UNRAVELLING THE QUASI-PERIODIC DYNAMICS OF GLOBAL INFLUENZA OUTBREAKS

Tsz Chai Fung*
University of Toronto
tszchai.fung@mail.utoronto.ca

Ryan J. Kinnear*
University of Waterloo
ryan@Kinnear.ca

Lei Sun*
University of Toronto
sunlei181@gmail.com

Richard Wu*
University of Waterloo
me@richardwu.ca
*

EXECUTIVE SUMMARY

This report seeks to understand and reveal the dynamics of global influenza outbreaks. The complicated patterns of shifted periodicity and significant reporting inconsistencies are the primary difficulties that we address. We perform a careful variable analysis which reveals firstly that countries in higher latitudes tend to be afflicted with earlier and more intense flu outbreaks; and secondly that flu outbreaks, while shifted in time each season, tend to *shift together* across different countries. We leverage these insights into a sophisticated time-series model capable of capturing the salient patterns in the observed flu reporting data. Significant hurdles regarding missing data are, at the same time, overcome with a rigorous supplementary imputation strategy. Ultimately, despite the inter-country and time-dependent heterogeneity in flu activity reportings as well as significant problems with missing data, our model is able to capture the outbreak dynamics and even provide plausible activity level estimates in cases where stretches of reporting data are entirely absent. This is a significant technical contribution, but we stress that countries, particularly less developed nations, need to improve influenza surveillance.

1 Background and Motivation

Throughout much of human history life has been "solitary, poor, nasty, brutish, and short" [2], and one of the most significant contributors to its nastiness has been infectious disease. Though modern man is largely spared from the most pernicious of these diseases, many underdeveloped areas of the world still suffer a terrible burden from widespread infectious disease, where lower respiratory tract infections (flu and pneumonia), Diarrhoeal diseases (e.g. Cholera and largely waterborne infections), and HIV/AIDS occupy three of the four leading causes of death (see [5], as well as Figure 7). We focus on infectious diseases specifically, even though cardiac diseases overall kill far more people worldwide, since more can be done in the immediate future to quarantine the spread of diseases.

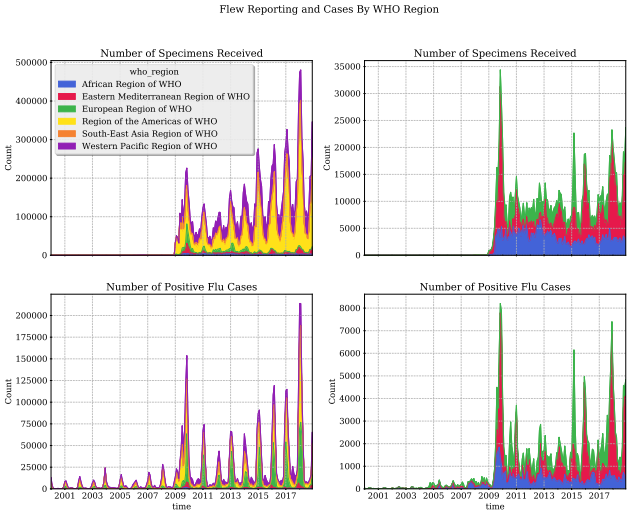
Our primary analysis will focus on influenza, since (1) flu and pneumonia remain primary causes of death from infectious disease in all areas of the world (Figure 7); (2) flu can lead to more serious complications like pneumonia which claims lives of people suffering immunodeficiencies (e.g. cancer patients, or HIV sufferers); (3) a deeper understanding of influenza outbreaks can lead to a more thorough understanding of disease outbreaks in general; and (4) influenza benefits from the most complete (though still quite limited) reporting.

2 Exploratory Data Analysis

This section briefly describes the basic patterns we have observed in the available datasets, as well as outlining the supplementary and external datasets we have used in the subsequent sections. A number of exploratory figures originally generated for this section have been relegated to the Appendix due to space constraints. Given that this report focuses on understanding flu outbreak intensities and patterns, the main dataset under consideration is the **influenza activity** dataset (`influenza_activity.csv`). This data consists of weekly data points (from year 2000 to 2018) and our focus is on the number of confirmed influenza cases, as well as the reports of "influenza-like-illness activity" levels (i.e. "No Report", “Sporadic", etc.) for each country. We do not account for different flu strains, seeking only to understand flu outbreak dynamics more broadly.

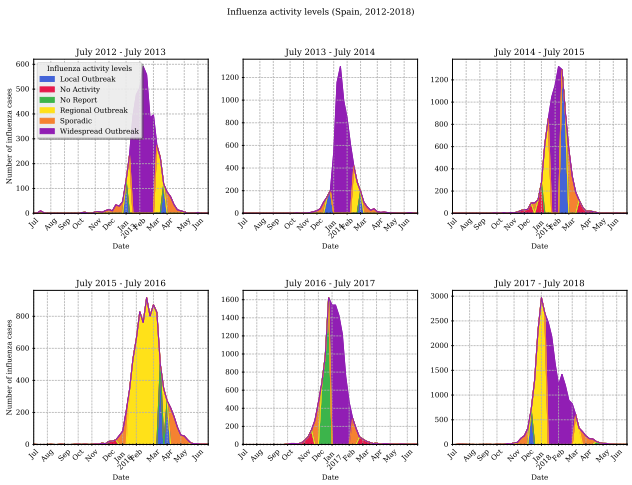
Preliminary analysis immediately reveals (consistent with everyone’s experience) that the numbers of influenza cases reported exhibit strong seasonal patterns (Figure 1a). Influenza activity is in general the most prevalent between October to April. On the

*Equal contribution.



(a) Flu Reporting by Region

Total number of specimens received and positive flu reported from different regions of the world across time. (TOP LEFT) Overall flu specimens, (BOTTOM LEFT) Overall number of positive flu cases, (RIGHT) zoomed in on low reporting regions.



(b) Flu Reports for Spain 2012 - 2018

Patterns in Flu reporting exist, and have potential for providing advanced outbreak warnings to nearby countries.

Figure 1: Worldwide Reporting Trends and Seasonality Patterns

other hand, when we extract countries from the southern hemisphere only, different patterns are observed: flu outbreaks usually occur from June to October instead. Globally, this simply indicates that flu outbreaks are more common in the winter months.

Apart from the regular seasonality patterns, the flu outbreaks also exhibit *quasi-periodicity*. By quasi-periodicity we mean that although the flu is typically prevalent in the winter, the beginning and end of major outbreaks are shifted in time by weeks or months, this can be seen in Figure 1b. We believe that this is one of the most salient features of flu activity dynamics, therefore a detailed analysis of this phenomenon is conducted in Section 4.1, and the methods by which we attempt to capture this phenomenon in our models is discussed in Sections 5.1 and 5.2.

One very undesirable characteristics of the influenza dataset is the reporting inconsistencies among countries. Most flu cases are reported from developed countries such as the US. In contrast, very few cases are detected in African countries, even if we have normalized the number of flu reported to the country populations. Obviously, it is not because developed countries are more prone to flu outbreak, but instead many developing countries do not posses the requisite resources and medical infrastructure to adequately identify cases of the flu. As a result, under-reporting (and missing data in general) is more prevalent in developing nations. We attempt to treat these issues rigorously (in particular, see Section 4.2).

Worse still, reporting inconsistencies are also heterogeneous over time. Figure 1a shows that the reporting of flu activity increased dramatically during and following the H1N1 (a "descendent" of the Spanish flu) pandemic of 2009. In fact, one can observe that the number of reported flu cases and outbreak activity levels in 2009 were abnormally high across all continents and regions as observed in Figure 1a. This data is important for developing flu forecasting models and to understand the spread of disease (Section), however, it is useless to provide reports only during the midst of pandemic outbreak. We can see from Figure 9 that the number of countries reporting flu activities declines sharply after 2014, which is an unsettling trend. It is critical for governments to continue consistent flu reporting in order to preempt outbreaks.

After identifying several basic features of the main influenza activity dataset, we explore several of the provided datasets that may reveal crucial information pertaining to flu outbreak intensities and patterns. We primarily draw upon the following datasets as a source of modelling covariates:

1. **Water quality** (`water_quality.csv`) provides for each country the yearly information (up to year 2015) of the percentage of the population using, respectively, at least basic water facilities, safely managed water facilities, and hand washing facilities (with soap). Due to the severity of the missing data problem and to align with the structure of the influenza dataset, which provides information up to year 2018, we have imputed additional values (see 4.2 for a detailed analysis).
2. **Sanitation** (`sanitation.csv`) provides the population (%) using at least basic and safely managed sanitation services respectively. The data structure and missing-value issue are similar to the water quality dataset with missing values imputed.

3. **Health indicators** (`health_indicators.csv`) includes basic yearly (up to 2017) demographic features. We find "population" to be useful information for normalizing the number of flu reports of each country to the number of flu reported. Population in 2018 is imputed based on a simple linear regression.

We further employ several external datasets which we believe are highly relevant to influenza, these datasets will be used alongside the aforementioned datasets without further explanation:

1. **Geographic information** (`country_info.csv`) from *Google Map Geocoding* summarizes the latitude and longitude for each country, as well as the continent each country belongs to.
2. **Human Development Index (HDI)** (`HDI.csv`) data is drawn from the primary source [6], useful for quantifying a general category of "rich" and "poor" beyond what is possible from the provided consumption data (`consumption.csv`). The few missing values in the HDI data were filled in by simple linear regression. HDI is typically used as a control variable that can absorb, for example, the impact of the development progress of a country to the under-reporting issue of the flu cases.
3. **Temperature** (`country_tmax.csv`) from National Oceanic and Atmospheric Administration (NOAA) records daily maximum temperature for each country. This allows us analyze how flu outbreaks are impacted by weather conditions, and to incorporate what we hypothesize as a key driving factor into our complete time-series model. Moreover, we are motivated by the fact that **weather can be effectively forecast** (at least relatively, by direct physical simulation), and therefore, if flu and weather are connected, flu outbreaks can be forecast as well.

3 Primary Hypotheses and Objectives

While the exploratory data analysis gives insights on the patterns of seasonal flu outbreaks, it still falls short on finding any underlying drivers of the pattern. For example, how are the flu outbreak patterns affected by various characteristics of a country? How does the quasi-periodicity of flu outbreaks interact among countries? Are there exogenous variables that would enable us to effectively forecast the outbreak of flu? Answers to these questions will help us build a more complete understanding of flu outbreak dynamics, and provide sound justification for the use of particular covariates in our final model. We formalize three questions of interest below:

1. **Flu outbreaks are connected to lower air temperature**, as observed in Section 2 that influenza outbreaks usually occur in winter. Based on this hypothesis, we expect that countries with higher latitudes are likely to have more severe flu outbreaks. Furthermore, we expect that higher latitudes will be associated with earlier seasonal outbreaks.
2. **Seasonal outbreak times are correlated across countries**. We have seen that the major outbreaks can be shifted in time by weeks or months between seasons, and we further hypothesize that the outbreaks across different countries are more likely to be shifted in same direction. This would indicate that the flu is either spread between countries, or that outbreaks are initiated by a common cause.
3. **Countries with poorer water facilities tend to have more severe flu outbreaks.** This is motivated by the common knowledge that hand-washing is the key means by which we can avoid contracting the flu. In particular, we hypothesize that the impact of water still exists after controlling the fact that richer countries tend to have better water facilities.

We begin in section 4.1 with a detailed analysis of covariate data, paying particular attention to the impact on the above hypotheses in order to understand and identify some underlying features that impact the intensities and patterns of flu outbreaks. This not only provides insights for policy makers to adopt measures to mitigate the severity of flu outbreaks, but also serves as a detailed covariate selection step. This detailed analysis is leveraged in the sequel in a more complete time-series model of flu activity and provides country-wide predictions and characterizations of future influenza outbreak characteristics (see Section 5.1).

Primary Modelling Objective The bulk of our modelling is focused on understanding the occurrence and intensity of influenza outbreaks. We emphasize that our ultimate aim is to construct a probabilistic model capable of capturing the salient features in the flu seasonal data, and that we are not necessarily driven to attempt to forecast the available time series, or achieve high cross-validation accuracy. Indeed, the data is marred so badly by missing data and under-reporting, that models designed purely to forecast the available data have little practical value. Instead, we seek a generative process which can be used to *plausibly estimate* what the *actual* (rather than reported) flu prevalence is over the reporting period, and to generate plausible future characteristics of influenza outbreaks. This estimation is necessarily extremely rough, and our model should accurately report on the uncertainty.

4 Detailed Auxiliary Modelling

In this section we describe in depth two of our important stepping stones. Firstly, Section 4.1 provides detailed analysis in support of our primary objectives, addresses the three main hypotheses proposed in Section 3 and justifies the use of particular covariates in our ultimate time series model of Section 5.1. Secondly, Section 4.2 details our approach to dealing with and correcting issues of missing data, which are pervasive in the given dataset, and indeed, a pertinent issue for global health data.

4.1 Covariate Analysis

To enable comprehensive covariates analysis that is useful in addressing the research hypotheses, we first need to *quantify* two *qualitative* technical terms mentioned in the research hypotheses: "(seasonal) flu outbreak severity/intensity" and "(seasonal) flu outbreak time". This is very challenging, especially since the reporting of flu activities are inconsistent across time and country, so we need to quantify them in a way such that we can *normalize* the reporting impacts. To draw appropriate conclusions, we then need to identify, justify, and fit statistical models to investigate the effects of several covariates (latitude, water facility, etc.) to the seasonal flu outbreak intensity and time.

In order to avoid unnecessary complication, we will focus this section's analysis only upon the northern hemisphere since northern/southern hemisphere flu seasons are complementary. Given that the seasonal flu starts around October and ends around March, in this particular analysis we will define study period t to be 1^{st} September of year $(t-1)$ up to 31^{st} August of year t.

Intensity measure via the Gini-Index For each country c and each year t, we will use the Gini-index as a measure of outbreak intensity. Since the influenza activity dataset provides weekly number of flu cases reported $\mathbf{y}_{ct} = (y_{ct}^{(1)}, \ldots, y_{ct}^{(W)})$ each c and t, the Gini-index is defined as

$$G_{ct} = \frac{\sum_{w=1}^{W}\sum_{k=1}^{W}|y_{ct}^{(w)} - y_{ct}^{(k)}|}{2W\sum_{w=1}^{W} y_{ct}^{(w)}} \tag{1}$$

where W is the number of recorded weeks. The statistic G_{ct} measures the concentration of the weekly number of flus reported in country c during study period t. A larger Gini-index represents a greater concentration, meaning that the majority of flus occur within the span of a few weeks, and therefore G_{ct} is a strong signal of flu outbreak severity. One of the principle advantages of using the Gini-index is that under-reporting issues are partially mitigated since $G_{ct} \in [0, 1]$, and is a *relative* measure within countries and within study periods.

Defining Outbreak Times The outbreak time H_{ct} will be quantified by a simple average weighted by the weekly number of flus reported, which also has the advantage of normalizing some under-reporting effects:

$$H_{ct} = \frac{\sum_{w=1}^{W} w y_{ct}^{(w)}}{\sum_{w=1}^{W} y_{ct}^{(w)}} \tag{2}$$

To assess the impacts of latitude, year, and water resource availability on flu intensities and outbreak times, we model the outbreak intensity G_{ct} and outbreak time H_{ct} respectively through two models: Firstly, a linear random effects model for intensity G_{ct}:

$$\begin{aligned} &G_{ct} = \beta_g^T \mathbf{x}_{ct} + U_t + W_{ct} \\ &U_t \sim \mathcal{N}(0, \sigma^2) \qquad W_{ct} \sim \mathcal{N}(0, \tau^2) \end{aligned} \tag{3}$$

and secondly a separate random slope model for H_{ct}:

$$\begin{aligned} &H_{ct} = \beta_h^T \mathbf{x}_{ct} + U_{t1} + U_{t2} L_c + W_{ct} \\ &(U_{t1}, U_{t2})^T \sim \mathcal{N}(0, \Gamma^2) \qquad W_{ct} \sim \mathcal{N}(0, \tau^2) \end{aligned} \tag{4}$$

where $\mathbf{x}_{ct}$ represents the covariates containing the latitude, continent, HDI, and water resource availability; L_c in the second model is the latitude of country c and we will see in the sequel why this is separated. The random effects are included to align with hypothesis 2: we want to investigate if there is a significant shift in the seasonal outbreak time *shared across countries*.

The seasonal flu outbreak intensity (Gini-index) and the outbreak time versus countries' latitudes are first presented, respectively, in Figure 2 and 3. In general, except for the European region, the positive relations between latitude and flu outbreak intensity as well as the negative correlation between latitude and outbreak time supports the first hypothesis of Section 3, that is, that outbreaks go hand in hand with lower temperatures. To gain more insight, we also provide similar plots including data points for the year 2009 alone since there was an outlier H1N1 outbreak that year. During this year, the flu intensities are particularly high and the outbreaks are particularly late. The anomalies are especially significant for countries nearer to the equator. This can be accounted for by the fact that the H1N1 flu originated from Mexico (latitude $\sim$ 23N) [1] during the summer (close to the end of study period) and spread globally including the countries on low latitudes. Although removing such data points yields slightly better fitting results, it does not lead to different conclusions, and so we do not present the additional results for lack of space.

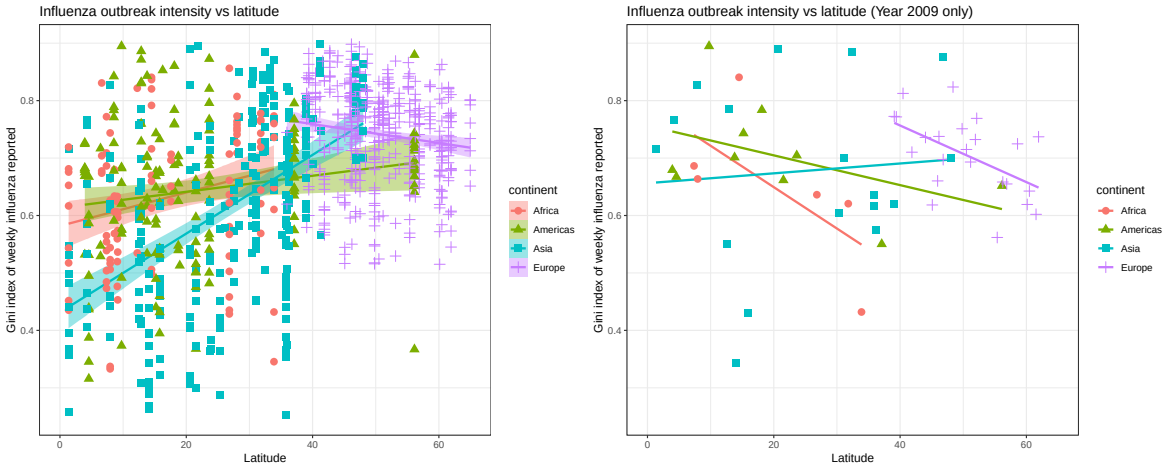


Figure 2: Influenza *outbreak intensity* vs latitude classified by continents.

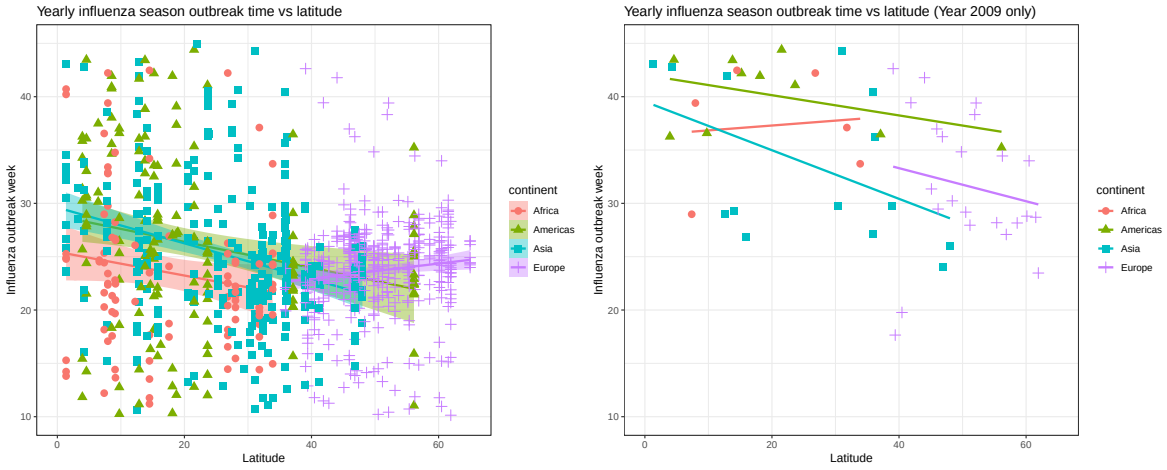


Figure 3: Influenza *outbreak time* vs latitude classified by continents.

4.1.1 Covariate Analysis Conclusions

Firstly, the results derived from the model of Equation (3) for the **flu outbreak intensity** are displayed in the left panel of Table 1. The observations are as follows:

1. Countries on a relatively high latitude are more prone to severe, high intensity flu outbreak, after allowing for some variables that are related to the latitudes of countries, such as the continent and the HDI. This can be seen from the positive, highly significant regression coefficient for the covariate "latitude".
2. Countries without high coverage of basic water facilities are more susceptible to severe flu outbreak, as indicated by two significant positive regression coefficients for "basic water facility".
3. The impact of the latitude to the outbreak intensity is greatly dependent on the continent. In particular, latitude has an opposite effect to the flu outbreak intensities for European countries. This is evidenced by the significant interaction coefficients.

And secondly, in the right panel of Table 1 shows the results of the model in Equation (4), **flu outbreak time**. Notable observations:

1. The outbreak times of seasonal flu are earlier for countries on high latitudes, as evidenced by the negative coefficient for "latitude".
2. The interactive effects between latitude and continent are significant, especially the relation between latitude and flu outbreak time, which is positive for European countries.

3. The variation in outbreaks time across years ($SD = 4.40$ for the shared random effect U_t) is comparable to those among countries ($SD = 5.00$ for W_{ct}). As a result we conclude that for each year there is a significant shift in the seasonal outbreak time that is *shared across countries*.

4. The magnitude of the shift in seasonal flu outbreak time increases with latitude. In other words, we are less certain about the flu outbreak time for countries in higher latitudes.

To further our analysis, we attempt to include yearly average winter temperature as a covariate in the above models, but we also realize a nearly perfect negative correlation between latitude and temperature. Hence, temperature and latitude produce similar impacts to the outbreak intensity and time, and it is difficult to include both covariates at once. Note that the temperature data for this analysis is condensed (the original data contains *weekly* temperatures instead of *yearly* aggregations), so the weather information may still be a useful predictor for our complete time series model, which uses the full *weekly* activity data (see Section 5.2).

Overall, this analysis supports the three aforementioned research hypotheses proposed in Section 3, except that the effects of latitude to the seasonal flu outbreak intensity and time for European countries are contrary to those suggested by hypothesis 1. **It is therefore legitimate for high latitude countries to pay additional attentions to flu outbreaks and devote extra resources for mitigation measures such as vaccines. Further, governments should be aware of any early outbreaks from other countries - this can be a signal of early flu outbreak for your own country!** We also realize the importance of including latitude (and/ or temperature information), water facility, continent and HDI to the predictive powers of seasonal flu activities and patterns.

	Influenza Outbreak Intensity G				Influenza Outbreak Time H			
	Coefficient	Std.Error	t-value	p-value	Coefficient	Std.Error	t-value	p-value
Intercept	0.5509	0.0455	12.1152	**0.0000**	29.6208	2.3959	12.3633	**0.0000**
Latitude	0.0043	0.0011	3.8619	**0.0001**	-0.1153	0.0569	-2.0277	**0.0429**
Continent (ref: Africa)								
- America	0.0431	0.0319	1.3513	0.1769	2.8331	1.4720	1.9247	0.0546
- Asia	-0.1065	0.0318	-3.3479	**0.0008**	4.2806	1.4718	2.9085	**0.0037**
- Europe	0.3138	0.0461	6.8001	**0.0000**	-3.9309	2.1364	-1.8399	0.0661
HDI	-0.0677	0.0563	-1.2018	0.2297	-6.2843	2.6015	-2.4157	**0.0159**
Basic water facility (ref: high)								
- no information	-0.0193	0.0248	-0.7768	0.4375	5.7715	1.1535	5.0034	**0.0000**
- low	0.0492	0.0191	2.5844	**0.0099**	-1.1015	0.8767	-1.2564	0.2093
- medium	0.0295	0.0107	2.7560	**0.0060**	0.5378	0.4903	1.0969	0.2730
Interaction (ref: Africa)								
- lat/Americas	-0.0012	0.0013	-0.9752	0.3297	0.0149	0.0578	0.2575	0.7969
- lat/Asia	0.0030	0.0012	2.4337	**0.0151**	-0.0400	0.0575	-0.6956	0.4868
- lat/Europe	-0.0058	0.0013	-4.4151	**0.0000**	0.1789	0.0606	2.9516	**0.0032**
Random effects U_t	Stddev				Stddev			
- Intercept	0.0210				4.3984			
- Latitude	–				0.0904			
- Residual W_{ct}	0.1086				4.9991			

Table 1: Model fitting results for the influenza outbreak intensity (LEFT) and the influenza outbreak time (RIGHT).

4.2 Estimating (imputing) Missing Data

While the previous subsection performs covariates analysis that is useful to thoroughly reveal the underlying patterns of flu outbreak and the importance of the covariates, such analysis does not include rigorous methodology in imputing missing covariates to avoid unnecessary complications. Since missing data is such a difficult issue, particularly in poor regions of the world, we follow a rigorous methodology for filling in missing data for some meaningful covariates: `water_basic`, `water_safe`, `water_soap`, `san_basic`, `san_safe`. These correspond to access statistics available from `sanitation.csv` and `water_quality.csv`. These covariates were explored in Section 4.1, and their use are also well motivated by medical considerations. The carefully imputed covariates are further incorporated in Section 5 and 6, which present a full time-series modeling of the flu patterns across countries.

Let C be the number of countries, R the number of WHO regions, and T be the number of years (in this case, 2000-2018). Ultimately, we view the missing data problem as a matrix completion problem for $P \in [0,1]^{C \times T}$, and in addition leverage some additional "side-information" provided by the HDI (denoted hdi_{ct}, for country c and time t). Note that we are not concerned with missing data in the HDI itself, as it is complete for almost all countries in our dataset. The natural region-based structure motivates us to employ a Bayesian hierarchical logistic regression model. Hierarchical models are effective for limited-data problems since each variable is

able to "borrow information" from the nearby regions. In the end, we fill in missing values for each covariate with the posterior mean from the model. We admit that each covariate is modelled separately for simplicity. Mathematically, our model can be described as

$$\begin{aligned}
\text{top level coefs: } & \mu_0 \sim \mathcal{N}(\hat{\mu_0}, \sigma_{\mu_0}^2),\ \beta_0 \sim \mathcal{N}(\hat{\beta_0}, \sigma_{\beta_0}^2), \\
\text{coef uncertainty: } & \sigma_\beta \sim \mathcal{N}_+(\hat{\sigma_\beta}),\ \sigma_\mu \sim \mathcal{N}_+(\hat{\sigma_\mu}), \\
\text{region coefs and uncertainty: } & \beta_r \sim \mathcal{N}(\beta_0, \sigma_\beta^2),\ \mu_r \sim \mathcal{N}(\mu_0, \sigma_\mu),\ \sigma_r \sim \mathcal{N}_+(\hat{\sigma_r}), \\
\text{observation model: } & P_{ct} \sim \mathsf{logit}^{-1}\mathcal{N}(\mu_r + \beta_r \mathsf{hdi}_{ct}, \sigma_r^2),\ c \in C_r.
\end{aligned} \tag{5}$$

Where $\mathcal{N}_+$ is a half-normal distribution (an intentionally light-tailed prior), and hatted variables (e.g. $\hat{\sigma_\mu}$) indicate that we have used a weakly informative data-dependent prior (e.g. the mean of some crude estimates, $10\times$ the data's standard deviation, etc.). The number of WHO regions is denoted by R, and the number of countries in each region is denoted by C_r. That the subscripts above range over all values is to be understood. Some of the variables P_{ct} are observed (which are what we condition on to obtain a posterior) and some of which are not (since the data is missing).

Weakly informative priors, light-tailed priors, and a limited number of variance parameters have been used partially for expedience, but also to aid with computation. We note that Automatic Differentiation Variational Inference (ADVI [4]) produced poor results for covariates with very little data (e.g. `water_soap`), but this problem disappeared when using Hamiltonian Monte Carlo (HMC [3]). We can also report that HMC sampling converged with only minor issues, as verified by checking the number of symplectic integration divergences (figure 4d), as well as computing Gelman-Rubin statistics with multiple MCMC chains – these diagnostics are included with `pymc3` (see [8]) and are reported automatically. Posterior predictive checks, as well as convergence diagnostics for ADVI and HMC generated during development are shown in Figure 4, where particular attention should be given to figure 4b which is representative of our final results (the similar figure in 4a is entirely preliminary and close inspection will admittedly reveal some oddities).

Our practical application of this model on the available data is somewhat crude, as we have not attempted to share data across covariates. However, highly accurate results from the covariate imputing stage are not critical to our overall model, particularly since the HDI is colinear with many of these variables. However, we continue to stress that **missing data is a non-trivial problem** and is important overall for the current application; we do not think it is acceptable to simply drop countries with under-reported statistics, or to fill values in with trivial estimates as **the countries suffering from poor reporting are often the ones that need the greatest consideration for mitigating infectious disease**. That is to say, social and health data is *not* missing at random.

5 Influenza Time-Series Modelling Methodology

With exploratory and technical issues out of the way, we are now able to describe our principle contributions.

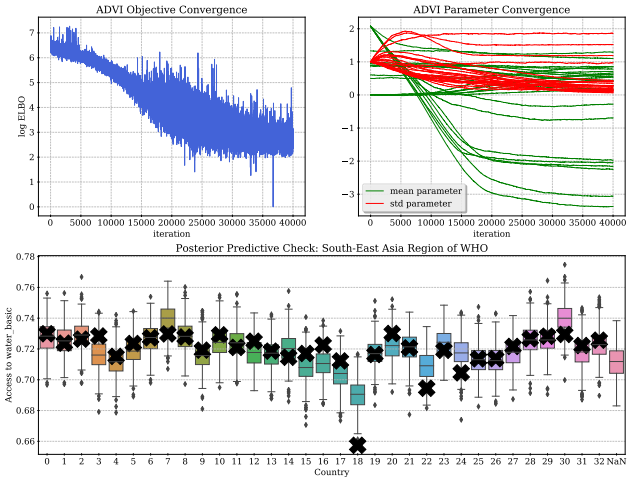
5.1 Bayesian Generalized Linear Count Regression with Quasi-Periodic Seasonality

Given influenza count data $f_c(t) \in \mathbb{N}$ for country c and time t, we consider a model in which $f_c(t) \sim \mathsf{Poi}(\lambda_c(t))$, that is, count data arises from a Poisson random variable with time-varying rate parameter. The (random) rate parameter itself is formed via a function of a linear combination of fixed covariates (e.g. sanitation availability), fixed but time varying covariates (sinusoids, weather data), random effects (simple random walk), and random state-based offsets. That is,

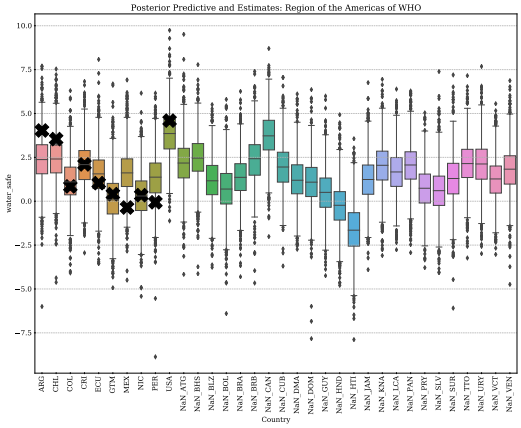
$$\begin{aligned}
f_c(t) &\sim \mathsf{Poi}(\lambda_c(t)) \\
\ln\lambda_c(t) &= \langle \beta_c^{(f)}, x_c^{(f)} \rangle + \langle \beta_c^{(T)}, x_c(t) \rangle + \sigma_c w_c(t) + S_{R_c(t)}^{(c)} \\
\sigma &\sim \Gamma(a, b),\ S_k^{(c)} \sim \exp(\phi_S);\ k = 1, \ldots, 6 \\
R_c(t) | L_c(t) &\sim \mathcal{MC}(P_{L_c(t)}),\ L_c(t) \sim \mathcal{CTMC}(Q)
\end{aligned} \tag{6}$$

where $\beta_c^{(f)}$ and $\beta_c^{(T)}$ are coefficients for non-random fixed and time varying covariates, $w_c(t)$ is a random walk (with jumps in weekly/monthly/yearly periods) having variance σ_c^2 which provides in some sense a measure of how well the fixed effects model the data, and $S_{R_c(t)}^{(c)}$ is an intensity level depending on the state of a Markov Chain $R_c(t)$. That c ranges over C, the collection of countries, is to be understood, and we will often drop the subscript in the sequel. Moreover, some hierarchical sharing, similarly to Section 4.2 can be deployed in this model as well, but we do not spell out the details.

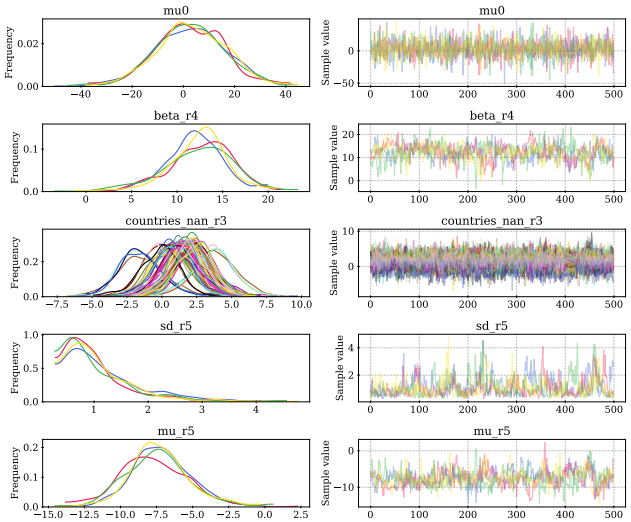
The notation $L(t) \sim \mathcal{CTMC}(Q)$ and $R(t)|L(t) \sim \mathcal{MC}(P_{L(t)})$ is used to mean that $L(t)$ (for "intensity *level*") is a stationary continuous time Markov Chain with rate matrix Q, and that $R(t)$ is a conditionally (on L) stationary discrete time markov chain with



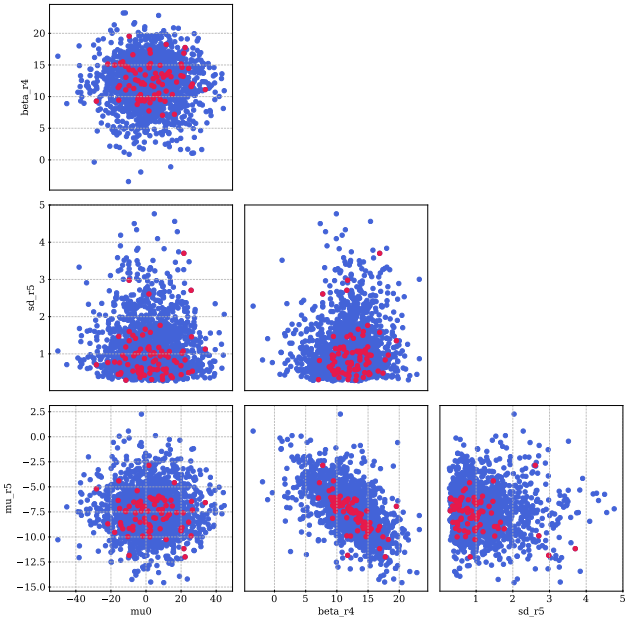
(a) ADVI Diagnostics



(b) Posterior Predictive Checks and Estimates of missing data.



(c) HMC Diagnostic Traceplots



(d) HMC Diagnostic Pairplots with Integrator Divergences

Figure 4: Missing Data: Posterior Predictive Checks, Estimates, and Convergence Diagnostics (UPPER-LEFT) model diagnostics and *preliminary* results on the target covariate `water_basic`. (UPPER-RIGHT) logit-scale proportion of individuals having access to safe water. Large black X markers in the left of the figure denote the means (across time and country) of known data points, we emphasize that **a lack of a black X indicates that the country has zero observed data**, and box plots denote posterior distributions (averaged on the time axis) sampled via HMC from our hierarchical model. (LOWER-LEFT) Illustrative traceplots for variables in `water_safe` model. We can see that the chain does not get "stuck" or meander inefficiently. (LOWER-RIGHT) Illustrative pairplot; divergences are marked red and do not concentrate in any one region – indicating robust results.

transistion matrix $P_{L(t)}$, which depends on $L(t)$. This encodes the random but quasi-periodic beginning and end times of each year's flu season, and is described in greater detail in Section 5.2. The reason that $R(t)$ is discrete is simply because our data is sampled on a weekly basis. The purpose of the Markov Chain is to drive the quasi-periodic "spikes" in the flu outbreak data: we believe that this is **one of the principle innovations of our approach** and overcomes significant difficulties of stability in autoregressive or other non-linear recursive models.

The linear component for the rate parameter $\lambda_c(t)$ is a fairly standard generalized linear model with fixed and random effects. Unfortunately, the use of a Poisson output is slightly restrictive as it is only a single parameter model (the mean and variance cannot be controlled separately). A fairly standard extension to this model is to instead use the two-parameter Negative-Binomial output distribution. However, we have chosen to avoid this modest complication for computational reasons: approximate inferences schemes including INLA [7] and ADVI had difficulties fitting the Negative Binomial model. We had some success using HMC for fitting simple Negative-Binomial models, but concluded the additional computational burden was not worth the benefits.

Dealing with Multi-Modality The attentive reader may notice that the likelihood of the above model is multimodal due to label-switching symmetries on S_k; this is a classic latent-variable headache and can cause serious computational problems for posterior inference, as well as rendering the posterior mean nearly meaningless. However, in our case this symmetry is naturally broken by enforcing an *ordering* on the variables, i.e. forcing the constraint $S_1 < S_2 < \cdots < S_6$, where now the state interpretation is natural: `No Activity` $<$ `Sporadic`, etc. In practice, this can be achieved with `pm.distributions.transforms.Ordered` in the case of `pymc3`, and an analogous transform in `stan`.

Fitting the Model The above described model is relatively complex, and we have resorted to an ad-hoc two stage estimation process. Firstly, the Markov chain component is estimated separately (see Section 5.2 and 5.3), and samples $R^n(t),\ n = 1, \dots N$ are drawn from the estimate. These samples are then fed into our software for fitting the Poisson regression component, obtaining samples $\lambda_t^{(c,n)};\ n = 1, \dots, N$ a final point estimate can produced via $\lambda_t^{(c)} = \frac{1}{N}\sum_{n=1}^{N} \lambda_t^{(c,n)}$, but it is generally preferable to plot each series in order to illustrate the model's uncertainty.

We also admit that some other model simplifications or minor modifications are occasionally deployed. In particular, we have actually used $\beta_s R_s(t)$ with $R_s(t) \in 0, 1$ providing a 1-hot state encoding for state s instead of the literal random variables $S_{R(t)}$, though this is essentially equivalent to the above formulation.

5.2 Hierarchical Markov Model $(R(t), L(t))$

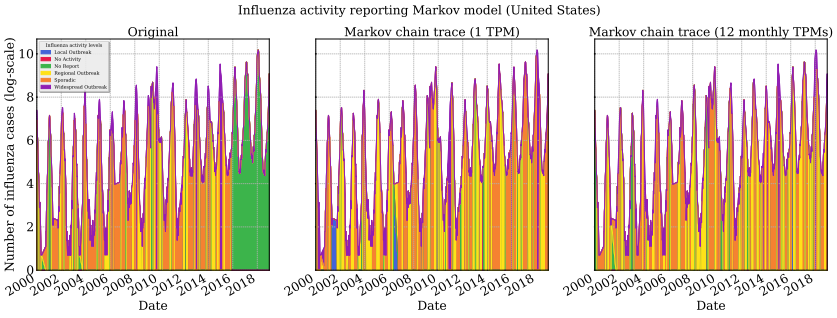
In reference to Section 5.1, $L(t)$ is a continuous-time Markov Chain (CTMC) with 2 states: `flu-season` and `off-season`, whereby the transition rate from `flu-season` to `off-season` should be faster than the reverse transition (i.e. flu seasons are shorter than off seasons). The discrete-time Markov Chain (DTMC) $R(t)$, conditionally stationary for each state in $L(t)$, models the week by week transitions between *influenza-like-illness (ILI) activity levels* reported by each country, specifically: `No Activity`, `Sporadic`, `Local Outbreak`, `Regional Outbreak`, `Widespread Outbreak`, and `No Report`. During flu seasons ($L(t) =$ `flu-season`) the chain $R(t)$ is intended to transition in some way through the higher intensity levels, while remaining at `No Activity` or `Sporadic` during the off season ($L(t)$ = `off-season`).

The Markov model serves two purposes: firstly as a generative process to fill in missing data (`No Report`) for countries that have never reported or stopped reporting ILI activity levels. Secondly, to **drive the quasi-seasonality in the Poisson count regression model** (Section 5.1). In regards missing data, see for example Figure 5a (left) we see that the USA stopped reporting activity levels after the week of November 9, 2015. Some countries do not report ILI activity levels *at all* (e.g. *France*, *Ukraine*, *Guyana*, etc.); We will see shortly that our Markov model is indeed capable of filling in plausible values for flu reports.

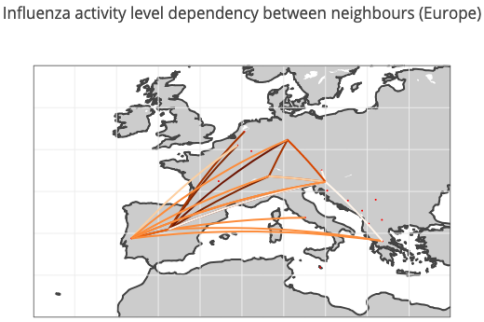
For our implementation, we admit that the Markov model is fit separately[2] from the Poisson regression model (see the bottom paragraph of Section 5.1) due to the complexity the latent variables. Computational techniques tailored to joint estimation of our model is an interesting topic for further work.

We first report initial results using maximum likelihood estimation (MLE) on a stationary DTMC $R_c(t)$ for every country c *without regard for seasonality*. That is, we estimated a single Transition Probability Matrix (TPM) $\hat{\Pi}_c$ for each country. We plot a sample trace generated from $R_{USA}(t)$ and $R_{Spain}(t)$ (where $R_c(0)$ is bootstrapped from the original data) in Figure 5a and 5c (middle). We observe that for $R_{USA}(t)$ the trace is able to generate plausible ILI activity levels for both the cross-validation period (2000-2015) and out-of-sample period (post-2015) (region in green: `No Report`). The trace from $R_{Spain}(t)$ is able to capture the `No Activity` (red) status in the valleys where flu cases are minimal and some of `Outbreak` statuses during the peaks. We note however that the more common `No Activity` status bleeds into actual outbreak spikes, a result of using a single stationary chain $R_{Spain}(t)$ for the entire series.

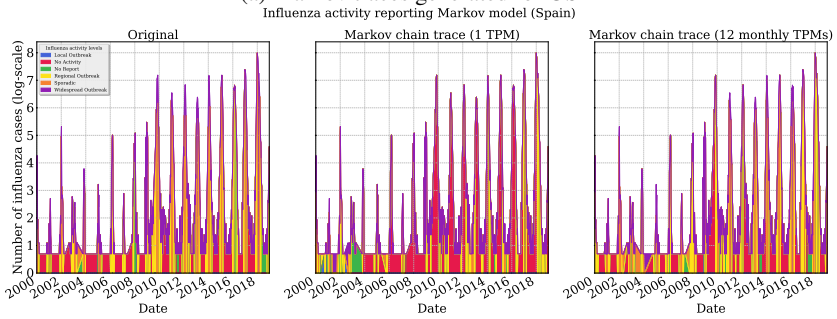
[2] this also helps with the distribution of labour



(a) Markov traces generated for USA



(b) Neighbour weights (Europe)



(c) Markov traces generated for Spain

Figure 5: (a) and (c): Original vs. Markov chain traces generated for *the USA* (TOP) and *Spain* (BOTTOM) from 2000-2018 on a weekly basis. The vertical axis is the log number of influenza cases whereby activity levels are broken down by colours. *The USA* is one example where reporting stops after a particular year. *Spain* is one of the countries with the most consistent ILI activity reporting. (LEFT) The original data for each country. (MIDDLE) A Markov Chain trace for each country c from a single DTMC $R_c(t)$. (RIGHT) A Markov Chain trace for each country c generated from monthly DTMCs $R'_{c,j}(t)$, $j = 1, \ldots, 12$.
(b): Trace lines representing learned importance weights as described in Section 5.3 between countries in the *South West Europe* WHO flu region. Smaller weights and larger weights correspond to lighter and darker shades of red, respectively. We observe that there is some cascading effect between *adjacent* countries.

In order to expand upon the single stationary model, we have replaced the originally intended continuous time component $L(t)$ of Section 5.1 and instead used a cyclo-stationary DTMC $R'(t)$ with a deterministic sequence of transition matrices $\hat{P}_c(t) \in \{1, 2, \ldots, 12\}$ where each TPM corresponds to activity levels in a calendar month, and we calculate the MLE from data for each country c. Sample traces for $R'_{USA}(t)$ and $R'_{Spain}(t)$ are shown in Figure 5c (right). We observe that for $R'_{Spain}(t)$ the additional parameterization allows us to capture more `No Activity` levels during valleys and more `Outbreak` levels at the peaks compared to the completely stationary chain $R_{Spain}(t)$.

Traces from these Markov simulations are generated offline and fed into the Poisson Regression described in Section 5.1.

5.3 Neighbourhood Markov Model for ILI Outbreak Levels

We propose a *Neighbourhood Markov Model* to (1) impute traces for countries with completely missing ILI outbreak data (e.g. *France*) and (2) estimate the shared behavior of flu outbreaks between countries in a *neighbourhood* as observed in Section 4.1.1. Given n countries and r disjoint neighbourhoods $\mathcal{N} = \{N_j : N_j \subseteq \{1, \ldots, n\}, j = 1, \ldots, r\}$, let $x_{c,t}$ denote the outbreak level reported by country c at time t, and Π_c as the 1-step transition probability matrix (TPM) for the DTMC $R_c(t)$ defined in Section 5.2, where $[\Pi_c]_{a,\cdot} = \pi_{c,a}$ is the transition probability vector out of activity level a. Then we infer $x_{c,t+1}$ the activity level at time $t+1$ as a function of its neighbours' outbreak levels in neighbourhood N_j where $c \in N_j$, that is

$$
\begin{aligned}
\alpha_{c,a,k} &\sim \mathcal{N}_+(\hat{\sigma}^2_{c,a,k}) \qquad k = 1, \ldots, 6 \\
\pi_{c,a} &\sim \text{Dir}(\alpha_{c,a,1}, \ldots, \alpha_{c,a,k}) \\
w_{c,j_\ell} &\sim \mathcal{N}_+(\hat{\sigma}^2_{c,j_\ell}) \qquad \ell = 1, \ldots, |N_j| \\
x_{c,t+1} \mid x_{c,t}, x_{j_1,t}, \ldots, x_{j_{|N_j|},t} &\sim \text{Cat}\Big(\frac{1}{Z}(\pi_{c,x_{c,t}} + \sum_{\ell=1}^{|N_j|} w_{c,j_\ell} \times \mathbb{1}_{x_{j_\ell}})\Big)
\end{aligned}
\tag{7}
$$

where $\mathbb{1}_k$ is the one-hot vector v where $v_k = 1$ and Z is some normalization constant such that the above parameterizes a valid Cat$(\cdot)$ distribution. Note that for each country c we also infer positive weights w_{c,j_ℓ} for each neighbour j_ℓ that measures the contribution of the neighbours' outbreak levels to the country's outbreak levels at the next time step. We can simultaneously fit the DTMC $R_c(t)$ as a direct result of the posterior of $\pi_{c,a}$. Considering the overwhelming amount of missing data, the high parameterization space makes this computationally infeasible. Instead we fix Π_c as the MLE TPMs we estimated in Section 5.2.

We choose to implement the model in Equation 7 as a shallow neural network in `PyTorch` which will allow us iterate through hundreds of epochs to convergence in a couple of minutes. The cost function is the categorical cross-entropy between the predicted transition probabilities $\hat{\pi}_{c,x_{c,t},\cdot}$ for $x_{c,t+1}$ and the actual activity level at $t+1$, augmented with an ℓ_1 norm on the weights assigned for its neighbours, that is

$$\mathcal{L}(\hat{\pi}_{c,x_{c,t}}, x_{c,t+1}; w_c) = -\sum_{k=1}^{6} \mathbb{1}_{[x_{c,t+1}=k]} \log \hat{\pi}_{c,x_{c,t},k} + \lambda \|w_c\|_1 \tag{8}$$

where $\mathbb{1}_{[a=b]}$ is the indicator function. We minimize Equation 8 for each neighbourhood, which we defined for simplicity as the WHO flu regions, with λ adjusted accordingly for each region to identify the salient neighbour weights. To illustrate the learned weights, we trace lines between countries in the *South West Europe* flu region in 5b where a higher weight corresponds to a darker shade of red.

Due to Equation 8 it does not make sense to minimize $x_{c,t+1} =$ `No Report` for the countries that have mostly (or all) `No Report` outbreak levels since we want to impute the probable outbreak levels commensurate to the number of flu cases. During training we introduce a slight bias that outbreak levels of the `No Report` countries follow the *distribution of its neighbours* by sampling from neighbours' outbreak levels in the previous time step for our new target $\tilde{x}_{c,t+1}$. We are then able to produce Markov traces for countries with entirely `No Report` outbreak activities from the weighted effects of neighbours for our Poisson Regression model described in Section 5.1.

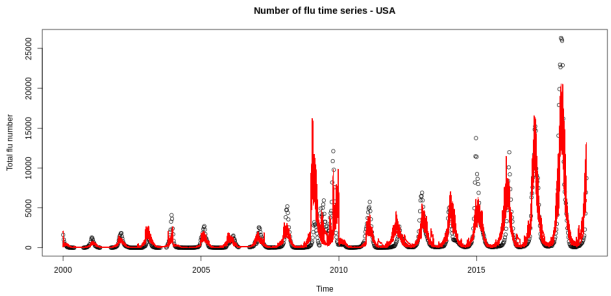
6 Analysis and Results

The estimation of our Poisson regression model of Section 5.1 is carried out via Integrated Nested Laplace Approximations (INLA) [7], a method capable of providing reasonable approximate inference quickly enough to iterate on our model. However, such estimates are known to underestimate the posterior variability, which is likely part of the reason we will see fairly thin credible intervals in our results. In order to incorporate traces from our Markov chain model, we are fitting one Poisson regression model for each (of 30) Markov traces, and drawing 100 posterior samples in each case. The visualizations in this section overlay each of these traces on top of the actual observed data in order to help visualize the variability in the posterior.

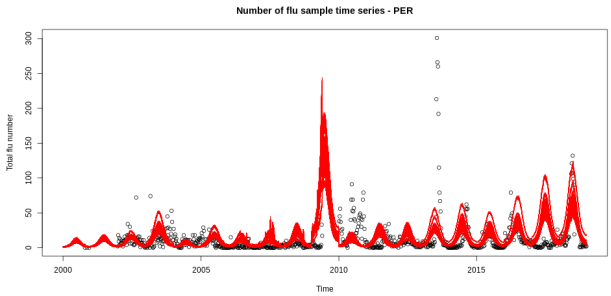
We estimate first a single observed data series (the USA) since this region has the highest data quality. Country specific fixed effects (water quality, HDI, etc.) are excluded in this first case, therefore the Poisson output intensity $\lambda_{USA}(t)$ is formed from a linear combination of (1) weather data (t_{max}), (2) Markov Chain samples $R_{\text{USA}}(t)$, (3) sinusoidal components, (4) (logarithm of) the population series and finally (5) a random effect: random walk with 1-year period. We assign simple Normal priors to each of the fixed-effect coefficients. The results are given in Figure 6a and Table 2.

From the figure, we can draw the qualitative conclusion that our model has the capacity to capture the complex flu activity dynamics. A more careful analysis of the tabulated results allows us to attempt to quantify these observations. Firstly, none of the posterior credible intervals contain the null hypothesis $\beta = 0$, leading us to conclude that each of our covariates is useful for explaining at least a portion of the variability of the output. Secondly, we believe our Markov model is correctly driving quasi-periodic state switching behaviour due to the ordering of the coefficients on the 1-hot state encodings conforming to intuition. Finally, the coefficient on t_{max} is positive, contrary to our initial expectations as guided by Section 4.1. There are many reasons why this may be the case, and we would consider it a topic for further exploration. We note that in Section 4.1 we observed a significant relationship between latitude, temperature and *flu intensity* as measured by the Gini-index, which is a subtly different measure than *the actual counts of flu cases*. In addition, since the incubation period for flu is at least a week long, we believe that there may be a time-lag effect relating to temperature which would be captured in the Gini-index statistic (since it evaluates the whole period) but not contemporaneously by our Poisson regression. Finally, and perhaps most obviously, a single temperature statistic for the entire US (a geographically diverse region) may simply not be enough to infer meaningful results.

Having concluded our analysis of the simple USA-only model, we expand into a larger region: WHO Americas. Even after imputing significant amounts of missing data, the influenza reporting in many countries is extremely limited, and we have only been able to generate complete results for a model including 8 countries in North and South America. In this inter-country model, we are including the same covariates as the USA only model (described above) as well as the additional `water_safe`, `HDI`, and `latitude` covariates, and interaction terms between latitude and the fixed sinusoids which helps with the latitude dependent seasonality. We have resorted to excluding a number of our other country-specific fixed effects (sanitation, etc.) due to constraints on computational resources. Finally, we include an additional cross-country random effect which quantifies the variability between each country in the model. m.Posterior samples are given in Figure 6b, and quantitative results in Table 3.



(a) US Results from Single Country Model



(b) Peru Results from Inter-Country (Americas) Model

Figure 6: Posterior Samples for Flu Count Data from the Model of Section 5.1
(LEFT) Posterior samples (100 samples for each of 30 Markov Chain traces) drawn for estimating our model on the US alone, i.e. without any country specific covariates. Good data availability makes this a simple target series. (RIGHT) Posterior samples drawn from Peru in an inter-country model for the America's WHO region. We emphasize that this region suffers from a significant number of missing data points that our model is attempting to correct, such as the imputed flu outbreak spikes from 2000 to 2002, and the 2009 pandemic that was poorly reported in Peru.

height	0.025quant	0.5quant	0.975quant
(Intercept)	-0.6737	-0.6522	-0.6308
tmax	0.0628	0.0639	0.0649
t_sin	0.9696	0.9748	0.9801
t_cos	1.8756	1.8895	1.9033
No Activity	-4.8249	-4.2039	-3.5834
Sporadic	-0.5572	-0.5379	-0.5186
Local	0.0762	0.0952	0.1142
Regional	0.2749	0.2937	0.3126
Widespread	1.0391	1.0568	1.0746
sd for year	0.5281	0.7007	0.9584

Table 2: US model posterior result

height	0.025quant	0.5quant	0.975quant
(Intercept)	33.8897	37.2264	40.5604
HDI	-16.0203	-15.4009	-14.7819
water_safe	-36.5779	-36.0238	-35.4703
tmax	0.0185	0.0194	0.0202
t_sin	0.0177	0.0231	0.0284
lat	-0.1975	-0.0766	0.0441
t_cos	-0.3311	-0.3243	-0.3175
No Activity	-4.0417	-3.9165	-3.7914
Sporadic	-0.6753	-0.6620	-0.6487
Local	0.1633	0.1760	0.1887
Regional	0.8820	0.8945	0.9069
Widespread	1.1928	1.2048	1.2167
t_sin:lat	0.0175	0.0177	0.0179
lat:t_cos	0.0412	0.0415	0.0418
sd for country code	3.6982	4.6137	5.8415
sd for year	0.5046	0.6654	0.9151

Table 3: Americas model posterior result

Similarly to the USA-only case the Figure 6b shows that the model is sufficiently expressive to model the flu dynamics. In addition, Peru is a region with a significant number of missing observations, which our model fills in with plausible estimates. In regards to the tabular results, we note that coefficients on `HDI` and `water_safe` are large in magnitude, but that this doesn't necessarily indicate that these features are extremely important, since we have not done any careful scaling of the covariate data. Instead, we are to conclude simply (since the credible intervals do not contain zero) that these covariates have at least some use in explaining inter-country variability. Similarly to the USA case, the ordering on the Markov Chain outbreak level state encodings corresponds to our expectation, even considering the differing reporting tendencies of each country.

6.1 Concluding Remarks

The systematic inconsistent reporting of data on influenza outbreaks and cases pose a challenge to analyzing and predicting the dynamics of the infectious disease. **We have rigorously accounted for the uncertainty in the data and attempted to produce meaningful insights into global influenza patterns.** The Poisson Regression mixed-effects model with an intensity-level Markov Chain component we've constructed is able to capture much of the nuances of global influenza outbreak patterns, such as the quasi-periodicity of flu seasons. A key takeaway from our analysis is that we should support country-level reporting of influenza data in order to develop a more complete picture of influenza, especially developing countries that have significant downsides when it comes to mortality rates related to influenza (see Figure 7). Finally, our model can be propagated forward in time to produce forecasts and therefore **we have made an important technical contribution: our model can be leveraged by policy makers to assist in preempting future flu epidemics, thereby mitigating the worldwide influenza disease burden.**

References

[1] Adrian J Gibbs, John S Armstrong, and Jean C Downie. From where did the 2009'swine-origin'influenza a virus (h1n1) emerge? *Virology journal*, 6(1):207, 2009.

[2] T. Hobbes and J.C.A. Gaskin. *Leviathan*. Oxford world's classics. Oxford University Press, 1996.

[3] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[4] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[5] World Health Organization. The top 10 causes of death, 2018.

[6] United Nations Development Program. Human development index data, 2018.

[7] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[8] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

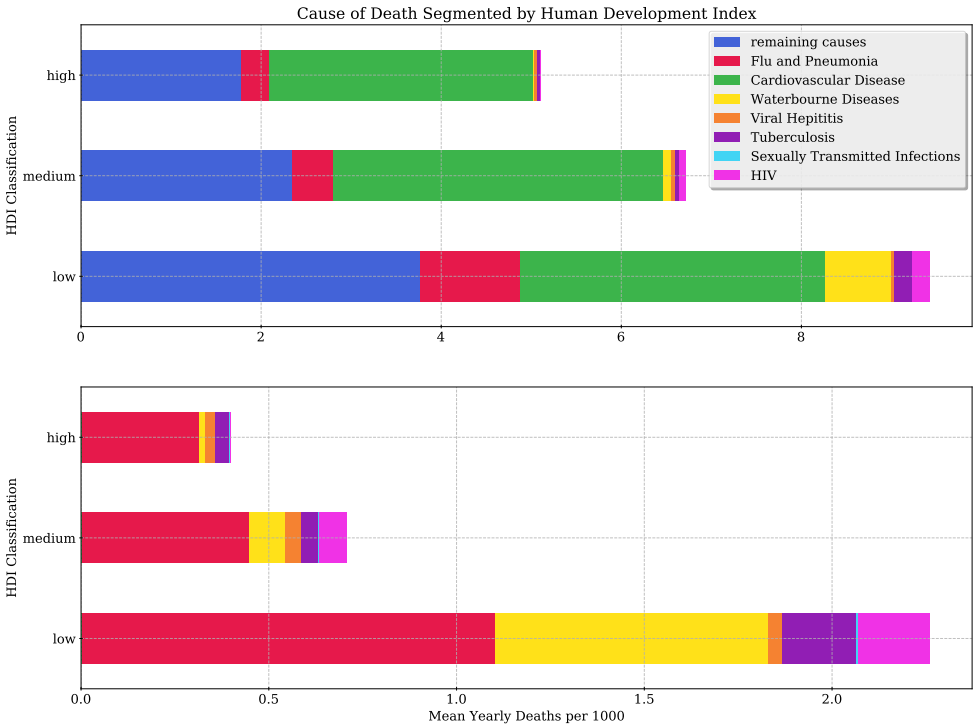
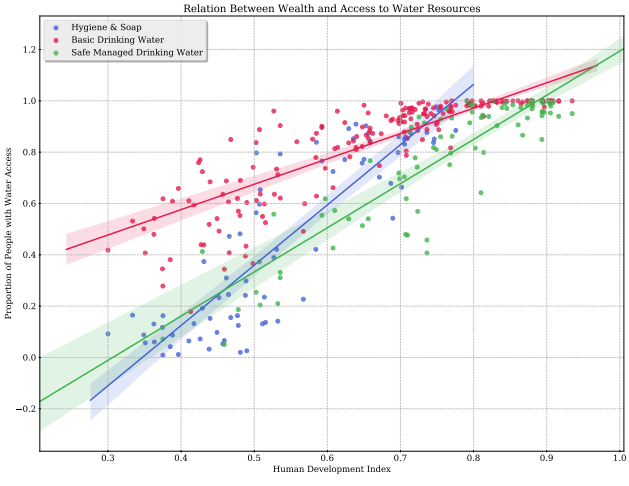


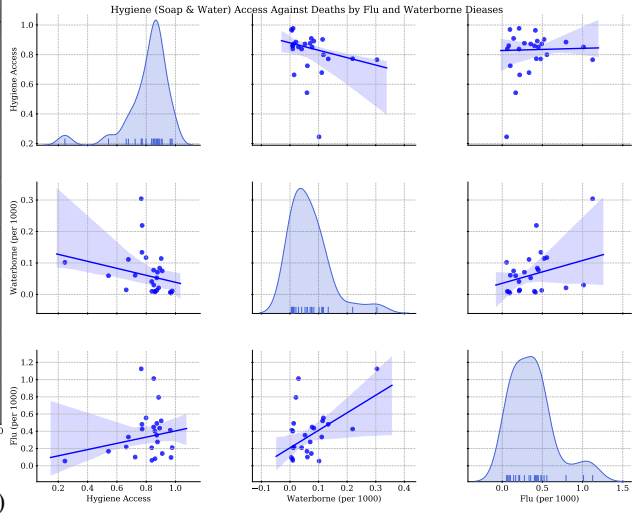
Figure 7: Cause of Death by HDI
(TOP) Total death rates per 1000 people in each HDI bucket. (BOTTOM) Similarly to the TOP figure, excluding cardiovascular diseases and "remaining causes". Causes of death were collected from the available data (`mortality.csv`) and joined HDI. The figure focuses in on infectious diseases, particularly ones which are relevant to our available dataset (i.e. Influenza, waterborne diseases, STIs / HIV). We see that flu and pneumonia remain primary causes of death from infectious disease in all areas of the world. A principle contributor to the "remaining causes" are cancers.

A Additional Figures

We cannot help but include here some additional figures, each of which has only minor or tangential importance to our main thesis.



(a) HDI Compared with Access to Water Resources



(b) Comparison of Hygiene Access to Flu and Waterborne Disease. Plot is produced by crude matchings and is marred by data collection and underreporting issues.

Access to different water resources (`water_quality.csv`) regressed against HDI. Each data point represents a single country with coordinates formed from the medians over the period $2000 - 2015$ of their per capita water access and HDI indicators.

Figure 8: Exploratory Regression Plots: Flu, HDI and Access to Hygiene Resources

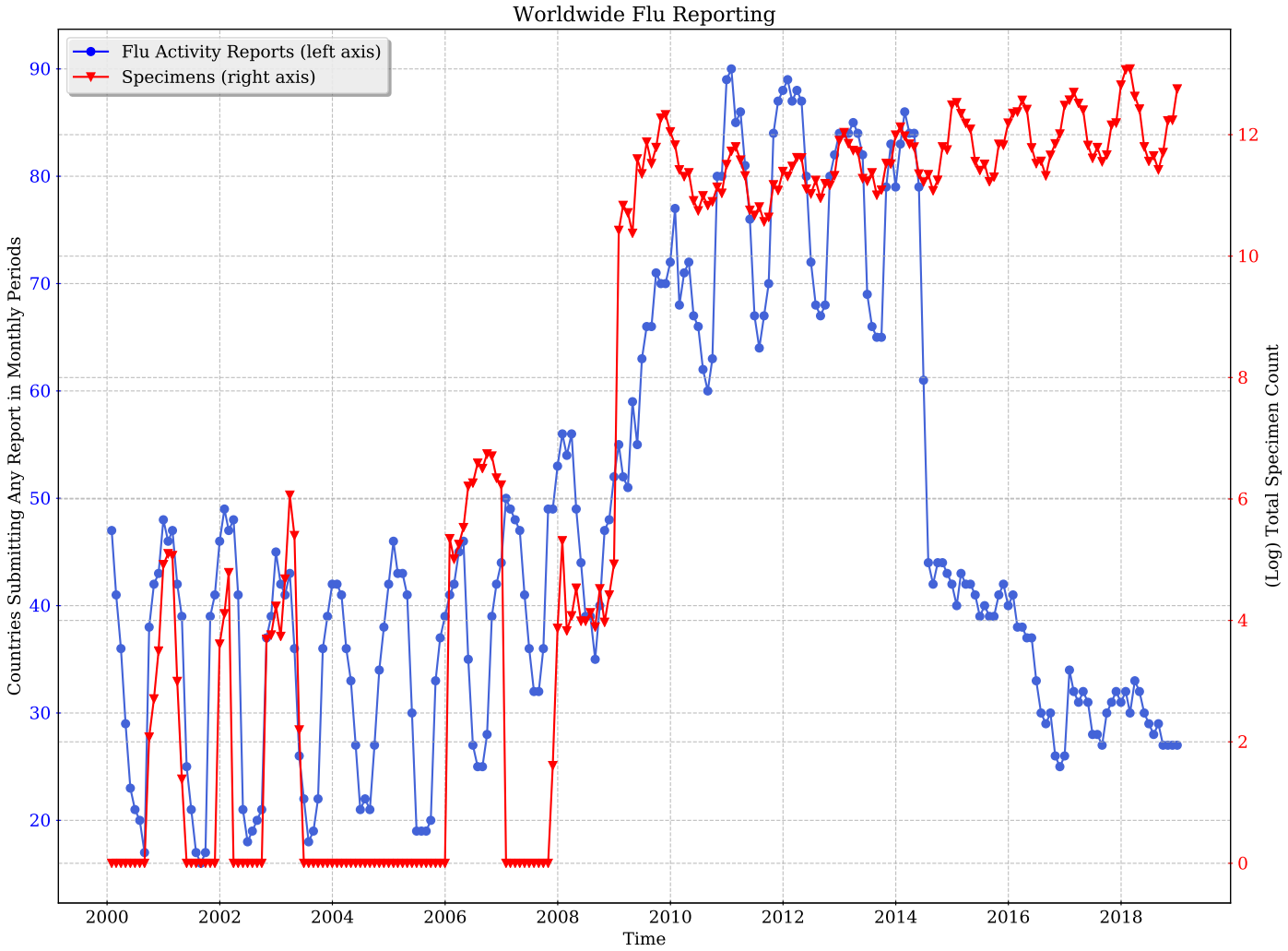


Figure 9: World Wide Flu Reporting

Flu reports are highly sporadic, and the rapid downward trend is unsettling.

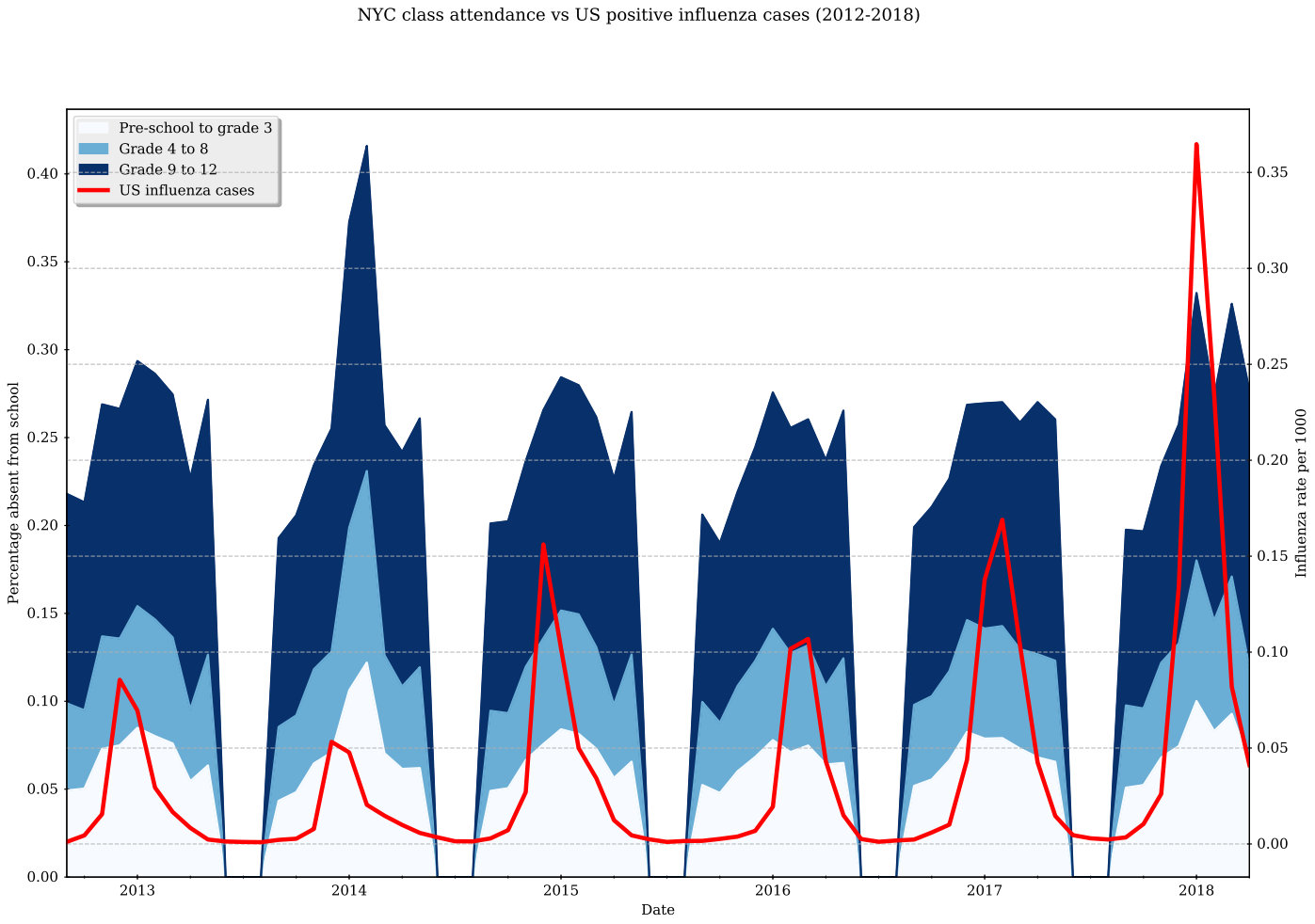


Figure 10: Potential Burden of Influenza on NYC Education
Flu outbreaks in the US (red) coincide with increased absentees from schools in New York City (blues). While causation conclusions cannot be effectively made, it seems that wealthy countries may be impacted by influenza in various ways (disease burden).