (2015)):

- ϕ_s and γ_t are independent random effects (i.e., effects with no spatial or temporal structure). These random effects have a generic $N(0, \sigma^2)$ form. For instance, $\phi_s \sim N(0, \sigma^2_\phi)$. This variance determines the amount of smoothing with small/large values favoring large/small amounts of smoothing. These capture state or year specific random effects. For instance, we might expect migration rates for Nevada state to be generally higher. Similarly, one could potentially expect migration rates for year 2015 to be generally lower.

- α_t is a random walk of order 2 process (RW2) on a yearly scale. Random walks are part of a larger family of Intrinsic Gaussian Markov Random Fields (IGMRF). IGMRFs are "improper", as they have precision matrices not of full rank.

  Intuitively, temporal random walk processes are processes that specify how the latent variables are dependent on the last k variables in time. k specifies the order of the random walk process. For instance, a random walk process with order 1 (RW1) indicates that an variable is dependent only on the last variable. Mathematically, let α_t be the latent variables, then

  $$\Delta_t = \alpha_t - \alpha_{t-1} \sim N(0, \sigma^2_\alpha)$$

  In other words, the value of the next latent variable in time is the value of the current latent variable in time plus a random independent increment.

The inverse of σ_α^2 is a precision parameter that determines the variation of the increments, or, equivalently, the amount of smoothing.

The logic can be applied to a random walk process of order 2. For a RW2 process, the conditional mean of a data point is dependent only on the last two data points with a precision parameter τ_y to estimate. We have:

- $\Delta^2\alpha_t = \Delta(\Delta\alpha_t)$, and $\Delta^2\alpha_t \sim_{iid} N(0, \tau_\alpha^{-1})$. Then:
- $E[\alpha_{t+k}|\alpha_1, \ldots, \alpha_t, \tau_\alpha] = (1+k)\alpha_t - k\alpha_{t-1}$
- $Prec(\alpha_{t+k}|\alpha_1, \ldots, \alpha_t, \tau_\alpha) = \tau_\alpha/(1 + 2^2 + \cdots + k^2)$

- One can specify similar processes for dependencies in space. In the model, θ_s is a local spatial smoothing term that is modeled as having an from intrinsic conditional autoregressive (ICAR) process. An ICAR model is an IGMRF but the dependency of the latent variable is now on the adjacent neighbors of a state (which is similar to a generalization of a RW1 process to a lattice). Let α_s be an latent variable in space with m neighbors. Under an ICAR model, the distribution of a variable α_s is:

$$\alpha_s|\boldsymbol{\alpha_{-s}}, \tau_\alpha \sim N(\tfrac{1}{m_\alpha}\textstyle\sum_{j:j\sim s}\alpha_s, \tfrac{1}{m_s\tau_\alpha}).$$

  Again τ_α is the precision parameter to be estimated in the model. In words, the model indicates that the mean of the latent variable is the average of the values of its neighbors and a random independent deviation. τ_α determines the magnitude of the variation of the deviation.

- δ_{st} is the type IV interaction described by Knorr-Held (2000). The interaction term that assumes the spatially ICAR effect and temporal RW2 effect interact at the yearly level. That is, the precision matrix Q_δ is the kronecker product of the precision matrix of the ICAR process and the precision matrix of the RW2 process: $Q_\delta = Q_\alpha \otimes Q_\theta$.

2.3 Mathematical formulation of the model

Having introduced the general space-time structure above, we proceed with applying the model to the case of emigration. Since emigration rates are probabilities that lie between zero and one, we model the *logit* of the emigration rates. The reason why we model the rates rather than the counts of migrants is practical: for Twitter data, we do not have the population size for each state. For instance, we do not have how many Twitter users there are in Washington state. Thus, a count model with an offset of population size is not feasible, and thus we follow Mercer et al. (2015) and models the logit of the rates.

Define $p_{s,t}$ as the emigration probability for an individual living in state s in year t. Also define $Y_{s,t}$ is the logit of $\hat{p}_{s,t}$ (the estimate of $p_{s,t}$), defined as $Y_{s,t} = \log(\frac{\hat{p}_{s,t}}{1-\hat{p}_{s,t}})$. we assess where these estimates arise from shortly.

Formally, denote:

- $Y_{s,t}^{\text{ACS}}$ as the logit of the migration estimates from ACS for state s, year t
- $Y_{s,t}^{\text{TWIT}}$ as the logit of the migration estimates from Twitter for state s, year t

The logits are continuous variables on the real line, and the data generating mechanism for both the ACS data and Twitter data are assumed to follow normal distributions:

- $Y_{s,t}^{\text{ACS}} \sim N(\mu_{s,t}, V_{s,t}^{\text{ACS}})$
- $Y_{s,t}^{\text{TWIT}} \sim N(\mu_{s,t} + B_{s,t}, V_{s,t}^{\text{TWIT}})$

where $V_{s,t}^{\text{ACS}}$ and $V_{s,t}^{\text{TWIT}}$ are the estimated variances that acknowledge the study design.

Notice the common mean component μ_{st} in both ACS and Twitter models. Because of this common mean component, we model the two processes *jointly*. That is:

$$\begin{bmatrix} Y_{s,t}^{\text{ACS}} \\ Y_{s,t}^{\text{TWIT}} \end{bmatrix} \sim N\left[\begin{pmatrix} \mu_{s,t} \\ \mu_{s,t} + B_{st} \end{pmatrix}, \begin{pmatrix} V_{s,t}^{\text{ACS}} & 0 \\ 0 & V_{s,t}^{\text{TWIT}} \end{pmatrix}\right]$$

In this model for the observed data $B_{s,t}$ represents the bias term for the Twitter estimates, and we assume that the ACS estimates are unbiased. Also, we assume that the covariance terms are 0 as the measurement errors of ACS and Twitter are independent because the data are drawn from independent samples, and measurement errors should be unrelated.

The choice would then becomes how to model μ_{st} and $B_{s,t}$. For the true process $\mu_{s,t}$, we use an extension of the Fay-Herriot model (Fay and Herriot, 1979) (also see (Mercer et al., 2015)), and specify $\mu_{s,t}$ as:

- A BYM2 spatial process: $\mu_{s,t} = \mu + \theta_s + \phi_s$
- A Random Walk 2 and IID temporal process: $\mu_{s,t} = \mu + \alpha_t + \gamma_t$
- A space-time main effect only process: $\mu_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t$
- A space-time interaction process: $\mu_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$

Where μ is an overall mean, θ_s is a spatial intrinsic conditional autoregressive process (ICAR), ϕ_s is a random IID intercept for each state, α_t is a random walk of order 2 process (RW2), γ_t is a random IID intercept for each year, δ_{st} is a structured interaction between the ICAR process and the RW2 process.

This statistical model shares information between contiguous neighbors and close time periods. The way such sharing is carried out is by specifying probability distributions that penalize contributions to the mean of $Y_{s,t}$ that are very different in areas that are geographically and/or temporally close. Hence, similarity in estimates is encouraged.

Similarly to how one models the common mean process $\mu_{s,t}$, one could model $B_{s,t}$ as:

- An ICAR spatial process: $B_{s,t} = \mu + \theta_s^b + \phi_s^b$
- A Random Walk 2 temporal process: $B_{s,t} = \mu + \alpha_t^b + \gamma_t^b$
- A space-time main effects process: $B_{s,t} = \mu + \theta_s^b + \phi_s^b + \alpha_t^b + \gamma_t^b$
- A space-time interaction process: $B_{s,t} = \mu + \theta_s^b + \phi_s^b + \alpha_t^b + \gamma_t^b + \delta_{st}^b$

Notice the superscripts b denote that these parameters correspond to the bias parameters. These different models specify different effects for the bias structure. On the one hand, one would wish to correctly specify the effects. For instance, if there were spatial dependency in the bias structure, the model should include such spatial effects. However, there is also the potential risk of over-specification, such as including a space-time interaction effect whereas there is none in the true bias structure. Extra parameters give wider intervals, which we would like to avoid as they may lead to worse predictive accuracy. It is difficult to derive substantively which choice would be better beforehand, and thus we use two approaches to adjudicate model selection. One is using model selection criteria, as discussed in subsection 2.4. A different approach but a separate exercise is to simulate hypothetical data and glean insights on how model mis-specification (either under-specification or over-specification) affects predictive accuracy, which will be shown in section 3.

2.4 Model selection and model evaluation

We fit the models using the *R-INLA* package in the statistical software *R* (Lindgren and Rue, 2015). Since the ACS data is representative, the analytical strategy is to use only the ACS data first to select the best model for $\mu_{s,t}$. We use three model selection criteria: log-CPO (higher indicates better fit), DIC (lower indicates better fit) and WAIC (lower indicates better fit) (also see (Gelman et al., 2014a; Krainski et al., 2018)). We explain each in detail below:

- Log-CPO represents the "leave-one-out" predictive measures of fit. The CPO value $P(y_i|\boldsymbol{y_{-i}})$, which is the probability density of an observed response y_i based on the model fit to the rest of the data ($\boldsymbol{y_{-i}} = y_1, y_2 \ldots y_{i-1}, y_{i+1}, \ldots y_n$). In other words:

$$p(y_i|\boldsymbol{y_{-i}}) = \int p(y_i|\theta)\pi(\theta|\boldsymbol{y_{-i}})d(\theta)$$

where $\pi(\theta|\boldsymbol{y_{-i}}$ is the posterior based on $\boldsymbol{y_{-i}}$

- The deviance information criterion (DIC) is a hierarchical modeling generalization of the Akaike information criterion (AIC) (Spiegelhalter et al., 2002). It is based on the expected log predictive density for a new data point (elpd), defined as:

$$E_f(log(p_{post}(\tilde{y_i})))$$

Specifically, the estimated DIC makes modifications to the elpd:

$$\hat{elpd}_{DIC} = log(p(y|\hat{\theta}_{Bayes})) - p_{DIC}$$

where p_{DIC} is the effective number of parameters, defined as:

$$p_{DIC} = 2(log(p(y|\hat{\theta}_{Bayes})) - E_{post}(log(p(y|\theta)))$$

where the expectation in the second term is an average of θ over its posterior distribution.

The actual quantity of DIC is defined in terms of the deviance rather than the log predictive density:

$$\text{DIC} = -2(log(p(y|\hat{\theta}_{Bayes}))) + 2p_{DIC}$$

Because of the negative term, lower values of DIC indicate better fit.

- The Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) is a more fully Bayesian approach for estimating the out-of-sample expectation, starting with the computed log pointwise posterior predictive density and then adding a correction for effective number of parameters to adjust for overfitting (Gelman et al., 2014a). The WAIC is based on the expected log pointwise predictive density for a new dataset (elppd), estimated by:

$$\hat{elppd}_{WAIC} = lppd - p_{WAIC}$$

where p_{WAIC} is calculated by:

$$p_{WAIC} = 2\sum_{i=1}^{n}(log(E_{post}p(y_i|\theta) - E_{post}(log(p(y_i|\theta)))))$$

which can be computed from simulations by replacing the expectations by averages over the posterior draws.

Similar to the DIC, lower values of WAIC indicate better out-of-sample predictive accuracy.

From these criteria we select a "best" model for the common mean process ($\mu_{s,t}$) from a model that uses only ACS data. Then following the specification of the common mean process, the thesis estimate the bias structure ($B_{s,t}$) with a joint model that uses both ACS and Twitter data. Note that these criteria serve as a general guidance only. The WAIC has not yet been shown to be valid with dependent data, while the DIC has been criticized by other scholars (see Gelman et al. (2014b)). Nevertheless, if one can find consistency across these measures, then one can argue that different criteria agree on which model has the best fit.

More substantively on the issue of predictive accuracy, the thesis evaluates the performance of the joint model over an "ACS-only" model using a "Leave-One-Year-Out" cross validation. That is, suppose that estimates from the ACS

in year t is the prediction target, we remove the latent variables from ACS in that year and compare the model predictions from the joint model with the predictions from the "ACS-only" model. The goal would be to recover the target estimates from the model predictions. Ideally, the prediction error for the joint model should be lower than the "ACS-only" model, which shows that Twitter information can help emigration prediction. This is particularly useful when we wish to forecast the future before survey data is available, or if we wish to fill in missing years without survey estimates to understand migration trends.

3 A Simulation Study to Assess the Model

3.1 Overview of simulation test

Before delving into the empirical data, it is important to understand if a joint model that analyzes two processes, one an unbiased process and one a biased process, improves prediction compared to a "simple model" that only analyzes the unbiased process.

To assess this, we use *simulated data* where we know the values of the true processes and the corresponding hyperparameters (e.g., the precision parameter for the ICAR part of a process). The primary goal of the simulations is to assess the predictive ability of the joint model compared to the simple model.

3.2 The simulation setup

Following the notation in the previous sections, consider a potential full space-time process where $X_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$. Potentially, one could simulate a spatial process (which only includes μ, θ_s, ϕ_s), a temporal process (which only includes μ, α_t, γ_t), a space-time main effects process without space-time interactions (which includes $\mu, \theta_s, \phi_s, \alpha_t, \gamma_t$), or a space-time interaction process (which includes all the terms).

Since we know the structure will be complex we focus on the full space-time interaction model for the true process, while for the bias process we consider different possibilities of a spatial process, a temporal process, a space-time main effects process, or a space-time interaction process.

Thus, the full range of possible scenarios are as follows. Let:

- $T_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$ be the true process
- $Y_{1,s,t} = T_{s,t} + \epsilon_{1,s,t}$ as an unbiased process that measures the true process with some random error

- $Y_{2,s,t} = T_{s,t} + B_{s,t} + \epsilon_{2,s,t}$ as a biased process that measures the true process with random error and a bias structure (i.e., $B_{s,t}$)

Then we compare the following scenarios for different types of $B_{s,t}$:

1. Spatial bias where $B_{s,t} = \theta_s^b + \phi_s^b$
2. Temporal bias where $B_{s,t} = \alpha_t^b + \gamma_t^b$
3. Space-time main effects bias where $B_{s,t} = \theta_s^b + \phi_s^b + \alpha_t^b + \gamma_t^b$
4. Space-time interaction bias where $B_{s,t} = \theta_s^b + \phi_s^b + \alpha_t^b + \gamma_t^b + \delta_{s,t}^b$

We specify the number of time periods to be 40, which represents a large enough number of periods to estimate temporal dependency. For the spatial geography, we use the 49 states in the United States (excluding Alaska and Hawaii) to simulate a spatial process within a connected geography. The geography thus mimics the empirical data, although the hyperparameters would almost definitely not be the same.

For the hyperparameters in the process, we set the hyperparameters as the following:

- For the parameters of the unbiased process:
  - Precision for spatial structure: $\tau_\theta = 1$, and the proportion of the marginal variance explained by the spatial effect $\Phi = 0.37$. This latter parameter is just a reparameterization of the spatial components (i.e., θ_s and ϕ_s), which leaves the underlying model components unchanged, also known as the BYM2 model (Riebler et al., 2016)
  - Precision for the random walk temporal structure: $\tau_\alpha = 1$
  - Precision for iid random temporal effects: $\tau_\gamma = 1$
  - Precision for space-time interaction: $\tau_\delta = 1$

- Precision for error term: $\tau_{\epsilon 1} = 4$

- For the parameters of the bias structure:

  - Precision for spatial structure: $\tau_\theta^b = 0.5$, and the proportion of the marginal variance explained by the spatial effect $\Phi^b = 0.54$.
  - Precision for random walk temporal structure: $\tau_\alpha^b = 0.5$
  - Precision for iid random temporal effects: $\tau_\gamma^b = 0.25$
  - Precision for space-time interaction: $\tau_\delta^b = 0.04$
  - Precision for error term: $\tau_{\epsilon 2} = 0.25$

  Note that the precision parameters of the bias structure are generally smaller compared to the true process, as often there is more variation in digital data which leads to larger variances. Also notice the subscript b to denote hyperparameters of the bias structure.

We then estimate a class of models that specify the bias structure as (1) spatial effects only; (2) temporal effects only; (3) space-time main effects without a space-time interaction; (4) space-time effects with a space-time interaction. We analyze a range of models to examine how sensitive the results are to mis-specification of the model. For instance, how would the results on predictive accuracy alter if the bias were spatial only, but the model specified the bias as temporal? As mentioned above, mis-specification could be either the case of under-specification (i.e., the model leaves out effects in the true process) or over-specification (i.e., the model specifies effects that are not in the true process), and it would be of importance to understand the implications of such mis-specifications for predictive accuracy.

Thus, as noted above, there are four types of biases, and there are four types of models. This yields a total of $4 \times 4 = 16$ scenarios. For example, one scenario

would be that the bias follows a space-time main effects structure, but the model specifies effects as spatial effects only.

3.3 Simulation Strategy

As the goal of the exercise is to determine whether the joint model has better predictive power than the simple model, we remove the last period of the unbiased process (i.e., time 40). This mimics an empirically relevant scenario where traditional surveys are not yet available but we already have digital data. For instance, the latest migration statistics for year 2019 may not yet be available, but from Twitter data we can already obtain estimates. The goal would be to predict the true process for this period (i.e., $T_{s,t} = T_{s,40}$).

To assess predictive accuracy, the thesis compares the root mean squared error (RMSE) of the simple model to the joint model. Specifically, the procedure of the simulations is as follows.

For each of the 16 scenarios:

1. Simulate the true process with the defined hyperparameters
2. Add the error term of the unbiased process ($\epsilon_{1,s,t}$) to the true process and create the unbiased process ($Y_{1,s,t}$)
3. Add the error term of the biased process ($\epsilon_{2,s,t}$) and the bias term ($B_{s,t}$) to the true process and create the biased process ($Y_{2,s,t}$)
4. Remove the last period of the unbiased process and code as missing values
5. Fit the model that only uses the unbiased process as data
6. Fit the joint model that uses both the unbiased process and the biased process as data

7. Use the posterior medians as the predicted values of the last period of the true process for both the simple model and the joint model

8. Compare the predicted values of the last period from both models to the last period of the true process and calculate the root mean squared error(s). Let $y_{s,t}$ be the true value for state s, year t and $\hat{y}_{s,t}$ be the predicted value from the model based on the posterior median, then:

$$RMSE_t = \sqrt{\frac{\sum_{i=1}^{S} \hat{y}_{s,t} - y_{s,t}}{S}}$$

9. Repeat the above steps 300 times and calculate the average across the simulations

One important detail is that in Bayesian models, one should specify priors for the hyperparameters for the spatial and temporal effects. We first use the default priors in INLA. In INLA, by default, the intercept of the model is assigned a Gaussian prior with mean and precision equal to 0. The priors for the hyperparameters of the random effects are specified with a lognormal distribution with parameters $\{1, 0.00005\}$ (Moraga, 2019). However, in a separate subsection below we also experiment with penalized complexity priors (PC priors), described in greater detail later (Simpson et al., 2017).

3.4 Results

3.4.1 Results with default priors

We first compare the predictive accuracy, measured by the RMSE, between the simple model and the joint model. The results are represented in Table 1. In the table, each row represents a different bias structure and model specification of the bias structure. For instance, the first row represents when the bias structure follows a spatial structure, and the model specification specifies only spatial

effects for the bias structure. In this particular case, the model is correctly specified. The columns represent the RMSE's for the simple model and the joint model.

In general, the joint model performs better in predictive accuracy than the simple model.

When the bias structure is spatial, all four joint models perform better than the simple model. However, the spatial model (the correct specification) yields the lowest RMSE. The "over-specified models" of the space-time main effects and space-time interaction models also perform relatively well. The temporal model, which does not include spatial effects, has the worse predictive accuracy of the four models. This is not surprising given that it is the only model that does not include spatial effects.

When the bias structure is temporal, the spatial model performs much worse than the simple model, as it mis-specifies the bias structure. The temporal model, space-time main effects model, and space-time interaction model are similar in predictive accuracy, and all are better than the simple model.

When the bias structure is space-time main effects (i.e., without a space-time interaction structure), again the spatial model performs worse than the simple model. The other three models again are similar in predictive accuracy, although it seems that the correctly specified model (i.e., specifying the bias structure as space-time main effects) yields a slightly lower RMSE. Nevertheless, all three models perform better than the simple model.

When the bias structure includes a space-time interaction component, the spatial model, the temporal model, and the space-time main effects model all perform worse than the simple model. However, the space-time interaction model has a very close RMSE to the simple model. This is reasonable as this is the only model that includes a space-time interaction effect in the model

and does not underspecify the effects. However, neither joint model has better predictive accuracy than the simple model.

In general, given a bias structure, the correctly specified model yields a better predictive accuracy than the simple model. "Over-specified models," such as including additional temporal effects when there is only spatially-structured bias, do not appear to decrease predictive accuracy much. However, "under-specified models" that leave out true processes in the model lead to much worse predictive accuracy. Thus, given that in empirical settings it is difficult to a-priori determine the bias structure, specifying the bias structure as a space-time interaction process appears to be a conservative but safer choice.

One thing to note is that when the bias follows a space-time interaction structure, even when the bias structure is correctly specified, the predictive accuracy of the joint model is not higher than the simple model which only uses the unbiased observations. One possibility is that in this special case perhaps there are too many parameters to estimate, especially for estimating the space-time interaction term. Nonetheless, in this worst case scenario a space-interaction interaction model for the bias structure can still achieve a predictive accuracy similar to the simple model. Another possibility is that the default priors are not adequate, and results could be improved using PC priors, which we would show is the case in subsection 3.4.2.

	Bias structure	Model for bias structure	RMSE simple model	RMSE joint model
1	space	space	1.77	0.74
2	space	time	1.77	1.52
3	space	spacetime main effects	1.77	0.78
4	space	spacetime interaction	1.77	0.77
5	time	space	1.81	4.07
6	time	time	1.81	1.68
7	time	spacetime main effects	1.81	1.75
8	time	spacetime interaction	1.81	1.52
9	spacetime main effects	space	1.78	3.74
10	spacetime main effects	time	1.78	1.73
11	spacetime main effects	spacetime main effects	1.78	1.56
12	spacetime main effects	spacetime interaction	1.78	1.75
13	spacetime interaction	space	1.77	3.95
14	spacetime interaction	time	1.77	2.07
15	spacetime interaction	spacetime main effects	1.77	2.03
16	spacetime interaction	spacetime interaction	1.77	1.78

Table 1: Comparison of predictive accuracy using default priors

3.4.2 Results using PC priors

One may wonder if the results are sensitive to the priors. Thus, this section experiments with PC priors. PC priors are a set of priors that are invariant to reparameterisations, are designed to support Occam's razor, and have excellent robustness properties (Simpson et al., 2017), suggesting better performance. Furthermore, PC priors are defined using probability statements about the parameter, which are easier to understand. PC priors penalise the complexity induced by deviating from the simpler base model and are formulated after the input of a user-defined scaling parameter for that model component, both in the univariate and the multivariate case, which are amenable to hierarchical models with many hyperparameters.

The construction of PC priors is based on the prior belief of the standard deviation of the precision parameter τ. The user provides a scale U which implicitly defines how far the model is allowed to deviate from the base model. In particular, the researcher chooses a prior such that $P(1/\sqrt{\tau}) > U) = \alpha$

where α is a predefined small probability, such as $\alpha = 0.05$. The appropriate U is derived after the type-2 Gumbel distribution for τ is integrated out.

We use independent PC priors for each of the hyperparameters, choosing $\alpha = 0.05$. As the thesis simulates from hypothetical data where the true hyperparameters are known, we can find the appropriate U that corresponds to $\alpha = 0.05$. From this testing, we can examine compared to the default priors, whether the "most accurate" PC priors aid predictive accuracy and recovery of the hyperparameters.

We start with the central concern on whether PC priors improve prediction. Table 2 compares the RMSE's of the simple model, joint model with default priors, and joint model with PC priors. For each row, the lowest RMSE is highlighted with bold text. As seen, when the bias structure is spatial, the PC priors and default priors produce similar results.

When the bias structure is temporal, PC priors appear to be produce lower RMSEs when the model includes a temporal effect (i.e., the temporal model, space-time main effects model, and space-time interaction model).

When the bias structure is both spatial and temporal, but without a space-time interaction, again the PC priors produce lower RMSEs when the model includes both spatial effects and temporal effects (i.e., the space-time main effects model and space-time interaction model).

Finally, for a space-time interaction bias structure, the PC priors again yield better RMSEs when all appropriate effects are included (i.e., the space-time interaction model).

In short, PC priors can produce better predictive accuracy than the default priors, but only when all process that are in the true process are included in the model specification (i.e., no under-specification). It is also important to discuss the implications of model mis-specification. Mis-specification due to

adding non-existent effects in the model appear to have minimal impact, but mis-specification due to leaving out existing effects in the true process can lead to large prediction errors. Thus, similar to the general implications regarding default priors, specifying the space-time interaction model that includes all the effects is the recommended approach, as it is often difficult for researchers to determine a-priori what spatial or temporal processes there are in the population.

Another thing to note is that under the case of a bias structure of space-time interaction, using default priors for the joint model did not yield better predictive accuracy than the simple model. However, as shown in this section, correctly constructed PC priors yield better RMSEs than the simple model.

	Bias structure	Model for bias structure	RMSE simple	RMSE joint (default priors)	RMSE joint (PC priors)
1	space	space	1.77	**0.74**	0.78
2	space	time	1.77	1.52	**1.48**
3	space	spacetime main effects	1.77	0.78	**0.77**
4	space	spacetime interaction	1.77	**0.77**	**0.77**
5	time	space	**1.81**	4.07	3.93
6	time	time	1.81	1.68	**1.47**
7	time	spacetime main effects	1.81	1.75	**1.50**
8	time	spacetime interaction	1.81	1.52	**1.39**
9	spacetime main effects	space	**1.78**	3.74	4.02
10	spacetime main effects	time	1.78	1.73	**1.62**
11	spacetime main effects	spacetime main effects	1.78	1.56	**1.48**
12	spacetime main effects	spacetime interaction	1.78	1.75	**1.39**
13	spacetime interaction	space	**1.77**	3.95	3.79
14	spacetime interaction	time	**1.77**	2.07	1.85
15	spacetime interaction	spacetime main effects	**1.77**	2.03	1.92
16	spacetime interaction	spacetime interaction	1.77	1.78	**1.42**

Table 2: Comparison of predictive accuracy using default priors

4 Empirical data using Twitter and ACS data

As the simulation results boost the confidence in the predictive power of a joint model over a simple model, we proceed with our case on state-level emigration rates in the United States.

The data in the thesis come from two main sources: (1) twitter data from the 1% historical archive of Twitter; (2) official statistics from the American Community Survey (ACS).

4.1 Data

4.1.1 Twitter data

The Twitter data used in this paper were assembled from a historical archive of a 1% sample of the Twitter stream (Archive.org, 2016) between January 1, 2010 and December 31, 2016. From this sample, we selected all tweets containing a geo-tag (i.e., information on latitude and longitude) that occurred within the United States, and use this information to place each tweet in one of 50 US states. Migration is defined as a change in residency between US states. The initial selection results in a sample of 554,229,541 tweets from 2,226,467 users.

While we have already discussed some of the reasons why Twitter data may be biased, there are also reasons to believe that these biases are not randomly distributed across time and space. First, one would expect the demographic composition of a state to be more or less stable across years, yielding consistent spatial effects. Second, the user base is not consistent over the period. While the mean number of geo-tagged tweets associated with each user is 267, these tweets do not necessarily occur evenly over time. On average, a user appears in a 24-week spread over a period of about a year. This means, for example, that the 2011-2012 interstate migration flows are estimated from a different set of users than the 2012-2013 flows. Thus, including time-specific effects when

capturing the bias becomes imperative. Third, geo-tagged tweets will pick up all kinds of movement – e.g., holiday travel, travel for business, and short-term mobility for education or family-related reasons – and not just the semi-permanent relocation associated with migration. Thus, different people will migrate for different reasons to and from different places. We can therefore expect to observe spatial patterns in the degree of bias in the Twitter data. For example, the popularity of destinations like Las Vegas and Miami might result in the overestimation of migration to Nevada and Florida.

For these reasons, one would expect migration estimates derived from Twitter data to be biased with respect to time and space. However, thesis shows that we can model the structure of these biases to better isolate the migration signal from the Twitter data.

4.1.2 American Community Survey (ACS) data

The ACS asks respondents to name the state where they currently live and the state/country where they were living one year before the interview. From the information on current and previous residence, the ACS produces estimates of state-to-state migration flows on an annual basis. We draw from ACS estimates for the years 2010-2016.

By aggregating the migration flows, one can obtain a point estimate for the emigration rate for each state (e.g., the number of migrants from Arizona is the sum of the migrants from Arizona $\rightarrow$ Florida, Arizona $\rightarrow$ Kentucky, etc.). We also compute the standard errors that acknowledge the survey design using the replicate weights in the ACS.

4.2 Obtaining estimates from Twitter data

Before starting the analysis, one would need to transform the raw Twitter data into emigration estimates for each state-year. To do this, we follow Zagheni et al. (2014) and uses the following procedure:

1. For each geo-tweet, we use the latitude and longitude to identify the US state from where the tweet was posted.

2. For each user and for each year, calculate the number of tweets posted in each US state.

3. For every two-year period (e.g., 2010 and 2011), discard users for whom the number of tweets posted in the modal state is less than three in at least one of the two years, or for whom the ratio between the number of tweets posted in the modal state and the number of tweets posted in the second modal state is less than three. For example, if a user posted 15 tweets in Washington state and eight tweets in Ohio in 2010, and 20 tweets in Washington state and three tweets in Ohio in 2011, the user would be discarded because the ratio in 2010 is less than three. This threshold is somewhat arbitrary, but was chosen based on the underlying goal of achieving a compromise between ensuring that the state of residence is identified accurately, while also maintaining a large sample of tweets.

4. For every two-year period, and for users who meet the threshold criteria described above, if the modal state in the first year is different from the modal state in the second year, the user is classified as a migrant. If the modal state is the same, the user is classified as a non-migrant. For the purposes of this thesis, we consider internal migration only. However, this approach could also be applied to international migration.

5. For every two-year period, t and $(t+1)$, calculate the estimated migration probability for the first year (defined as $\hat{p_t}$) in the state as $N_{Migrants}/N_{Users}$.

4.3 Assessing bias of Twitter estimates using the ACS

4.3.1 Overview of bias diagnostics

After obtaining the raw estimates from Twitter, it would be helpful to compare these estimates, which are drawn from a non-representative sample, to estimates from the ACS, which are based on a sample that is representative of the US population.

We assess the degree of bias using a simple bias ratio formula. More specifically, we define BR as the bias ratio. Let $\hat{p}_{s,t}$ be the estimate of the emigration probability for state s, year t. Let $\hat{p}_{s,t}^{\text{TWIT}}$ be the raw estimates from Twitter and $\hat{p}_{s,t}^{\text{ACS}}$ be the official estimates from ACS.

Then:

$$BR_{s,t} = \frac{\hat{p}_{s,t}^{\text{TWIT}}}{\hat{p}_{s,t}^{\text{ACS}}}$$

The bias ratio is used to assess the relative discrepancy between the Twitter estimates and the ACS estimates. If the Twitter estimates were perfectly in line with the the ACS estimates, the bias ratio would be one. Values larger than one indicate that the raw Twitter estimates overestimate the emigration rate (e.g., a bias ratio of 1.15 indicates an over-estimation of 15%), whereas values smaller than one indicate that the Twitter estimates underestimate the emigration rate (e.g., a bias ratio of 0.78 indicates an underestimation of 22%). The goal of this section is to assess the bias structure of the Twitter estimates. Simply knowing that Twitter estimates are biased does not help us improve the accuracy of our estimates and predictions of emigration trends; instead, we need to leverage

the potential spatial/temporal variation and spatial/temporal correlation of the bias in our statistical estimation model.

4.3.2 Visual diagnoses of bias over space and time

The first step is to visually diagnose the bias for each state over time. Figure 1 plots the bias ratios across states for each year. Each subplot shows the map of the bias ratios for a given year. The colors indicate the size of the bias. Red colors are associated with bias ratios that are larger than one; the darker the red color, the larger the bias. Conversely, blue colors indicate bias ratios that are smaller than one. Gray colors indicate that data are missing in the state for the year, which occurs in 2010 only.

As Figure 1 shows, there are red colors, but no blue colors. This pattern indicates that the Twitter estimates of internal migration rates are always higher the the ACS estimates. This is to be expected, given that, on average, Twitter users are younger than the general population (Mislove et al., 2011) and have higher mobility rates.

Furthermore, although the Twitter estimates tend to be upwardly biased, this bias is not randomly distributed. There appears to be spatial variation in the distribution of bias, as states vary consistently in their degree of bias. For instance, it appears that states like Alaska and Nevada have much larger biases over time (i.e., darker reds for each subplot), while states like Alabama have smaller biases over the period considered.

A potential diagnosis of the spatial correlation is that there are similar degrees of bias in the New England area and similar degrees of bias in the West Coast region. While these patterns are not obvious, we can see that these areas tend to have clusters of red in states that share neighbors.

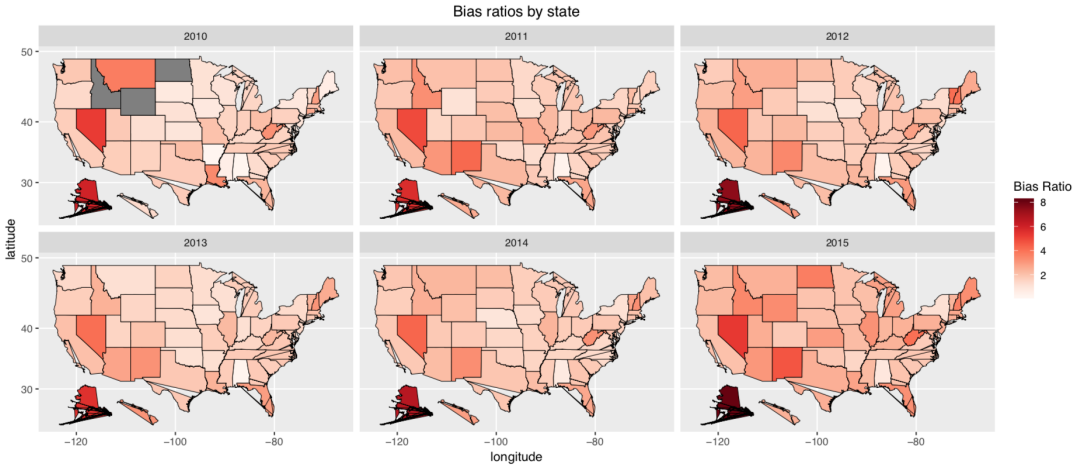

Figure 1: Map of bias ratios for each state across years (darker red colors indicate larger upward bias)

An alternative visual diagnosis is based on an examination of how these biases vary across states. We plot the bias ratios over the years in Figure 2. Each subplot represents a state, which is geographically mapped to its relative position in the United States. Within each subplot, the lines represent the bias ratios over the years for each state.

We can see that the relative positions of the lines are generally consistent over the years. This finding again suggests that there is spatial variation in the degree of bias, as states with higher overall levels of bias tend to have higher levels of bias across the years. For instance, Alaska and the District of Columbia have higher bias ratios, while Alabama has a very low bias ratio.

Additionally, there is evidence of temporal variation. The uptick that can be observed in many subplots leads us to conclude that, in general, the level of bias is higher in 2015 and is lower in 2010.

Finally, we see that most of the lines are relatively smooth with only a few bumps, which suggests that there is a temporal correlation, whereby the bias of the current year is related to the bias in the previous year.

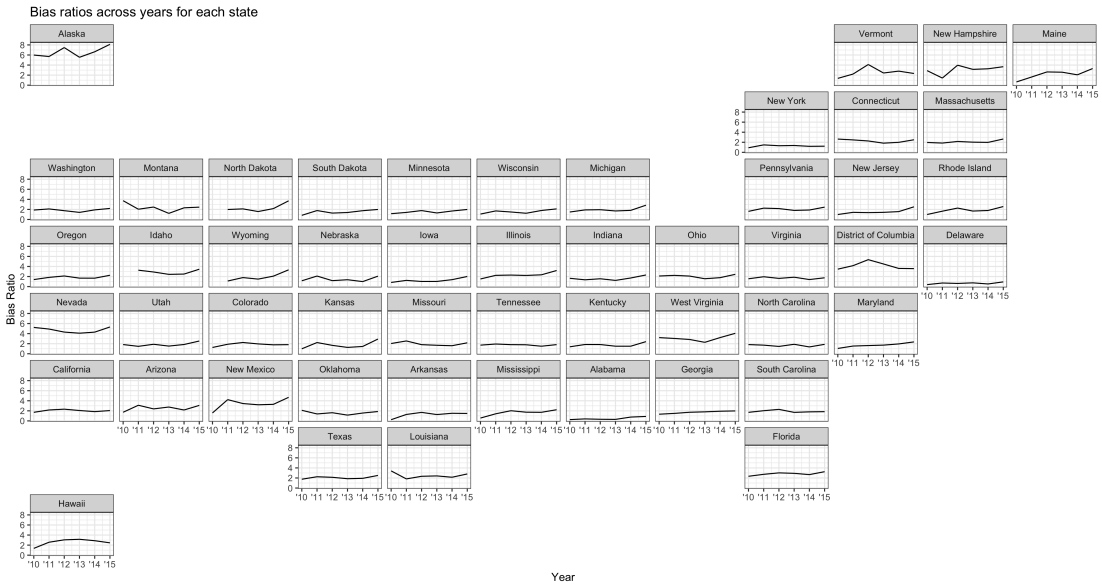

Figure 2: Lineplots of bias ratios within each state across years

We observed that the Twitter estimates are always higher than the ACS estimates, which indicates that the level of bias is overestimated if we use the raw estimates from Twitter. Nonetheless, the diagnostics also show that there is spatial and temporal dependency in the biases, which suggests that by capturing this spatio-temporal structure of the bias, we may be able to combine the Twitter and ACS estimates to predict emigration rates. The visual diagnoses give us some initial confidence that the use of a space-time model can help us capture the bias statistically, and, in turn, improve the accuracy of our predictions.

4.4 Selecting the best model for the true process and the bias process

Recall that the modeling strategy is to model the bias structure in addition to the true emigration process. Thus, before we model the structure of the bias,

we need to model the structure of the true emigration process. To accurately estimate this true emigration process, we construct a set of models from the unbiased ACS data to prevent a contamination of biases (i.e., "ACS-only models"). We consider different modeling options, including a spatial-only model, a temporal-only model, a space-time main effects model, and a space-time interaction model.

We compare the fit statistics for these "ACS-only models" in order to select the best model for the true emigration process (i.e., $\mu_{s,t}$). To reiterate, as Twitter data includes potential bias, it is better to consider only ACS data as the unbiased source of ascertaining the true process. As Table 3 shows, the space-time interaction model has the highest log-CPO, the lowest DIC, and the lowest WAIC; which suggests that it is the model with the best fit for the true emigration process. Thus, although each fit criteria has its criticisms, they do agree in this analysis on the best model.

Table 3: Comparison statistics for ACS-only models

	Spatial	Temporal	Space-time main effects	Space-time interaction
log-CPO	234	-134	233	324
DIC	-479	266	-479	-634
WAIC	-471	268	-471	-641

Next, the thesis takes on the task of statistically modeling the bias. After specifying the space-time interaction structure for the true emigration process, we explore different models that incorporate the bias structure (i.e, $B_{s,t}$). Again we consider a space-only model, a time-only model, a space-time main effects model, and a space-time interaction model.

As Table 4 shows, the space-time interaction joint model best captures the bias structure, as it has the highest log-CPO, the lowest DIC, and the lowest WAIC. From the comparison statistics, we select the joint model that specifies the true process as a space-time interaction process, with the bias structure

also having a space-time interaction process. In the next section, we evaluate whether this joint model outperforms the best "ACS-only model" in terms of forecasting and prediction.

Table 4: Comparison statistics for joint models

	Spatial	Temporal	Space-Time main effects	Space-Time interaction
log-CPO	361	382	411	424
DIC	-485	-549	-621	-723
WAIC	-480	-546	-613	-724

4.5 Forecasting/predicting emigration rates

4.5.1 Overview

To determine whether a joint model that utilizes Twitter data outperforms an ACS-only model, one needs to compare the prediction error from both models. As mentioned in Section 2.4, the thesis uses a "Leave-One-Year-Out" validation to test the ability to recover missing years from the model predictions. This means that we have one year for which official statistics are not available, and can only be predicted from existing models. This approach is also in line with the approach of the simulations.

To assess the prediction error, we use the Root Mean Squared Error (RMSE, a measure of absolute prediction error) and the Mean Absolute Prediction Error (MAPE, a measure of relative prediction error) to evaluate the performance of the models.

Regarding RMSE, let $p_{s,t}^{\text{ACS}}$ be the emigration rate for the ACS target year (i.e., the year with ACS data removed), and $\hat{p}_{s,t}$ be the predicted emigration rate from the model. Then the RMSE for year t would be:

$$RMSE_t = \sqrt{\sum_{s=1}^{51}(\hat{p}_{s,t} - p_{s,t}^{\text{ACS}})^2}$$

A lower RMSE indicates a more accurate prediction. We calculate the RMSE

for each year to see how well the models perform.

Conversely, since emigration rates tend to be low (often around 5%), the MAPE measures the percentage of the error compared to the target. Specifically, the MAPE for year t would be:

$MAPE_t = \sum_{s=1}^{51} |(\hat{p}_{s,t} - p_{s,t}^{\text{ACS}})/p_{s,t}^{\text{ACS}}|$

A lower MAPE indicates a more accurate prediction. We calculate the MAPE for each year to see how well the models perform.

4.5.2 Results on prediction error

We compare the RMSEs for both models for each year in Figure 3. In the plot, the horizontal axis represents the year for which we calculate the RMSE, and the vertical axis is the value of the RMSE. The red line represents the RMSE for the ACS-only model, while the blue line represents the RMSE for the joint model that utilizes the Twitter data. It is clear that for every year, the joint model has a lower RMSE than the ACS-only model.

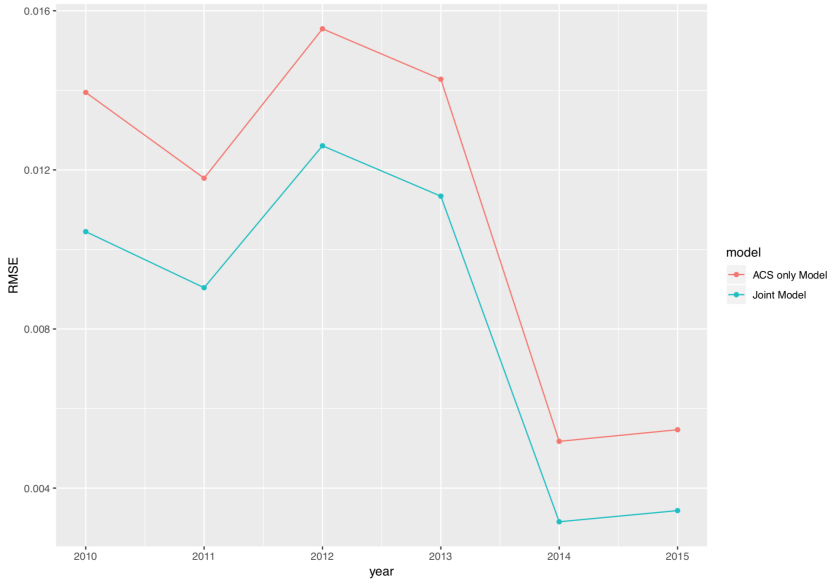

Figure 3: Comparison of RMSE of the joint model and the ACS-only model for each year

The thesis then compares the MAPEs for the two models in Figure 4. Again, we see that for every year, the joint model outperforms the "ACS-only model."

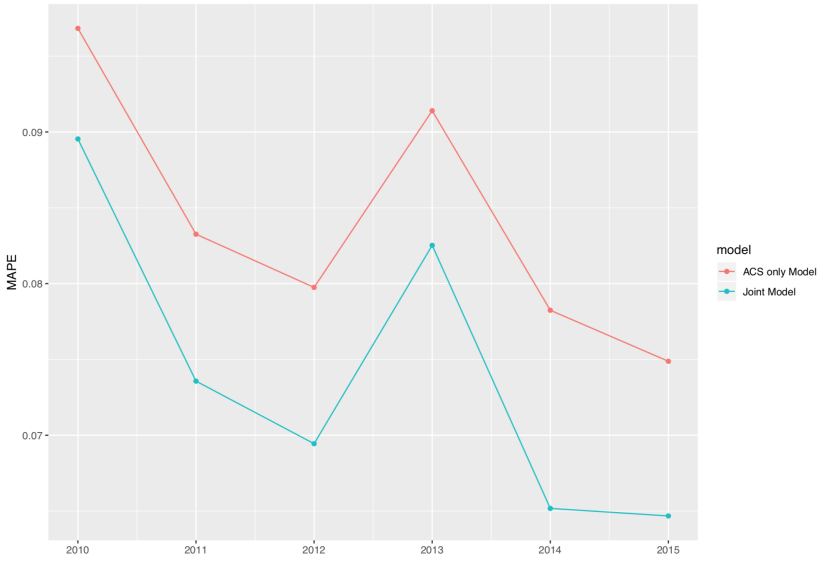

Figure 4: Comparison of MAPE of the joint model and the ACS-only model for each year

These results are encouraging, as they show that for both an absolute measure of error and a relative measure of error, the joint model outperforms the "ACS-only model." The findings therefore indicate that using Twitter data can improve the accuracy of predictions more than using ACS data only. Regardless of whether the goal is forecasting/nowcasting or filling in missing years for which there are no official statistics, it appears that models that include both digital data and traditional survey data produce better outcomes than models that rely on only one type of data. These results are in line with the results on the simulations, as they show that a joint model that draws from unbiased and biased sources of data can aid prediction.

4.6 Robustness checks with PC priors

The above analyses provided empirical evidence on how digital data can be combined with traditional surveys to predict migration processes. Nevertheless,

one may be interested whether the results are sensitive to priors, since the above analyses used the default priors in INLA. In particular, the results from the simulations indicated that correctly specified PC priors can further increase predictive power.

To reiterate, in PC priors, for each precision parameter, the user provides an U value and an α value. Generally, for precision parameter τ, the user specifies (U, α) so that $P(1/\sqrt{\tau} > U) = \alpha$.

However, unlike in the simulations, one cannot a-priori know what the appropriate value of U is. Thus, in this robustness check, we set $\alpha = 0.05$ and experiment with different values of U: $U = \{0.5, 1, 5, 10, 20\}$. Similar to the results from the simulations, we define these as independent priors for each hyperparameter. We then rerun the validation procedure and recomputes the RMSE to test if different priors affect the prediction error of the models.

Figure 5 shows the RMSE's under different PC priors. The results suggest that the prediction errors, whether from the default priors or the PC priors, are all better than the ACS-only model (i.e., the red solid line). In fact, in low values of U, the RMSE from models using PC priors are almost identical to results from the default priors, as on the plot the RMSE's overlap. However, under large values of U (e.g., 5 or 10) the predictive power further increases, as the results show lower values of RMSE. From the insights from the simulations, these are probably priors that are closer to the "correctly-specified" priors. Still, the main insight that regardless of the priors the joint model always outperforms the ACS-only model remains the same.

Figure 5: Comparison of default priors and different PC priors

5 Discussion and Conclusion

Our social lives have been increasingly digitalized. As a result of these technological developments, the amount of information on population and social processes that is available to social scientists is growing rapidly. While studies that use digital sources are proliferating (Zagheni et al., 2014; Hawelka et al., 2014; Jurdak et al., 2015; Fiorio et al., 2017; Hughes et al., 2016), a key problem that social scientists of all disciplines face is that these data come from specific groups in the general population. While they overlap to some extent, the users

of cellphones are not the same as the users of Twitter or the users of email. Although each digital source leaves traces of human activity with a high degree of spatial and temporal granularity, none can provide a representative sample of the general population.

Concerns about the non-representativeness and the biased inferences of digital data have plagued scientists who wish use these data to conduct rigorous research. This thesis's view on this issue is that to utilize digital data effectively, we need to conceptualize digital data as complementing, rather than replacing, traditional sources. We contend that recent advances in the statistical sciences allow us to make scientifically rigorous inferences based on a combination of representative and non-representative sources.

We showed from simulated data that a joint model that draws from unbiased and biased sources of data can aid prediction. Then from the empirical data, we showed that Twitter estimates are always more upwardly biased than ACS estimates. If one analyzed Twitter data alone, the results we would produce would be highly misleading. Nevertheless, the thesis also showed that by combining Twitter and ACS data, this bias can be modeled statistically. By decomposing the bias into spatial and temporal processes, we were able to estimate the bias structure, and incorporate information drawn from both Twitter and the ACS into a joint model that enhances prediction accuracy.

Although we illustrated how the method might be applied by using Twitter and ACS data to measure emigration rates, the method can be used to study other issues while drawing from other data sources. The requirements for the generic method would be (1) to establish a "gold standard estimate" drawn from official statistics with representative estimates, such as survey data; (2) while also taking into account (often biased) data that provide fine-grained information on the locations of individuals over time. This has been showed in

the simulations, as the synthetic data were purely hypothetical and were not tied to processes of migration or Twitter.

Thus, the method has many potential applications. First, although this thesis used Twitter data in our example, we also showed that the method can combine a wide range of unbiased and biased sources of data, such as cellphone data, social media data, administrative records, and various sources of geographical and temporal information.

Second, the generic method can be applied to cases beyond than that of emigration rates in the US. The model is a generic approach that merely specifies two processes: one that is unbiased and one that is structurally biased. Future applications of the model might include measuring immigration rates or migration stocks, or replicating the results in other countries.

Finally, the generic method is not limited to including only two sources of data. Although the thesis illustrated the method using Twitter and ACS data, the method could be easily extended to incorporate multiple sources of data. The method could, for example, take into account census data, large survey estimates from organizations, Twitter data, cellphone records, or email histories; while using different bias terms for each source of data.

In an age when humans are routinely leaving behind digital traces, combining new and traditional sources of data and methods will enable population scientists to investigate social processes in innovative ways. We look forward to future studies that adopt a perspective similar to the approach in this thesis.

References

Archive.org (2016). Archive.org of twitter 1% sample. https://archive.org/details/twitterstream. Accessed: 2016-09-30.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125.

Chantavanich, S. (2003). Culture of peace and migration: Integrating migration education into secondary school social science curriculum in thailand. *Asian Research Center for Migration, Bangkok*.

Clark, W. A., Deurloo, M. C., and Dieleman, F. M. (2000). Housing consumption and residential crowding in us housing markets. *Journal of Urban Affairs*, 22(1):49–63.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association*, 74:269–277.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., and Vinué, G. (2017). Using twitter data to estimate the relationship between short-term mobility and long-term migration. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 103–110. ACM.

Gelman, A., Hwang, J., and Vehtari, A. (2014a). Understanding predictive information criteria for bayesian models. *Stat. Comput.*, 24(6):997–1016.

Gelman, A., Hwang, J., and Vehtari, A. (2014b). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Greenwood, M. (1981). *Migration and Economic Growth in the United States*. Academic Press.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Hughes, C., Zagheni, E., Abel, G. J., Sorichetta, A., Wi'sniowski, A., Weber, I., and Tatem, A. J. (2016). Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data: Feasibility study on inferring (labour) mobility and migration in the european union from big data and social media data.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PloS one*, 10(7):e0131469.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015). Space-Time smoothing of complex survey data: Small area estimation for child mortality. *Ann. Appl. Stat.*, 9(4):1889–1905.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the united states. *Journal of Economic perspectives*, 25(3):173–96.

Molloy, R., Smith, C. L., and Wozniak, A. (2017). Job changing and the decline in long-distance migration in the united states. *Demography*, 54(2):631–653.

Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press.

Moretti, E. (2013). Real wage inequality. *American Economic Journal: Applied Economics*, 5(1):65–103.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and application*. Chapman and Hall/CRC Press, Boca Raton.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Sweeney, S. H. and Goldstein, H. A. (2005). Accounting for migration in regional occupational employment projections. *The Annals of Regional Science*, 39(2):297–316.

Tolstokorova, A. et al. (2009). Who cares for carers?: Feminization of labor migration from ukraine and its impact on social welfare. *International Issues & Slovak Foreign Policy Affairs*, 18(01):62–84.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2018). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*. To Appear.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 439–444, New York, NY, USA. ACM.

Zagheni, E. and Weber, I. (2012). You are where you e-mail: Global migration trends discovered in email data. In *WebSci '12 Proceedings of the 4th Annual ACM Web Science Conference*, pages 22–24.