Thesis proposal

Pancasting: forecasting epidemics from provisional data

Logan Brooks
Computer Science Department
Carnegie Mellon University

Thesis Committee:
Roni Rosenfeld (chair)
Ryan Tibshirani
Zico Kolter
Jeffrey Shaman (Columbia University)

Abstract

Infectious diseases remain among the top contributors to human illness and death worldwide [71, 45]. While some infectious disease activity appears in consistent, regular patterns within a population, many diseases produce less predictable epidemic waves of illness. Uncertainty and surprises in the timing, intensity, and other characteristics of these epidemics stymies planning and response of public health officials, health care providers, and the general public. Accurate forecasts of this information with well-calibrated descriptions of the associated uncertainty can assist stakeholders in tailoring countermeasures, such as vaccination campaigns, staff scheduling, and resource allocation, to the situation at hand, which in turn could translate to reductions in the impact of a disease.

Domain-driven epidemiological models of disease prevalence can be difficult to fit to observed data while incorporating enough details and flexibility so that the observed data can be explained well. Meanwhile, more general statistical approaches can also be applied, but traditional modeling frameworks seem ill-suited for irregular bursts of disease activity, and focus on producing accurate single-number estimates of future observations rather than well-calibrated measures of uncertainty on more complicated functions of the data. The first part of the proposed work develops more flexible variants of simple statistical approaches that increase the flexibility of both point predictions and probability distribution forecasts.

Epidemiological surveillance systems commonly incorporate a data revision process, whereby each measurement may be updated multiple times to improve accuracy as additional reports and test results are received and data is cleaned. The second part of the proposed work discusses how this process impacts proper forecast evaluation and visualization. Additionally, it extends the models above to "backcast" how existing measurements will be revised, which in turn can be used to improve forecast accuracy.

Often, there are multiple available sources of estimates of a disease's prevalence, which vary in geographical and temporal scope and resolution, accuracy, and timeliness, and each of which may exhibit its own peculiarities. The final part of the proposed work further generalizes the above methodology to incorporate multiple data sources with similar temporal scopes and resolutions, in order to produce better forecasts than are possible with a single data source alone.

Contents

Chapter 1

Introduction

Much of this chapter is based on material from [8].

Infectious diseases and the motivation for forecasting

Despite modern medical advances, infectious diseases remain among the top causes of human illness and death worldwide, and pose major threats even in high-income countries [71, 45, 42]. Within the scope of infectious diseases, leading contributors include lower respiratory infections (e.g., with pneumonia or influenza) and diarrheal diseases (e.g., from foodborne bacteria and viruses) [71, 45, 42, 19]. Some infectious disease activity occurs in consistent, regular "endemic" patterns within a population, but many diseases produce less predictable "epidemic" waves of illness. Uncertainty and surprises in the timing, intensity, and other characteristics of these epidemics stymies planning and response of public health officials, health care providers, and the general public, and contributes to a high health and economic burden.

For instance, in the United States and other temperate regions, lower respiratory infection activity various classes of respiratory and circulatory disease, such as lower respiratory infections, present fairly uniform "baseline" patterns repeating each year, punctuated by sharp spikes in prevalence often associated with influenza epidemics [58, 65, 66, 77, 38]. Influenza epidemics typically occur once a year during the "influenza season" (roughly from October to May in the Northern Hemisphere), but vary in timing, intensity, and other traits; these "seasonal" epidemics are associated with an estimated 250 000 to 500 000 annual deaths worldwide [72], with a range of 3000 to 56 000 deaths in the US alone [12, 65, 57]. Additionally, influenza "pandemics", which are rare global outbreaks of especially novel influenza viruses [14, 13], can cause deaths on even greater scales [68, 33]. Potential countermeasures include [46] adjusting scheduling and providing on-site child care for health workers to better handle increased patient loads; canceling or rebooking less urgent medical appointments and procedures, admitting emergency department patients to inpatient hallway beds and reconfigured or alternative spaces; transferring patients to other facilities to avoid or reduce overcrowding; producing and tuning composition of vaccines; manufacturing, allocating, and redistributing antiviral medication, respirators, and other resources; and launching or modifying campaigns to promote vaccination, effective hand-washing practices, wearing face masks [55, 1, 64, 61, 63, 17], and other beneficial behaviors, targeted to sick individuals, their close contacts, or health workers, in order to curtail the spread and consequences of infections. The design and effectiveness of these efforts depends on the range of expectations for and ultimate reality of an epidemic's size, timing, and other characteristics.

Accurate and reliable forecasts of this information could provide early warning, bolster situational awareness, and assist in designing countermeasures, which in turn may reduce the overall impact of infectious disease. While the idea of epidemic modeling and forecasting is not new, recent years have seen growing interest driving government initiatives that standardize datasets, tasks, and metrics to improve forecast usability, address decision-maker needs, attract and assist external modelers, and allow for rigorous evaluation and comparison. These efforts include the U.S. government's Dengue Forecasting project, CHIKV (Chikungunya virus) Challenge, and a series of influenza forecast comparisons. This document will focus on these influenza forecasting testbeds and corresponding surveillance systems in the US.

The Centers for Disease Control and Prevention (CDC) monitors influenza prevalence with several well-established surveillance systems [11]; the recurring nature of seasonal epidemics and availability of historical data provide promising opportunities for the formation, evaluation, and application of statistical models. Starting with the 2013/2014 "Predict the Influenza Season Challenge" [4] and continuing each season thereafter as the Epidemic Prediction Initiative's FluSight project [5], CDC has solicited and compiled forecasts of influenza-like illness (ILI) prevalence from external research groups and worked with them to develop standardized forecast formats and quantitative evaluation metrics. Targets of interest include disease prevalence in the near future, as well as features describing the timing and overall intensity of the disease activity in the season currently underway. Policymakers desire not only in point predictions of these quantities, but full distributional forecasts; recent initiatives solicit both types of estimates, but base evaluation on customized log scores of distributional forecasts.

1.1 Models of disease dynamics

Various approaches to influenza epidemic forecasting are summarized in literature reviews [16, 47, 67] and descriptions of the CDC comparisons [4, 5]. Some common approaches are described below, with references to work applicable to the current FluSight project and related seasonal dengue forecasting tasks, emphasizing more recent work that may not be listed in the above three literature reviews:

Mechanistic models: describe the disease state and interaction between individuals with causal models, as well as the surveillance data generation process.

Compartmental models (e.g., [59, 60, 24, 76, 35])**:** break down the population into a number of discrete "compartments" describing their characteristics (e.g., age, location) and state (e.g., susceptible to, infectious with, or recovered from a particular disease), and describe how the occupancy of these compartments changes over time, either deterministically or probabilistically. In many of these models, this division describes solely the state with respect to a single disease, ignoring details regarding age, spatial dynamics, and mixtures of ILI diseases, but keeping the number of parameters to infer low. Methods to fit these models to data include variants of particle and ensemble Kalman filters [75], naïve importance sampling [7], iterative augmented-state filtering [28, 40, 29], general Bayesian frameworks [49] (using JAGS [53], Stan [10], etc.), filtering using linear noise approximation [79, 78], and Gaussian process approximations [9].

Agent-based models (e.g., [47, 18])**:** also known as individual-based models, these approaches use more detailed descriptions of disease state and/or individual characteristics and behavior, which are not easily simplified into a compartmental form, typically studied using computation-heavy simulations. These approaches usually include many more parameters than compartmental models, which may be set based on heuristics or additional data sources and studies, or, alternatively, inferred based on the surveillance data, often by using Markov chain Monte Carlo (MCMC) procedures. Developing effective inference techniques is an active area, with scalability to large populations still on the frontier of research, requiring special inference techniques and/or likelihood approximations [48].

Phenomenological models: also referred to as statistical models, these approaches describe the surveillance data without directly incorporating the epidemiological underpinnings.

Direct regression models (e.g., [69, 15, 7, 56])**:** attempt to estimate future prevalence or targets of interest using various types of regression, including nonparametric statistical approaches and alternatives from machine learning literature.

Time series models (e.g., [50, 27, 44, 41, 39, 74, 73, 32, 22])**:** represent the expected value of (transformations of) observations and/or underlying latent state at a particular time as (typically linear) functions of these quantities at previous times and additional covariates, paired with Gaussian, Poisson, negative binomial, or other noise distributions. This category includes linear dynamical systems and frameworks such as SARIMAX.

Complicated mechanistic approaches such as agent-based models are often too complex to efficiently fit to surveillance data and are instead less strictly “calibrated” based on summary measures, which may not produce a close match to the surveillance observations. Instead, mechanistic forecasting approaches have focused on simpler compartmental models and frameworks for fitting them to surveillance data. However, oversimplified compartmental models often cannot tightly match the surveillance data for an entire season simultaneously [7], which can degrade inference quality. Some degree of mismatch can be attributed to observation models that do not reflect important details of the surveillance system, which are discussed in the next section. Another contributor is the rigidity of compartmental models with deterministic state transitions and shallow-tailed observational noise, leading to overconfident forecasts; some paths forward are to incorporate variance inflation factors [59], overdispersed observational noise [41], stochastic variants or process noise [36], random walk discrepancy terms [49], error breeding procedures [52], more complex models with appropriate filtering algorithms [51], or use of improper conditioning procedures to combat overconfidence.

Phenomenological models, on the other hand, offer a wide range of general-purpose methods designed around efficient, straightforward inference. Univariate response models are extremely flexible, but seem inappropriate when the target of interest is a function of a surveillance time series. The most popular statistical time series methods, falling within “alphabet soup” frameworks such as SARIMAX and GARCH, directly model the time series, but sacrifice some flexibility by focusing on linear dynamics and Gaussian noise.

Chapter 2 expands the phenomenological front and moves toward the mechanistic one, presenting methods that incorporate the flexibility of univariate response models into ARI time series models, and ways to tailor these models to epidemiological settings to resemble a compartmental model.

1.2 Models of observations

Most epidemic modeling work, including much of the epidemic forecasting literature, tends to focus on the disease transmission dynamics, with the nature of surveillance system modeled with a very simple observational noise term. However, recognizing some details of surveillance systems is essential when performing retrospective forecast preparation and evaluation, and using these details to inform models can improve forecast accuracy. For example, surveillance data may contain spikes around holidays, which may be explained by differences in health care seeking behavior producing artifacts in the surveillance system, and/or due to changes in disease transmission behavior. Chapter 2 touches on some model settings that can be used to acknowledge holiday effects.

A more fundamental issue is that the ground truth from traditional surveillance systems used for evaluation is not available in real-time for use in forecasts. It takes time for symptoms to be recorded, diagnoses to be made, lab tests to complete; for health care workers to prepare and submit reports; and for public health officials to compile, clean, summarize, and publish the data. Furthermore, a case might only be reported after recovery or death, but recorded with a time closer to the onset of symptoms. In short, there is a trade-off between the accuracy of an observation and its timeliness. Traditional surveillance systems often address this problem by publishing an initial observation for a given time once the level of reliability is deemed acceptable, then later reporting a revised value or sequence of revised values that improve the expected accuracy. After some time, the observation may be finalized in the surveillance system or considered stable enough to be interchangeable with the finalized value and used as ground truth for forecast evaluation. Chapter 3 discusses how this revision process impacts proper retrospective forecast evaluation, and how forecast accuracy can be improved by modeling the revision process.

In recent years, a number of novel digital surveillance sources and derived estimators have been prepared using internet search query data, social media activity, web page hits, self-reported illness, self-reporting, internet-integrated monitoring and testing devices, electronic health records, insurance claims, or some combination along with traditional surveillance data. These estimates are not used as ground truth for evaluation, but may have better timeliness and resolution, and offer opportunities for improved forecasting. Some of these sources undergo a similar revision process as more traditional surveillance data; others may be available so quickly that the time period corresponding to a given observation has not ended before initial estimates are available. Chapter 4 discusses how to incorporate these additional data sources within the modeling framework presented in the previous chapters.

1.3 Epidemiological surveillance data

Epidemiological surveillance data exhibit a number of behaviors which are problematic for traditional time series methods without tailoring:

Rare or one-time events cause major shifts in reported disease prevalence, including

- **invasions:** introduction of a disease into an area that has not encountered it before;
- **novel strain pandemics:** epidemics with wider geographical spread or high incidence, often occurring at unseasonal times of year, caused by mutations in a strain of a disease that result in more effective transmission;
- **mass vaccination and eradication campaigns:** coordinated efforts by public health officials to drastically increase the proportion of the population that is vaccinated against a disease; and
- **sudden shifts in reporting practices or suitability:** changes in reporting requirements; the type or number of reporting health care providers (in a passive surveillance system); disease definitions, testing procedures, testing equipment, or testing sensitivity to prevalent disease strains; reporting frequency, geographical and temporal scope and resolution, disease specificity; among other changes;

Seasonality in transmissibility which results in irregular seasonal behavior in case counts: epidemic waves of varying heights and times that usually occur with some wide “on-season” time window (in addition to more predictable background seasonality for which sinusoidal or seasonal autoregressive terms and Gaussian-like noise seem more appropriate)

Nonadditive holiday effects on health care seeking, reporting, or disease transmission rates;

Data revisions to past surveillance data are common, as the reporting delay for cases may vary based on the attending health care provider and duration of illness, suspected cases of a disease may be included in early estimates but ruled out later, and, for rapidly available datasets, the time window for data aggregation may include times in the future (e.g., later days of the current week) which are necessarily not observed yet; and

Ragged data availability , used here to refer differences among surveillance signals in geotemporal and demographic resolution, availability, and reliability patterns; timeliness of release; and underlying stimuli, complicate the creation and use of models incorporating multiple signals simultaneously.

The proposed work focuses on building models that are appropriate given the last four aspects of epidemiological surveillance data. Chapter 2 focuses on building nonparametric univariate time series models that factor in holiday effects and seasonality in transmissibility, ignoring the fact that data revisions occur and additional surveillance signals may be available. Chapter 3 deals with the modeling of data revisions. Chapter 4 discusses incorporation of additional data sources with differing availability patterns.

Chapter 2

Probabilistic forecasting of the spread of epidemics

Stakeholders desire accurate and reliable forecasts of disease prevalence in the next few weeks, and of summary statistics about the timing and intensity of epidemics. The goal is to improve situational awareness and decision-making regarding, for example, hospital staffing and scheduling impacting readiness for surges in the number of inpatients, or the timing of a vaccination campaign. Each of the prediction targets could be handled separately: one model could be built to forecast disease prevalence next week, another to forecast the week when prevalence is highest, and so on. However, we focus on a more unified approach: first, forecasting the distribution of the disease prevalence trajectory for the entire season, then extracting the corresponding distributions for the targets of interest. This chapter discusses methods of forecasting the future of a trajectory given observed values of this trajectory in the past.

Given past observations $Y_{1..t}$ of a univariate surveillance time series $Y_{1..T}$ for a semi-regular seasonal epidemic, we want to estimate the distribution of future trajectories, $Y_{t+1..T}$. The distributional aspect of the forecast is important: many time series methods treat conditional mean estimates as "first class" and add Gaussian observational and/or process noise as a matter of convenience; we seek a more flexible noise model able of capturing heavy tails and multimodality. Furthermore, any conditional mean estimates that are produced should have a flexible, nonparametric flavor. Producing a sample from the distribution for $Y_{t+1..T}$ is sufficient; we do not need an explicit representation of the model.

One approach is to borrow from well-known, flexible univariate regression and density estimation models and repurpose them for time series estimation. A simple procedure allows us to sample from an estimate of $Y_{t+1..T} \mid Y_{1..t}$

based on samplers for estimates of one-step-ahead conditional distributions $Y_{t+1} \mid Y_{1..t}$, $Y_{t+2} \mid Y_{1..t+1}$, $\ldots$, $Y_T \mid Y_{1..T-1}$:

- Draw $Y_{t+1}^{\text{sim}} \sim Y_{t+1} \mid Y_{1..t}$
- Draw $Y_{t+2}^{\text{sim}} \sim Y_{t+2} \mid Y_{1..t}, Y_{t+1} = Y_{t+1}^{\text{sim}}$ (using model for $Y_{t+2} \mid Y_{1..t+1}$)
- Draw $Y_{t+3}^{\text{sim}} \sim Y_{t+3} \mid Y_{1..t}, Y_{t+1,t+2} = Y_{t+1,t+2}^{\text{sim}}$ (using model for $Y_{t+3} \mid Y_{1..t+2}$)
- $\ldots$
- Draw $Y_T^{\text{sim}} \sim Y_T \mid Y_{1..t}, Y_{t+1..t} = Y_{t+1..T-1}^{\text{sim}}$ (using model for $Y_T \mid Y_{1..T-1}$)

Record $Y_{t+1..T}^{\text{sim}}$ and repeat this process to obtain additional simulated futures. There are essentially no restrictions on the models selected for $Y_u \mid Y_{1..u-1}$ for each u.

One natural approach is to first directly estimate the conditional distribution $\Psi^{[u]} \mid \mathbf{\Phi}^{[u]}$, where $\Psi^{[u]}$ is a (potentially u-specific) function of $Y_{1..u}$ from which Y_u can be recovered given $Y_{1..u-1}$ (e.g., $\Psi^{[u]} = \Delta Y_u = Y_u - Y_{u-1}$ or $\Psi^{[u]} = \log Y_u$), and $\mathbf{\Phi}^{[u]}$ is a (potentially u-specific) vector of features derived from $Y_{1..u-1}$. During simulation, $Y_{1..u-1}^{\text{sim}}$ will be used to calculate corresponding simulated feature values $\mathbf{\Phi}^{[u],\text{sim}}$, which are used to draw a simulated transformed value $\Psi^{[u],\text{sim}}$, from which a corresponding simulated value Y_u^{sim} can be recovered. Nonparametric methods along these lines include:

Kernel delta density, which draws ΔY_u^{sim} from an estimate of the conditional density for $\Delta Y_u \mid \mathbf{\Phi}^{[\text{KDD},u]}$ based on smoothing kernel methods with some heuristic modifications, where $\mathbf{\Phi}^{[\text{KDD},u]}$ is a vector of heuristically constructed and weighted features for time u derived from $Y_{1..u-1}$, and

Quantile autoregression using locally linear quantile regression and corrupting noise, which selects $\Psi^{[u],\text{sim}}$ as the sum of a random estimated conditional quantile and (optionally) some smoothing noise, where the conditional quantile is estimated for $\Psi^{[u]} \mid \mathbf{\Phi}^{[\text{QARlinear},u]}, \mathbf{\Phi}^{[\text{QARkernel},u]}$ as a linear function of covariates $\mathbf{\Phi}^{[\text{QARlinear},u]}$, with training data weighted with a smoothing kernel on covariates $\mathbf{\Phi}^{[\text{QARkernel},u]}$.

2.1 Kernel delta density

Content in this section is taken from or based on material from [8].

Kernel density estimation and kernel regression use smoothing kernels to produce flexible estimates of the density of a random variable (e.g., $f_{Y_{t+1..T}}$) and the conditional expectation of one random variable given the value of another (e.g., $\mathbb{E}[Y_{t+1..T} \mid Y_{1..t}]$), respectively; we can combine these two methods to obtain estimates of the conditional density of one random variable given another. One possible approach would be to use the straightforward estimate

$$\hat{f}_{Y_{t+1..T}|Y_{1..t}}(y_{t+1..T} \mid y_{1..t}) = \frac{\sum_{s=1}^{S} I^{[1..t]}(y_{1..t}, Y_{(1..t)+(\Delta t)_s}) O^{[t+1..T]}(y_{t+1..T}, Y_{(t+1..T)+(\Delta t)_s})}{\sum_{s=1}^{S} I^{[1..t]}(y_{1..t}, Y_{(1..t)+(\Delta t)_s})},$$

where $\{1..S\}$ is the set of fully observed historical training seasons, and $I^{[1..t]}$ and $O^{[t+1..T]}$ are smoothing kernels describing similarity between "input" trajectories and between "output" trajectories, respectively. However, while basic kernel smoothing methods can excel in low-dimensional settings, their performance scales very poorly with growing dimensionality. During most of the season, neither $Y_{1..t}$ nor $Y_{t+1..T}$ is low-dimensional, and the current season's observations are extremely unlikely to closely match any past $Y_{(1..t)+(\Delta t)_s}$ or $Y_{(t+1..T)+(\Delta t)_s}$. This, in turn, can lead to kernel density estimates for $Y_{t+1..T}$ based almost entirely on the single season s with the closest $Y_{(1..t)+(\Delta t)_s}$ when conditioning on $Y_{1..t}$, and excessively narrow density estimates for $Y_{t+1..T}$ even without conditioning on $Y_{1..t}$. The high-dimensional output issue is already resolved by chaining together estimates of conditional densities with univariate outputs: $f_{\Delta Y_u|Y_{1..u-1}}$ for each u from $t+1$ to T, where $\Delta Y_u = Y_u - Y_{u-1}$. Estimating single-dimensional densities requires relatively little data. However, this reformulation exacerbates the high-dimensional input problem since we are conditioning on $Y_{1..u-1}$, which can be considerably longer than $Y_{1..t}$. We address the high-dimensional input problem by approximating $f_{\Delta Y_u|Y_{1..u-1}}$ with $f_{\Delta Y_u|\mathbf{\Phi}^{[\text{KDD},u]}}$ where $\mathbf{\Phi}^{[\text{KDD},u]}$ is some low-dimensional vector of features derived from $Y_{1..u-1}$. The straightforward conditional density estimation method described above for $Y_{t+1..T} \mid Y_{1..t}$ can be applied to the chained distributions $\Delta Y_u \mid \mathbf{\Phi}^{[\text{KDD},u]}$, although literature indicates that this approach is suboptimal [23].

The conditional density estimates above were developed based on combining kernel regression and univariate kernel density estimation techniques, it can also be understood as sampling from a joint kernel density estimate over input and output variables using a product kernel. A slightly more complicated take on the former viewpoint has been found to yield faster theoretical and simulated statistical convergence rates [23]. The latter interpretation offers additional alternatives such as deriving results from a joint density estimate based on a kernel that is not the product of an input and output kernel, as well as copula techniques. These approaches have been incorporated in a separate epidemiological forecasting system working directly with the higher-dimensional inputs and outputs rather than the one-step-ahead approach [56]. A host of work on kernel conditional density estimation offers avenues to improving these kernel delta density approaches, as well as resolving the original issues regarding high dimensionality.

We use two sets of choices for the approximate conditional density function and summary features to form two versions of the method.

Markovian delta density: approximates the conditional density of ΔY_u given $Y_{1..u-1}$ with its conditional density given just the previous (real or simulated) observation, Y_u:

$$\begin{aligned}\hat{f}_{Y_{t+1..T}|Y_{1..t}}(y_{t+1..T} \mid y_{1..t}) &= \prod_{u=t+1}^{T_2} \hat{f}_{\Delta Y_u|Y_{1..u-1}}(\Delta y_u \mid y_{1..u-1}) \\ &= \prod_{u=t+1}^{T_2} \hat{f}_{\Delta Y_u|Y_{u-1}}(\Delta y_u \mid y_{u-1}) \\ &= \prod_{u=t+1}^{T_2} \frac{\sum_s I^{[u]}(y_{u-1}, Y_{u-1+(\Delta t)_s}) \cdot O^{[u]}(\Delta y_u, \Delta Y_{u+(\Delta t)_s})}{\sum_s I^{[u]}(y_{u-1}, Y_{u-1+(\Delta t)_s})},\end{aligned}$$

where $I^{[u]}$ and $O^{[u]}$ are Gaussian smoothing kernels. The first equality corresponds to the chain rule of probability on the actual (not estimated) densities; the second incorporates the Markov assumption (i.e., selects $\mathbf{\Phi}^{[u]} = [Y_{u-1}]$); and the third gives our choice of estimators for the conditional densities $\hat{f}_{\Delta Y_u|Y_{u-1}}$ for each u. The bandwidth of each $I^{[u]}$ and $O^{[u]}$ is chosen separately using bandwidth selection procedures for regular kernel density estimation of Y_{u-1} and ΔY_u, respectively. (Specifically, we use the `bw.SJ` function from the `R`[54] built-in `stats` package, with `bw.nrd0` as a fallback in the case of errors. These functions do not accept weights for the inputs; it may be possible to improve forecast performance by incorporating these weights or by using other approaches to select the bandwidths.) Note that density estimates for ΔY_u are based on data from past seasons on week u only, allowing the method to incorporate seasonality and holiday effects (for holidays that consistently occur at the same time of year).

Forecasts are based on Monte Carlo simulations of $Y_{t+1..T} \mid Y_{1..t}$ using the chained one-step-ahead procedure described in the previous section. This process is illustrated in Figure 2.1. Repeating this procedure many times yields a sample from the model for $Y_{t+1..T} \mid Y_{1..t}$; stopping at 2000 draws seems sufficient for use in our ensemble forecasts, while at least 7000 are needed to smooth out noise when displaying distributional target forecasts for the delta density method in isolation. Any negative simulated weighted %ILI (wILI) values in these trajectories are clipped off and replaced with zeroes.

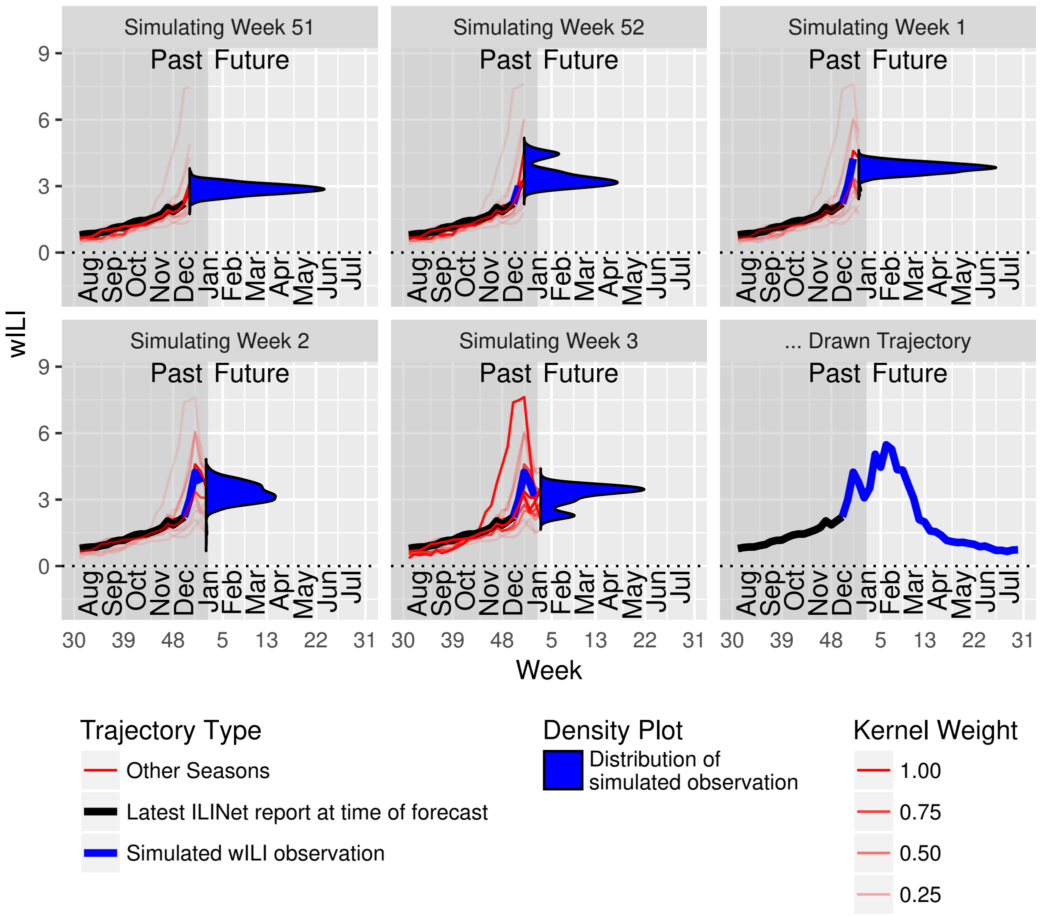

Figure 2.1: **The delta density method conditions on real and simulated observations up to week $u-1$ when building a probability distribution over the observation at week u.** This figure demonstrates the process for drawing a single trajectory from the Markovian delta density estimate. The past data $Y_{1..t}$, which incorporates observations through week 48, is shown in black. Kernel smoothing estimates for future values at times u from $t+1$ to T_2 are shown in blue, as are simulated observations drawn from these estimates. Past seasons' trajectories are shown in red, with alpha values proportional to the weight they are assigned by the kernel I^u.

Extended delta density: approximates the conditional density of ΔY_u given $Y_{1..u-1}$ with its conditional density given four features:

- the previous wILI value, Y_{u-1};
- the sum of the previous k^u wILI values, roughly corresponding to the sum of wILI values for the current season;
- an exponentially weighted sum of the previous k^u wILI; values, where the weight assigned to time u' is $0.5^{t'-u'}$; and
- the previous change in wILI value, ΔY_{u-1}.

The approximate conditional density assigns each of these features a weight (0.5, 0.25, 0.25, and 0.5, respectively) in order to reduce overfitting and emphasize some relative to the others, and incorporates data from other weeks close to u (specifically, within l^u weeks; the choice of l^u is discussed in a later section) with a truncated Laplacian kernel. We selected these weights and other settings, such as kernel bandwidth selection rules, somewhat arbitrarily based on intuition and experimentation on out-of-sample data; a cross-validation subroutine could be used to make the selection as well, but would multiply the amount of computation required. In case the resulting product of Gaussian and Laplacian kernels is too narrow, we mix its results with a wide boxcar kernel which evenly weights all data from time $u - l^u$ to $u + l^u$:

$$\begin{aligned}
&\hat{f}_{\Delta Y_u | Y_{1..u-1}}(\Delta y_u \mid y_{1..u-1}) \\
&= 0.9 \cdot \frac{\sum_s \sum_{u'=u-l^u}^{u+l^u} 0.7^{|u'-u|} \left[I_1^u(y_{u-1}, Y_{(u'-1)+(\Delta t)_s})\right]^{0.5} \cdots O^u(\Delta y_u, \Delta Y_{(u')+(\Delta t)_s})}{\sum_s \sum_{u'=u-l^u}^{l^u} 0.7^{|u'-u|} \left[I_1^u(y_{u-1}, Y_{(u'-1)+(\Delta t)_s})\right]^{0.5} \cdots \left[I_4^u(\Delta y_{u-1}, \Delta Y_{(u'-1)+(\Delta t)_s})\right]^{0.5}} \\
&+ 0.1 \cdot \frac{\sum_s \sum_{u'=u-l^u}^{u+l^u} O^u(\Delta y_u, \Delta Y_{(u')+(\Delta t)_s})}{\sum_s \sum_{u'=u-l^u}^{u+l^u} 1}.
\end{aligned}$$

Using data from $u' \neq u$ incorporates additional reasonable outcomes for Δy_u by incorporating past wILI patterns with different timing, but risks including some very unreasonable possibilities produced by repeatedly drawing from the same u' rather than following seasonal trends with increasing u''s. For example, when a portion of a past season that is more similar to itself with a slight time shift than to any other past season, it may be selected for multiple consecutive u's and produce an unreasonable trajectory. This could potentially occur when drawing data from the relatively flat regions of wILI trajectories of many seasons, or when incorporating observations around an unusually early, late, high, or low peak. To prevent this possibility, we combine the natural estimate for Y_u arising from the density estimate for ΔY_u with a random draw $Y_{\text{uncond}u}$

from the unconditional density estimate for Y_u (using a Gaussian kernel and only data from week u):

$$Y_u^{\text{sim}} = 0.9 \cdot (Y_{u-1} + \Delta Y_u^{\text{sim}}) + 0.1 \cdot Y_u^{\text{uncond}}.$$

2.2 Quantile autoregression

Locally linear quantile regression offers an alternative approach to modeling $Y_u \mid Y_{1..u-1}$ offering greater flexibility in covariate relationships and better anticipated behavior with larger numbers of covariates; "corrupting" its output with random noise is one way to address potential issues with discrete outputs that do not cover the entire support of Y_u. Basic linear quantile regression estimates the τth conditional quantile of some variable Y given covariates X as a linear function of X; locally linear quantile regression additionally allows for weighting of training instances based on a smoothing kernel on another set of covariates X' (potentially overlapping with X). Additionally, the same types of transformations can be applied on the output and covariates as in the kernel smoothing case. A specification of a simple corrupted locally linear quantile autoregression approach could consist of:

- $\Psi^{[u]}$: a transformation of Y_u from which we can recover Y_u (potentially using information from $Y_{1..u-1}$),
- $\mathbf{\Phi}^{[\text{QARlinear},u]}$: a set of features (derived from $Y_{1..u-1}$) to use in the linear combination estimating some quantile of Y_u,
- $\mathbf{\Phi}^{[\text{QARkernel},u]}, K^{\mathbf{\Phi}^{[\text{QARkernel},u]}}$: a set of features (derived from $Y_{1..u-1}$) and corresponding smoothing kernel (or "weighted" smoothing kernel as used in extended delta density) that assigns weights to training instances, and
- $K^{\Psi^{[u]}}$, a smoothing kernel that defines the distribution of additive corrupting noise.

The corresponding sampling procedure for Y_u^{sim} is:

1. Draw quantile level $\tau \sim U[0,1]$.
2. Compute estimate $\hat{q}$ of the level τ quantile of $\Psi^{[u]} \mid \Phi^{[\text{QARlinear},u]}, \Phi^{[\text{QARkernel},u]}$ using locally linear quantile regression.
3. Draw $\epsilon \sim K^{\Psi^{[u]}}$ from corrupting noise distribution.
4. Let $\Psi^{[u],\text{sim}} = \hat{q} + \epsilon$.

5. Let Y_u^{sim} be the value of Y_u given by $\Psi^{[u]} = \Psi^{[u],\text{sim}}$ and $Y_{1..u-1}$.

Quantile autoregression has already been formulated and studied from a theoretical perspective and applied to economic datasets [37]. A recent application to flu forecasting [70] studied different data weighting approaches based on time of season. Similarly, the proposed work focuses on customizing quantile autoregression approaches to epidemiological data.

2.2.1 Connection to smoothing kernel approaches

The family of corrupted locally linear quantile autoregression approaches subsumes the considered delta density approaches after mirroring any heuristic modifications to the kernel conditional density estimates. Consider a kernel conditional density estimate of $\Delta Y_u \mid \mathbf{\Phi}^{[\text{KDD},u]}$ using covariate kernel $K^{\mathbf{\Phi}^{[\text{KDD},u]}}$. If the response kernel $K^{\Delta Y_u}$ is replaced with the degenerate Dirac delta distribution, the resulting kernel conditional "density" estimates are just weighted empirical distributions. The corresponding quantiles are weighted sample quantiles of ΔY_u with weights based on $K^{\mathbf{\Phi}^{[\text{KDD},u]}}$; this coincides with the estimated quantiles of the locally linear/constant quantile regression model with the same $\Psi^{[u]}$, $\mathbf{\Phi}^{[\text{QARlinear},u]} = (1)$ (the model only fits an "intercept"), $\mathbf{\Phi}^{[\text{QARkernel},u]} = \mathbf{\Phi}^{[\text{KDD},u]}$, and $K^{\mathbf{\Phi}^{[\text{QARkernel},u]}} = K^{\mathbf{\Phi}^{[\text{KDD},u]}}$. The sampling procedures also coincide: drawing from a weighted empirical distribution function gives an equivalent distribution to selecting a weighted sample quantile with a level randomly distributed on the unit interval. ("Sample quantile" here is restricted to quantiles of the type outputted by quantile regression; for a finite number of quantile levels, there will not be a unique associated sample quantile and the one selected may vary across implementations, but these levels are drawn with probability 0. For other types of quantiles, e.g., from continuous quantile functions [26], this is normally not the case.) Using $K^{\Delta Y_u}$ instead of the Dirac distribution is equivalent to just adding additional noise to a draw from the weighted empirical distribution; thus, the smoothing kernel approach can be completely mimicked by a corrupted locally linear quantile regression approach using the same $K^{\Delta Y_u}$ as the corrupting noise distribution.

2.2.2 Incorporating covariates inspired by mechanistic models

While quantile regression can be restricted and post-processed to match the output of the kernel conditional density method, it is natural to favor use of

$\mathbf{\Phi}^{[\text{QARlinear},u]}$ covariates not only in appeal to more general statistical arguments regarding scaling with higher dimensionality inputs and boundary bias, but also due to similarities with domain-driven mechanistic models when incorporating autoregressive terms. Furthermore, additional covariates can be constructed to strengthen this resemblance while maintaining the flexibility of quantile modeling and smoothing kernel weighting.

Epidemiological compartmental models are a popular class of mechanistic model that divides a population into a fixed number of "compartments" and considers all individuals within each compartment to behave identically. System dynamics are characterized by the manner in which individuals are added, removed, or flow between different compartments. For example, "SIRS" compartmental models represent population state by the number or proportion of individuals in each of three states: those

- Susceptible to infection with some disease,
- Infectious and spreading the disease, and
- Recovered from the infectious stage of a disease and currently immune to future reinfection;

Susceptible individuals can become Infectious by interacting with Infectious individuals, Infectious individuals transition to Recovered over time, and Recovered individuals can become Susceptible again due to waning immunity or mismatches of antibodies with currently circulating strains of a pathogen; these possible transitions are the basis for the initialism "SIRS". A simple deterministic, continuous-time, proportion-based SIRS model can be specified with the following system of differential equations:

$$\begin{aligned}
s'(t) &= -s(t)\cdot\beta i(t) + r(t)\cdot\mu \\
i'(t) &= +s(t)\cdot\beta i(t) - i(t)\cdot\gamma \\
r'(t) &= +i(t)\cdot\gamma - r(t)\cdot\mu \\
s(0) &+ i(0) + r(0) = 1, s(0) \geq 0, i(0) \geq 0, r(0) \geq 0,
\end{aligned}$$

where

- $s(t)$, $i(t)$, and $r(t)$ are the proportions of the population in the Susceptible, Infectious, and Recovered states, respectively, at time t;
- β is the rate at which any individual experiences contact with another person in which the latter could potentially spread an infection to the former (assumed to be the same across all pairs of individuals, regardless

of their current state), potentially modulated by the current weather (i.e., $\beta(\mathbf{w})$ where $\mathbf{w}$ is a vector of weather variables) or other data;

- μ is the rate at which recovered individuals become susceptible again;
- γ is the rate at which infectious individuals recover; and
- the conditions on the state at $t = 0$ are preserved as invariants for all other t.

The underlying proportions $s(t)$, $i(t)$, and $r(t)$ are latent; a simple noiseless observation model assumes that infectious individuals produce some kind of reported health care events at a steady rate, with no false positives from the other compartments:

$$y(t) = i(t) \cdot N\rho,$$

where

- $y(t)$ is the number of reported health care events at time t,
- N is the population size, and
- ρ is the rate at which infectious individuals generate reported health care events.

Already, this formulation suggests the use of models with linear autoregressive terms, as changes to compartment occupancy depend linearly or quadratically on the current occupancy, and the observations depend linearly on compartment occupancy. However, the latent dynamics and quadratic terms complicate the relationship; fortunately, a few manipulations will allow us to fully characterize the dynamics of $y(t)$ without any reference to latent state, revealing a very direct relationship with linear autoregressive and additional auxiliary terms. These manipulations are likely more widely familiar in the context of differential equations than discrete-time difference equations, so we examine the former first then establish parallels in the latter.

Our ultimate goal is to express $y'(t)$ as a causal function of $y(t)$ (i.e., a function depending only on $y(\tau)$ for $\tau \leq t$). First, note that

$$\begin{aligned} y'(t) &= i'(t) \cdot N\rho \text{ and} && \text{(derivatives are linear)} \\ i(t) &= \frac{1}{N\rho} y(t) && \text{(scale both sides of } y(t) \text{ definition)} \end{aligned}$$

so we can instead seek to express $i'(t)$ as a causal function of $i(t)$ and quickly obtain $y'(t)$ as a causal function of $y(t)$. Next, observe that

$$\begin{aligned} i'(t) &= \beta s(t)i(t) - \gamma i(t) \\ &= \beta[1 - i(t) - r(t)]i(t) - \gamma i(t), \qquad \text{(proportions sum to 1)} \end{aligned}$$

so we just need to express $r(t)$ as a causal function of $i(t)$. Rearranging the equation for $r'(t)$ and applying an integrating factor approach, we find that

$$\begin{aligned} r'(t) &= i(t) \cdot \gamma - r(t) \cdot \mu \\ \mu r(t) + r'(t) &= \gamma i(t) \\ \mu e^{\mu t} r(t) + e^{\mu t} r'(t) &= \gamma e^{\mu t} i(t) \\ e^{\mu t} r(t) &= \gamma \int_{t_0}^{t} e^{\mu \tau} i(\tau)\,\mathrm{d}\tau + C \\ r(t) &= \gamma \int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau)\,\mathrm{d}\tau + Ce^{-\mu t}, \end{aligned}$$

for

- a time t_0 which is arbitrary for this derivation, but which we must select to be in the range of times for which observations are available, to ensure the integral involves only observed values of its argument, and
- a constant of integration $C \geq 0$ determining the initial conditions;

thus, $r(t)$ can be represented as a scaled exponential moving average of $i(t)$ (a causal function of $i(t)$) plus an exponential decay term. Applying the earlier observations gives

$$\begin{aligned} i'(t) &= \beta[1 - i(t) - r(t)]i(t) - \gamma i(t) \\ &= \beta[1 - i(t) - \gamma \int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau)\,\mathrm{d}\tau - Ce^{-\mu t}]i(t) - \gamma i(t) \\ &= (\beta - \gamma)\left[i(t)\right] - \beta\left[i^2(t)\right] - \beta\gamma\left[\int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau)\,\mathrm{d}\tau \cdot i(t)\right] - \beta C\left[e^{-\mu t} i(t)\right] \end{aligned}$$

and

$$\begin{aligned} y'(t) &= i'(t) \cdot N\rho \\ &= N\rho(\beta - \gamma)\left[i(t)\right] - N\rho\beta\left[i^2(t)\right] - N\rho\beta\gamma\left[\int_{t_0}^{t} e^{-\mu(t-\tau)} i(\tau)\,\mathrm{d}\tau \cdot i(t)\right] - N\rho\beta C\left[e^{-\mu t} i(t)\right] \\ &= (\beta - \gamma)[y(t)] - \frac{\beta}{N\rho}\left[y^2(t)\right] - \frac{\beta\gamma}{N\rho}\left[\int_{t_0}^{t} e^{-\mu(t-\tau)} y(\tau)\,\mathrm{d}\tau \cdot y(t)\right] - \beta C\left[e^{-\mu t} y(t)\right]. \end{aligned}$$

The discrete-time analogues of the key equations above and some additional transformations follow:

$$
\begin{aligned}
s_{t+1} &= s_t - \beta s_t i_t + \mu r_t \\
i_{t+1} &= i_t + \beta s_t i_t - \gamma i_t \\
r_{t+1} &= r_t + \gamma i_t - \mu r_t \\
s_0 + i_0 + r_0 &= 1, s_0 \geq 0, i_0 \geq 0, r_0 \geq 0 \\
y_t &= N\rho i_t \\
\Delta y_{t+1} = y_{t+1} - y_t &= (\beta - \gamma)\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau \cdot y_t\right] - \beta C\left[(1-\mu)^{t-1} y_t\right] \\
y_{t+1} &= (1 + \beta - \gamma)\left[y_t\right] - \frac{\beta}{N\rho}\left[y_t^2\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau \cdot y_t\right] - \beta C\left[(1-\mu)^{t-1} y_t\right] \\
\frac{\Delta y_{t+1}}{y_t} &= (\beta - \gamma)\left[1\right] - \frac{\beta}{N\rho}\left[y_t\right] - \frac{\beta\gamma}{N\rho}\left[\sum_{t_0}^{t-1}(1-\mu)^{t-1-\tau} y_\tau\right] - \beta C\left[(1-\mu)^{t-1}\right].
\end{aligned}
$$

The last few equations motivate the use of the bracketed quantities on the right as covariates in a regression for the response variable given on the left. Unfortunately, the last two bracketed quantities have a nonlinear dependence on the parameter μ and so μ can not be immediately selected using linear (quantile) regression; instead, a value of μ can be selected from domain literature or with hyperparameter search, or, for additional flexibility, multiple versions of the bracketed quantities with different possible μ values can be included simultaneously in the same regression. Additional manipulations of equations might allow for more efficient estimation of μ; alternatively, they might reveal some nonidentifiability of μ showing that any arbitrary selection of μ within some wide constraints would be equally valid, eliminating this concern altogether. If such nonidentifiability does not hold for the current model, it might hold for a version of the model incorporating birth and death rates and/or false positive reporting rates; repeating or generalizing the above analysis to this case and other types of compartmental models would be of interest regardless.

The primary goal of this effort is to inform construction of a higher-quality, easily fit model for y_t; any interpretation regarding the latent state is suspect in such a simplistic model, particularly causal or counterfactual reasoning, and especially if some parameters are nonidentifiable. Still, it is notable that we can recover (estimates of) γ, β, $N\rho$, and the latent state, at least in this purely deterministic setup; for example, consider the formulation involving $\frac{\Delta y_{t+1}}{y_t}$:

$$
\begin{aligned}
\theta_1 &= \beta - \gamma \\
\theta_2 &= -\frac{\beta}{N\rho} \\
\theta_3 &= -\frac{\beta\gamma}{N\rho} \\
\theta_4 &= -\beta C \\
\gamma &= \frac{\theta_3}{\theta_2} && = \frac{-\beta\gamma/(N\rho)}{-\beta/(N\rho)} && \text{(using definitions of } \theta_2,\ \theta_3\text{)} \\
\beta &= \theta_1 + \frac{\theta_3}{\theta_2} && = \theta_1 + \gamma && \text{(using definition of } \theta_1\text{)} \\
N\rho &= -\frac{\theta_1}{\theta_2} - \frac{\theta_3}{\theta_2^2} && = -\frac{\beta}{\theta_2} && \text{(using definition of } \theta_2\text{)} \\
C &= -\frac{\theta_4}{\theta_1 + \frac{\theta_3}{\theta_2}} && = -\frac{\theta_4}{\beta} && \text{(using definition of } \theta_4\text{)} \\
i_t &= \frac{y_t}{N\rho} \\
r_t &= \gamma \sum_{t_0}^{t-1} (1-\mu)^{t-1-\tau} i_\tau + C(1-\mu)^{t-1} && && \text{(parallel of continuous-time result)} \\
s_t &= 1 - i_t - r_t.
\end{aligned}
$$

The derivations above present some exciting possibilities for fitting compartmental models using standard regression routines, which may scale more readily than particle filter and MCMC approaches. While this derivation is based on a deterministic model, quantile autoregression provides for a flexible noise model which acts like process noise on i_t; other regression methods such as linear regression and generalized linear models provide additional options. However, observational noise in y_t and process noise in s_t and r_t is not considered; the former is especially important when dealing with noisy signals so momentary fluctuations are not mistaken for trends. This suggests a few potential branches of further investigation. First, preparing retrospective forecasts for real and/or simulated data using the currently bracketed covariates with various choices of μ could suggest whether forecast quality could benefit from a more thorough investigation. Second, a class of models or fitting technique that resolves the μ estimation problem would eliminate some potential reliability issues and may speed up computation. Third, the lack of true observational noise is somewhat jarring; spectral methods for predictive linear dynamical models (e.g., hidden Markov models (HMMs) [25], kernelized HMMs [62], and linear dynamical systems (Kalman filtering) [6]) may provide a guide to incorporating multivariate observational and process noise into the above derivation, and tools from quantile filtering [31] for maintaining a nonparametric noise model; alternatively, additional features and regularization may be sufficient to reduce sensitivity to observational noise encountered. Finally, the primary appeal of these manipulations is the potential to enable better scalability of nonparametric autoregressive models or other approaches when including multiple locations, demographic groups, or virus types, or to incorporate weather covariates, preferably using a generalized derivation and computational framework. One part of the proposed work is to perform the initial evaluation step above and investigate some of the latter points if merited.

2.2.3 Incorporating covariates to model multiplicative holiday effects

Content in this subsection is taken from or based on material from [8].

Holidays can impact the spread, observation, and impact of a disease. For example, reduced school and workplace contact may reduce disease transmission, patients may not seek or may delay medical care for less serious issues, and some health care providers may not be open or operate with reduced staffing. The delta density methods described above attempt to match holiday behaviors by restricting training windows around major holidays to focus on data from the same, or nearby, weeks of the year. This reduction in the amount of training data might actually degrade performance. A more direct model of the holiday effects may allow a model to match holiday behavior with less data, and simultaneously remove the perceived need for narrow training windows.

CDC's wILI measure is an estimate of the proportion of health care visits in an area that are due to ILI. Sharp rises and drops in wILI are common from early or mid-December to early January (roughly coinciding with a four week period beginning with epi week 50), with either the season's peak or a lower, secondary peak commonly occurring on epi week 52. This pattern appears to arise from at least two factors:

- spikes downward in the number of non-ILI visits during the holiday season (corresponding to increases in wILI), perhaps caused by patients choosing not to visit the doctor for less serious issues on holidays, and

- decreases in the average number of ILI visits at the end of the holidays, perhaps due to decreased transmission of ILI during holidays, which make the preceding increases in wILI appear even sharper.

Similarly, there are spikes or minor blips downward in the average number of non-ILI visits (which can result in small increases in wILI) associated with Thanksgiving Day; Labor Day; Independence Day; Memorial Day; Birthday of Martin Luther King, Jr.; Washington's Birthday; Columbus Day; and perhaps other holidays. The spike upward in wILI at Thanksgiving can push wILI unexpectedly over the onset threshold, and holiday effects may help explain the surprising frequency at which peaks occur on epi week 7 but not neighboring weeks. Additional age-specific patterns may be obscured by this analysis of aggregate ILI and non-ILI visit counts.

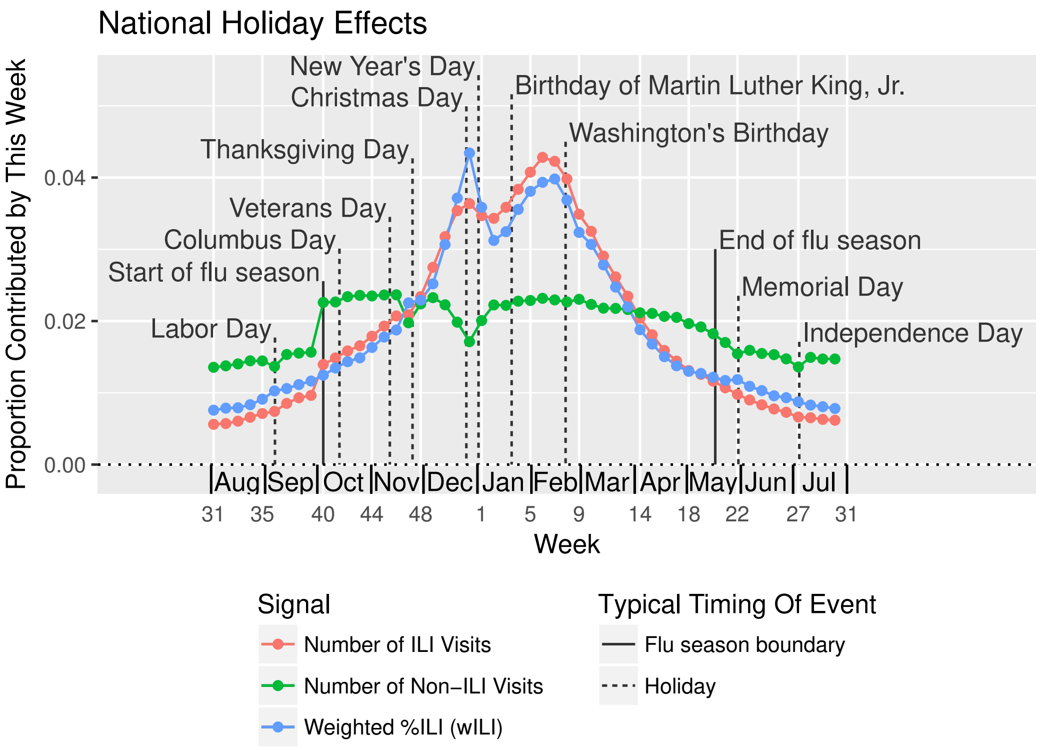

Figure 2.2: **On average, wILI is higher on holidays than expected based on neighboring weeks.** Weekly trends in wILI values, as expressed by the contribution of a each week to a sum of wILI values from seasons 2003/2004 to 2015/2016, excluding 2008/2009 and 2009/2010 (which include portions of the 2009 influenza pandemic), show spikes and bumps upward on and around major holidays. (U.S. federal holidays are indicated with event lines.) The number of non-ILI visits to ILINet health care providers spikes downwards on holidays (disproportionately with any drops in the number of ILI visits), contributing to higher wILI. The number of ILI visits generally declines in the second half of the winter holiday season, causing winter holiday peaks to appear even higher relative to nearby weeks. In addition to holiday effects, we see that average ILINet participation jumps upward on epi week 40, and gradually tapers off later in the season and in the off-season.

The goal of future exploration would be to incorporate holiday indicator covariates, lags of these covariates, or other features into the quantile autoregression approach, alongside response transformations such as $\Phi^{[u]} = \frac{\Delta Y_{u+1}}{Y_u}$, to obtain exact or approximate multipliers on reporting and transmission rates.

Chapter 3

Modeling surveillance data revisions

The past discussion assumed that, when forecasting future measurements of disease prevalence, we have access to these same desired measurements for all times in the past. In reality, this exact data is not immediately accessible, as accurate measurements may take weeks or years to be completed. However, to enable decisionmakers to quickly assess and respond to a situation, epidemiological surveillance systems often publish a sequence of tentative versions of each complete measurement, with later versions more accurate on average. The existence of multiple versions of measurements has significant implications for proper forecast evaluation and analysis, and explicitly accounting for the revision process can improve model forecasts:

- When estimating the performance of a proposed model by mimicking the forecasts it would have made in the past, it is important that we input the version of each measurement that would have been available at the time of each forecast; otherwise, accuracy estimates will almost surely be too high since the evaluation was based on higher accuracy input data.

- Visualizing past forecasts together with completed measurements can cause confusion when the version of the measurements fed into the forecast has significant error; plotting the available version alongside the complete measurements and forecast can eliminate this confusion.

- Forecast performance can potentially be improved by modeling the data revision process in addition to future observations, especially when a small change in past observations can cause a large change in the prediction target or associated forecast evaluations (as is sometimes the case for some timing and overall intensity targets), or when there is a high degree of error in earlier versions of measurements.

The above discussion and traditional time series forecasting methods such as SARIMA model the distribution of future observations of a time series of interest, $Y_{t+1..T}$, as a function of past observations of that time series, $Y_{1..t}$. However, in some settings, we do not have access to $Y_{1..t}$ itself but instead a sequence of tentative reports, $Y_1^{(1)}, Y_{1..2}^{(2)}, \ldots, Y_{1..t-1}^{(t-1)}, Y_{1..t}^{(t)}$, each adding a new (tentative) observation and revising previous values. For example:

- ILINet is a network of health care providers that voluntarily submit reports to CDC, which cleans and aggregates the data. Providers may differ in timeliness and frequency of reporting, and new providers may enter the system and might provide a chunk of data, and the aggregate measure of ILI prevalence is updated as additional providers submit or revise their data. CDC adjusts for the fact that different versions will be based on different numbers of providers by reporting the *proportion* of visits due to ILI, but earlier versions can still be biased, as slower or less frequent reporters may serve different populations with higher or lower typical ILI proportions than earlier reporters. The revisions may also be correlated across time, as a lower frequency or slower huge provider or group of similar providers may report a chunk of multiple weeks at the same time. CDC may also perform data cleaning, which can affect the entire season at the same time; for example, they may remove all data from a particular provider.

- FluSurv-NET is a surveillance network for laboratory-confirmed influenza hospitalizations. Many of the issues above still are applicable; for example, differences in types of laboratory test used, testing location, testing capacity, hospital administration, etc., can contribute to differences in timeliness of reporting between hospitals. Reporting may not take place until after a patient is discharged, which spreads reports apart further based on uncontrollable factors regarding duration of patients' illnesses. Additionally, reports may be revised after cases are ruled out as additional tests are performed. The combined effect is that the initially reported hospitalization rates are always or nearly always lower than the finalized figures, typically 50% of the finalized value, while later versions have a growing chance of overestimating the finalized value but are closer to it on average.

- Gross domestic product (GDP) and gross national product (GNP) estimates can also be revised over time. Previous work has named different types of updates and addressed the task of forecasting these updates in the context of Kalman filtering [34, 30, 43, 3].

Our goal is to build a distributional forecast of the entire, finalized time series of interest, $Y_{1..T_2}$, effectively leveraging information from tentative measurements $Y_1^{(1)}, Y_{1..2}^{(2)}, \ldots, Y_{1..t}^{(t)}$ and completed measurements $Y_{1..T_1}$ (where $T_1 \leq t$ and $T_1 < T_2$). That is, we want to jointly "backcast" (a.k.a. "backforecast", "back-forecast") $Y_{T_1..t}$ and forecast $Y_{t+1..T_2}$, and append the results to observations $Y_{1..T_1-1}$. There is an approach very similar in nature to the future trajectory simulation procedure above. We can simulate a random trajectory $Y_{1..T_2}^{\text{sim}}$ from the distribution of $Y_{1..T_2}$ given all tentative data $Y^{(1..t)}$ by chaining together $T_2 - T_1$ 1-step-ahead simulations:

- Let $Y_{1..T_1}^{\text{sim}} = Y_{1..T_1}$
- Draw $Y_{T_1+1}^{\text{sim}} \sim Y_{T_1+1} \mid Y^{(1..t)}, Y_{1..T_1} = Y_{1..T_1}^{\text{sim}}$
- Draw $Y_{T_1+2}^{\text{sim}} \sim Y_{T_1+2} \mid Y^{(1..t)}, Y_{1..T_1+1} = Y_{1..T_1+1}^{\text{sim}}$
- Draw $Y_{T_1+3}^{\text{sim}} \sim Y_{T_1+3} \mid Y^{(1..t)}, Y_{1..T_1+2} = Y_{1..T_1+2}^{\text{sim}}$
- . . .
- Draw $Y_{T_2}^{\text{sim}} \sim Y_{T_2} \mid Y^{(1..t)}, Y_{1..T_2-1} = Y_{1..T_2-1}^{\text{sim}}$

That is, we simulate the first observation Y_1, then feed that simulated value Y_1^{sim} into a model for Y_2, then feed the resulting value Y_2^{sim} along with Y_1^{sim} into a model for Y_3, and so on. The model selected for $Y_u \mid Y(1..t), Y_{1..u-1}$ is once again arbitrary, but it is often convenient to consider direct models of $\Psi^{[u]} \mid \mathbf{\Phi}^{[u]}$, where $\Psi^{[u]}$ can now depend on $Y^{(1..t)}, Y_{1..u}$ such that Y_u is recoverable, and $\mathbf{\Phi}^{[u]}$ is a feature vector prepared from $Y^{(1..t)}, Y_{1..u-1}$. The kernel delta density and locally linear quantile autoregression approaches have analogues in this setting: kernel residual density and quantile ARX (autoregression with exogenous variables):

Kernel residual density: uses kernel smoothing methods to estimate the conditional distribution of residuals $Y_u - \hat{Y}_u$ given some covariates when $u \leq t$, and of deltas $Y_u - Y_{u-1}$ when $u > t$.

Quantile ARX: uses quantile regression to estimate the conditional distribution of Y_u given a selection of features from $Y_{1..u-1}$ and $Y^{(1..t)}$.

3.1 Kernel residual density

Content in this section is taken from or based on material from [8].

The kernel residual density method chains together draws from conditional density estimates of $Y_u - \hat{Y}_u \mid \mathbf{\Phi}^{[u]}$ for u from T_1 to t and of $\Delta Y_u \mid \mathbf{\Phi}^{[u]}$ for u from $t+1$ to T_2, where $\mathbf{\Phi}^{[u]}$ is a function of $Y_{1..u-1}$ and $Y^{(1..t)}$. The delta density method can be seen as a special case where $T_1 = t$; $\hat{Y}_{1..t} = Y_{1..t}$, i.e., past values $Y_{1..t}$ are all treated as known and are simply duplicated in the simulated trajectories; and $\hat{Y}_{t+1..T_2} = Y_{t..T_2-1}$, i.e., the estimator Y_u when $u \geq t+1$ is the lagged version Y_{u-1}, which is filled in with a simulation except for when $u = t+1$ and so the complete observation $Y_{u-1} = Y_t$ has been observed. Each later residual $Y_u - Y_{u-1}$ corresponds to a delta in the delta density approach, ΔY_u.

Figure 3.1 shows sample forecasts over wILI trajectories generated by each of these approaches and compares them to some alternative methodologies.

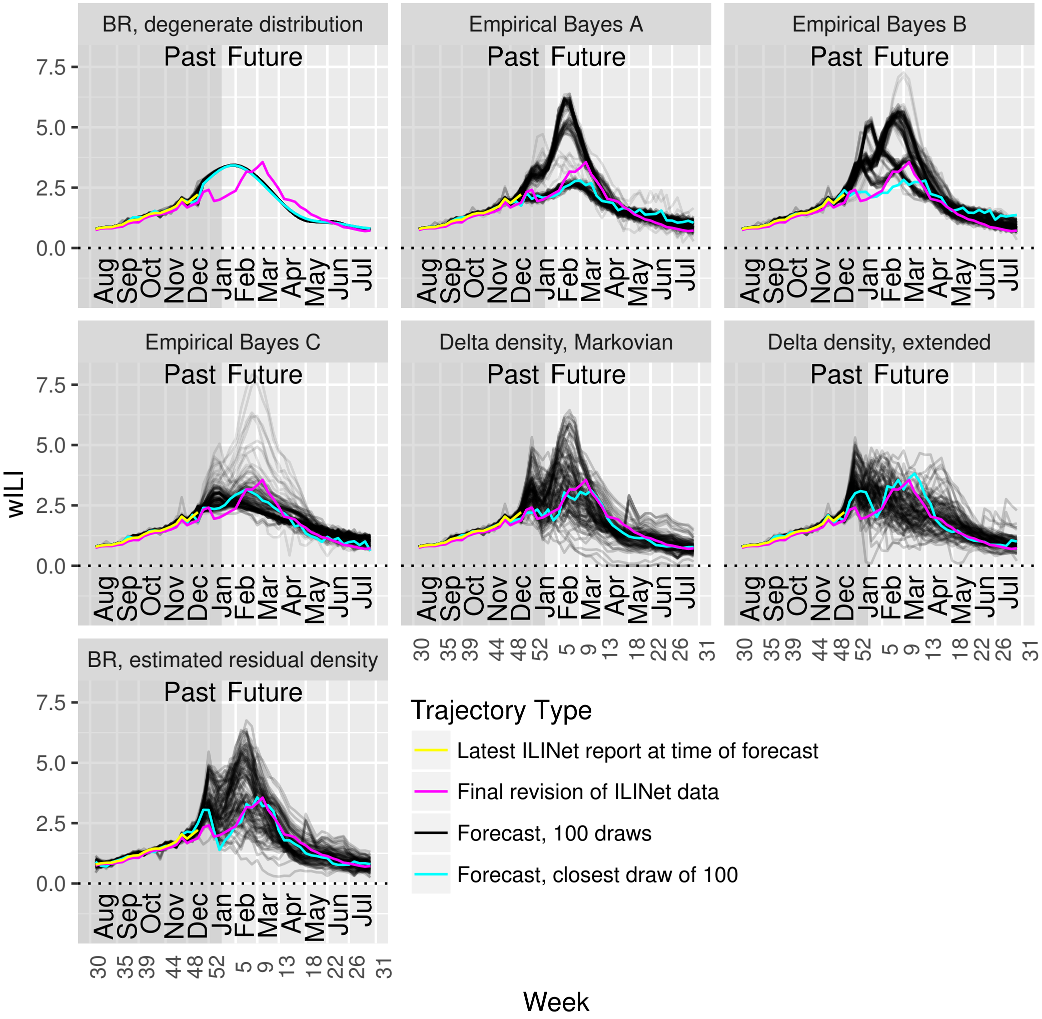

Figure 3.1: **Delta and residual density methods generate wider distributions over trajectories than methods that treat entire seasons as units.** These plots show sample forecasts of wILI trajectories generated from models that treat seasons as units (BR, Empirical Bayes) and from models incorporating delta and residual density methods. Yellow, the latest wILI report available for these forecasts; magenta, the ground truth wILI available at the beginning of the following season; black, a sample of 100 trajectories drawn from each model; cyan, the closest trajectory to the ground truth wILI from each sample of 100.

Figure 3.2 shows cross-validation performance estimates for the extended delta density method based on the following input data:

Ground truth, no nowcast: the ground truth wILI for the left-out season up to the forecast week is provided as input, resulting in an optimistic performance estimate;

Real-time data, no nowcast: the appropriate wILI report is used for data from the left-out season, but no adjustment is made for possible updates; this performance estimate is valid, but we can improve upon the underlying method;

Backcast, no nowcast: the appropriate wILI report is used for data from the left-out season, but we use a residual density method to "backcast" updates to this report; this performance estimate is valid, and the backcasting procedure significantly improves the log score;

Backcast, Gaussian nowcast: same as "Backcast, no nowcast" but with another week of simulated data added to the forecast, based on a Gaussian-distributed nowcast; and

Backcast, Student *t* nowcast: same as "Backcast, Gaussian nowcast" but using a Student t-distributed nowcast in place of the Gaussian nowcast.

Backcast, ensemble nowcast: same as the previous two but using the ensemble nowcast (which combines "no nowcast" with "Student t nowcast").

For every combination of target and forecast week, using ground truth as input rather than the appropriate version of these wILI observations produces either comparable or inflated performance estimates.

Using the "backcasting" method to model the difference between the ground truth and the available report helps close the gap between the update-ignorant method. The magnitude of the performance differences depends on the target and forecast week. Differences in mean scores for the short-term targets are small and may be reasonably explained by random chance alone; the largest potential difference appears to be an improvement in the "1 wk ahead" target by using backcasting. More significant differences appear in each of the seasonal targets following typical times for the corresponding onset or peak events; most of the improvement can be attributed to preventing the method from assigning inappropriately high probabilities (often 1) to events that look like they must or almost certainly will occur based on available wILI observations for past weeks, but which are ultimately not observed due to revisions of these observations. The magnitude of the mean log score improvement depends in part on the resolution of the log score bins; for example, wider bins for "Season peak percentage" may reduce the improvement in mean log score (but would also shrink the scale of all mean log scores). Similarly, the differences in scores may be reduced but not eliminated by use of multibin scores for evaluation or ensembles incorporating uniform components for forecasting.

Using the heavy-tailed Student t nowcasts or nowcast ensemble appears to improve on short-term forecasts without damaging performance on seasonal targets. The Gaussian nowcast has a similar effect as the other nowcasters except on the "1 wk ahead" target that it directly predicts: its distribution is too thin-tailed, resulting in lower mean log scores than using the forecaster by itself on this target.

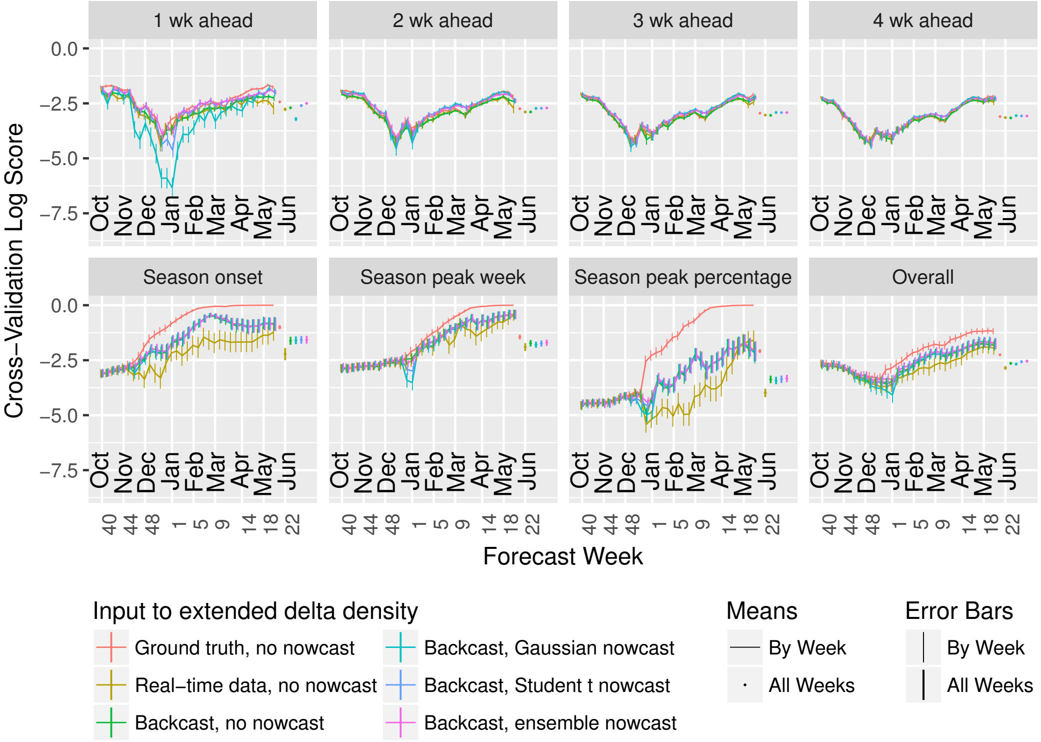

Figure 3.2: **Using finalized data for evaluation leads to optimistic estimates of performance, particularly for seasonal targets, "backcasting" improves predictions for seasonal targets, and nowcasting can improve predictions for short-term targets.** Mean log score of the extended delta density method, averaged across seasons 2010/2011 to 2015/2016, all locations, all targets, and forecast weeks 40 to 20, both broken down by target and averaged across all targets ("Overall"). Rough standard error bars for the mean score for each target (or overall) appear on the right, in addition to the error bars at each epi week.

3.2 Quantile ARX

Another candidate is a regularized, potentially corrupted, locally linear quantile regression on a subset of the conditioning covariates. One option is to simulate quantiles of Y_u as a linear function of the following covariates, along with a data weighting kernel and corrupting noise specification:

Name	Type	Description	Notation
Stable@u-4	Input	Stable/simulated value 4 weeks before	$Y_{u-4}/Y^{\text{sim}}_{u-4}$
Stable@u-3	Input	Stable/simulated value 3 weeks before	$Y_{u-3}/Y^{\text{sim}}_{u-3}$
Stable@u-2	Input	Stable/simulated value 2 weeks before	$Y_{u-2}/Y^{\text{sim}}_{u-2}$
Stable@u-1	Input	Stable/simulated value 1 weeks before	$Y_{u-1}/Y^{\text{sim}}_{u-1}$
Latest@u-1	Input	Latest value for 1 week before	$Y^{(t)}_{u-1}$
Latest@u	Input	Latest value for given week	$Y^{(t)}_{u}$
Latest@u+1	Input	Latest value for 1 week after	$Y^{(t)}_{u+1}$
Second-Latest@u	Input	Second-latest value for given week	$Y^{(t-1)}_{u}$
Stable@u	Output	Stable/simulated value for given week	Y_u/Y^{sim}_u

Table 3.1: One potential choice of $\mathbf{R}^{[u],\text{QARXlinear}}$ and $\Psi^{[u]}$.

Figure 3.3 visualizes this availability-dependent selection with a Bayes net.

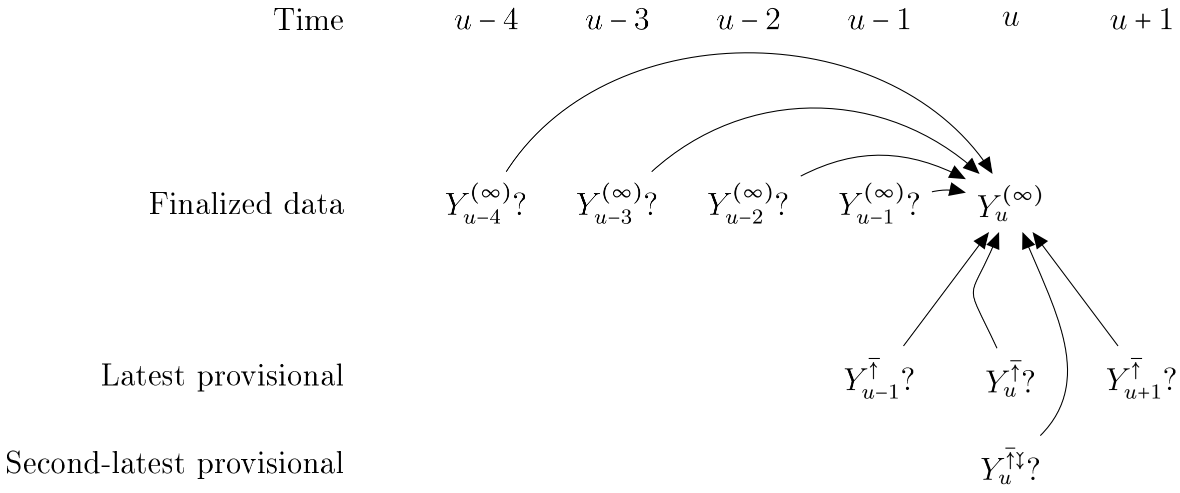
Time
$u-4$
$u-3$
$u-2$
$u-1$
u
$u+1$
Finalized data
$Y^{(\infty)}_{u-4}$?
$Y^{(\infty)}_{u-3}$?
$Y^{(\infty)}_{u-2}$?
$Y^{(\infty)}_{u-1}$?
$Y^{(\infty)}_{u}$
Latest provisional
$Y^{\bar{\upharpoonleft}}_{u-1}$?
$Y^{\bar{\upharpoonleft}}_{u}$?
$Y^{\bar{\upharpoonleft}}_{u+1}$?
Second-latest provisional
$Y^{\bar{\upharpoonleft}\downharpoonright}_{u}$?

Figure 3.3: Bayes net corresponding to earlier covariate table. Here, u could refer to a past, present, or future week, not just the current week. Question marks denote covariates that are included if available (observed/simulated) at test/application time. The $\bar{\upharpoonleft}$ symbol refers to the latest version of a wILI measurement available at test time (if there are any versions available), while $\bar{\upharpoonleft}\downharpoonright$ refers to the second-latest version of a wILI measurement available at test time (if there are ≥ 2 versions available).

Usually, we will start simulating with u's where most of this data is available, but at higher u some of the covariates will be excluded due to unavailability. For example, when simulating Y_{t+1}, the above covariate set would incorporate only $Y_t(t)$ and $Y^{\text{sim}}_{t-3..t}$. Training instances for the quantile regression model map these test covariates to the following training covariates:

- $Y^{(t+\Delta t)}_{(u-1..u+1)+\Delta t}$ corresponding to available $Y^{(t)}_{u-1..u+1}$
- $Y^{(t-1+\Delta t)}_{u+\Delta t}$ corresponding to available $Y^{(t-1)}_{u}$
- $Y_{(u-4..u-1)+\Delta t}$, corresponding to available $Y^{\text{sim}}_{u-4..u-1}$ or $Y_{u-4..u-1}$

The training set is limited to those instances where all of the above covariates are available. Weights can be assigned to training instances to encourage use of data from similar times of year and similar values of the covariates. Regularization is incorporated to prevent overfitting and remain robust in the face of collinearities. (Collinearities can arise, e.g., when the training set used fills in holes in records for $Y^{(1..t)}$ with other values from $Y^{(1..T)}$ or $Y_{1..T}$.)

Consider a more restrictive set of covariates: $Y^{(t)}_u$, if available, and Y^{sim}_{u-1} (or Y_{u-1} if available). Then the above process draws from conditional distributions that resemble a state space filter; for example, $Y_{T_1} \mid Y^{(t)}_{T_1}, Y_{T_1-1}$, using natural Markov assumptions, would be equivalent to $Y_{1..T_1} \mid Y^{(t)}_{T_1}, Y_{1..T_1-1}$, but would not consider information from observations for subsequent epiweeks such as $Y^{(t)}_{T_1+1..T_2}$. Since dependencies between data updates to observations for nearby weeks, we may want to ensure that this information is included. One simple way would be to simply add more elements from $Y^{(t)}_{T_1+1..T_2}$ as covariates when available, but this might lead to issues with fitting too many parameters, e.g., at T_1. An alternative would be to add a backward pass that parallels a state space smoothing algorithm; this approach may not be feasible when using complicated transformations or data weights. Yet another path would be to add subsequent values such as Y_{u+1} to the conditioning covariates and perform fitting and sampling using algorithms for the Multiple Quantile Graphical Model [2].

Chapter 4

Incorporating additional surveillance sources into spatiotemporal modeling

The previous two chapters discuss models for a single source of surveillance data that reports (multiple versions of) a single measurement for each time in the past for a particular location. This approach forgoes useful information available from additional traditional surveillance sources and a number of novel digital surveillance sources such as search query volume, social media activity, page hits, illness self-reporting, internet-integrated monitoring and testing devices, electronic health records, and insurance claims. Furthermore, the disease prevalence for each location of interest is forecast in isolation, which again neglects some available observations, and decreases the fidelity of any forecasts about the joint behavior of multiple locations. To address these issues, we can generalize the above approaches to forecast multiple data sources and/or locations at once, incorporating information from multiple auxiliary data streams. The remainder of the proposed work is to develop a joint modeling and simulation approach that incorporates dependencies across sources and locations using a careful selection of covariates, and to explore methods of incorporating dependencies between noise terms of sources and locations. In both single-location and multi-location settings, this task is often referred to as "nowcasting" or "nearcasting" when performing inference on times $u > t$ beyond the latest provisional data $\mathbf{Y}_{1..t}^{(t)}$ that correspond to or are close to the time of inference, taking advantage of some lower-latency external data $\mathbf{X}_u$ that is already available. Performing joint inference for $\mathbf{Y}_{T_1..T_2}$ covering times before, near, and after t combines the tasks of backcasting, nowcasting, and forecasting, and henceforth is referred to as "pancasting".

Baseline model

In the spatiotemporal pancasting context, one baseline for comparison is to treat each location in isolation, eschewing all exogenous covariates. A multi-location trajectory $\mathbf{Y}^{\text{sim}}_{1..T_2}$ would be formed simply by binding together location-specific simulations $Y^{\text{sim}}_{l,1..T_2}$ for all locations l.

Spatially isolated model with existing nowcasters

A limited number of additional data sources with sufficient temporal availability and matching resolution can be easily and directly added to the baseline model. For example, in the context of ILI forecasting, the ILI-Nearby [20, 21] system combines several novel data sources and autoregressive models into publicly accessible real-time "nowcasts" of CDC surveillance data, as well as weekly historical and retrospective nowcasts starting in the 2010/2011 flu season. Historical data about revisions of the CDC data is available starting sometime during the 2009/2010 season, so there is not much of a loss if training is limited to times when ILI-Nearby is available. The ILI-Nearby data can be treated in the same way as a provisional data point: added to the list of covariates used in the QARX pancasting routine. Figure 4.1 visualizes this availability-dependent selection with a Bayes net for a single location l.

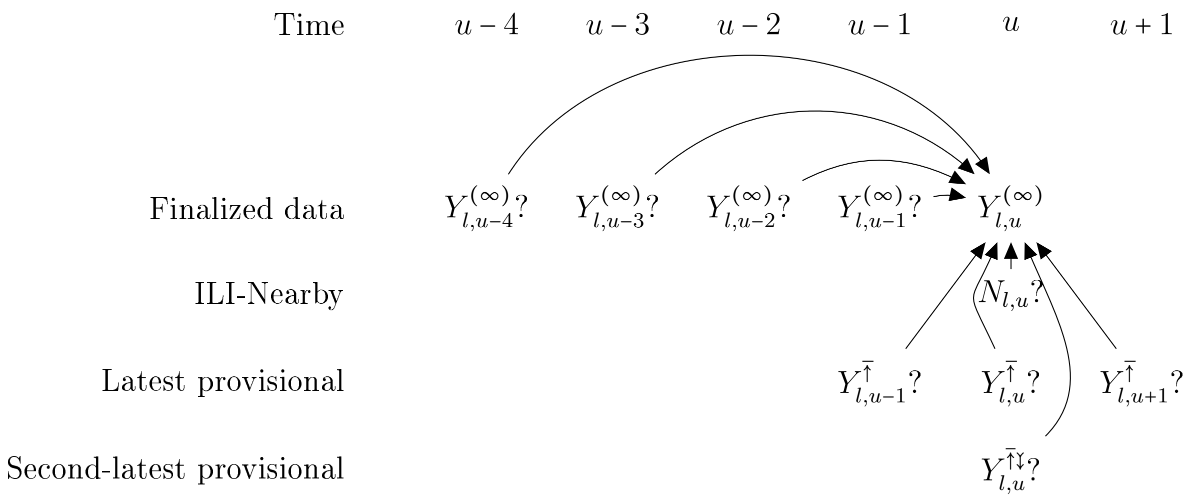

Figure 4.1: Bayes net corresponding to earlier covariate table. Here, u could refer to a past, present, or future week, not just the current week. Question marks denote covariates that are included if available (observed/simulated) at test/application time. The $\overline{\uparrow}$ symbol refers to the latest version of a wILI measurement available at test time (if there are any versions available), while $\overline{\uparrow}\,\overline{\downarrow}$ refers to the second-latest version of a wILI measurement available at test time (if there are ≥ 2 versions available).

Spatial covariates, independent quantile levels

Another straightforward extension to the above model is to augment the list of covariates used for predicting $Y_{l,u}$ with data from other locations $l' \neq l$. If there are a large number of such l', appropriate regularization is essential. One potential issue is that the quantile level $\tau_{l,u}$ drawn to generate $Y_{l,u}$ is independent of the quantile levels $\tau_{l',u}$ for other locations l'; this corresponds to conditional independence of $Y_{l,u}$ and $Y_{l',u}$ given a common set of spatial covariates. If observations are available at a fine enough time resolution, this assumption does not seem too objectionable; any biological events in l' in a sufficiently small time interval u have little opportunity to impact biological events in l in the same time window, and events from times $1..u-1$ can be incorporated in the covariate sets. We may still be concerned about potential reporting effects, e.g., tied to media reporting or variations in holiday impact.

Avenues for exploration

There are many additional avenues for exploration. Incorporating several new data sources directly, rather than relying partially on an external nowcasting solution, involves dealing with different ranges of data availability and reliability. Incorporating spatial interaction may require careful selection of a small number of spatial covariates and/or regularization to avoid overfitting. The independent quantile level noise model could be replaced with a more convincing noise model, e.g., using copulas, graphical models, multivariate quantiles, or continuous-time models. Additionally, modifications to the quantile regression formulation, such as non-crossing constraints, better covariate selection, and better kernel selection, could also contribute performance improvements. The goal of the remainder of the proposed work is to investigate one or multiple of these directions.

Chapter 5

Timeline

Dec 2018 – Feb 2019: set up comparison framework, initial methods

- Evaluation framework for backcasts, forecasts
- Robust quantile regression implementation
- Adding basic SIRS-inspired covariates
- Baseline & simple spatiotemporal methods

Mar 2019 – May 2019: pursue most promising directions

Forecasts: more complex mechanistic-inspired models

Backcasts: filtering → proper smoothing

Spatiotemporal pancasts: feature selection, dealing with missingness, and/or multivariate noise model

Jun 2019 – Jul 2019: additional analysis, evaluation, writing

Aug 2019: thesis oral

Bibliography

[1] Allison E Aiello, Rebecca M Coulborn, Vanessa Perez, and Elaine L Larson. Effect of hand hygiene on infectious disease risk in the community setting: a meta-analysis. *American journal of public health*, 98(8):1372–1381, 2008.

[2] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. The multiple quantile graphical model. In *Advances in Neural Information Processing Systems*, pages 3747–3755, 2016.

[3] S Borağan Aruoba. Data revisions are not well behaved. *Journal of money, credit and banking*, 40(2-3):319–340, 2008.

[4] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S. Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, Paola Velardi, Alessandro Vespignani, and Lyn Finelli. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16:357, 2016.

[5] Matthew Biggerstaff, Michael Johansson, David Alper, Logan C Brooks, Prithwish Chakraborty, David C Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the united states. *Epidemics*, 2018.

[6] Byron Boots. Spectral approaches to learning predictive representations. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2012.

[7] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Computational Biology*, 11(8):e1004382, 2015.

[8] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology*, 14(6):e1006134, 2018.

[9] Elizabeth Buckingham-Jeffery, Valerie Isham, and Thomas House. Gaussian process approximations for fast inference from infectious disease data. *Mathematical biosciences*, 2018.

[10] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[11] Centers for Disease Control and Prevention. Overview of influenza surveillance in the united states, 2013.

[12] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Estimating seasonal influenza-associated deaths in the United States | seasonal influenza (flu) |CDC, 2016.

[13] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Basics | Pandemic Influenza (Flu) | CDC. `https://www.cdc.gov/flu/pandemic-resources/basics/index.html`, 2016.

[14] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Pandemic Influenza | Pandemic Influenza (Flu) | CDC. `https://www.cdc.gov/flu/pandemic-resources/`, 2017.

[15] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav V Marathe, et al. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 262–270. SIAM, 2014.

[16] Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A Chitale, and F Ellis McKenzie. Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130, 2014.

[17] Benjamin J Cowling, Kwok-Hung Chan, Vicky J Fang, Calvin KY Cheng, Rita OP Fung, Winnie Wai, Joey Sin, Wing Hong Seto, Raymond Yung, Daniel WS Chu, et al. Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial. *Annals of internal medicine*, 151(7):437–446, 2009.

[18] Suruchi Deodhar, Jiangzhuo Chen, Mandy Wilson, Manikandan Soundarapandian, Keith Bisset, Bryan Lewis, Chris Barrett, and Madhav Marathe. Flu caster: A pervasive web application for high resolution situation assessment and forecasting of flu outbreaks. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 105–114. IEEE, 2015.

[19] Charbel El Bcheraoui, Ali H Mokdad, Laura Dwyer-Lindgren, Amelia Bertozzi-Villa, Rebecca W Stubbs, Chloe Morozoff, Shreya Shirude, Mohsen Naghavi, and Christopher JL Murray. Trends and patterns of differences in infectious disease mortality among us counties, 1980-2014. *JAMA*, 319(12):1248–1260, 2018.

[20] David C Farrow. *Modeling the Past, Present, and Future of Influenza.* Phd thesis, Carnegie Mellon University, 2016.

[21] David C. Farrow and Roni Rosenfeld. Multiple resolution nowcasting of influenza through sensor fusion. *Manuscript in preparation*, 2018.

[22] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, 10(11):e1003892, November 2014.

[23] Bruce E Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004.

[24] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLOS Computational Biology*, 11(5):e1004239, May 2015.

[25] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[26] Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[27] Michael Höhle, Sebastian Meyer, and Michaela Paul. *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*, 2017. R package version 1.13.1.

[28] Edward L Ionides, C Bretó, and Aaron A King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

[29] Edward L Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015.

[30] Jan PAM Jacobs and Simon Van Norden. Modeling data revisions: Measurement error and dynamics of “true” values. *Journal of Econometrics*, 161(2):101–109, 2011.

[31] Michael S Johannes, Nick Polson, and Seung M Yae. Quantile filtering and learning. 2009.

[32] Michael A. Johansson, Nicholas G. Reich, Aditi Hota, John S. Brownstein, and Mauricio Santillana. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Scientific Reports*, 6:33707, September 2016.

[33] Niall PAS Johnson and Juergen Mueller. Updating the accounts: global mortality of the 1918-1920" spanish" influenza pandemic. *Bulletin of the History of Medicine*, pages 105–115, 2002.

[34] Juan Manuel Julio et al. Modeling data revisions. *Borrador de Economía*, (641), 2011.

[35] Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Type-and subtype-specific influenza forecast. *American journal of epidemiology*, page 1, 2017.

[36] Aaron A King, Matthieu Domenech de Celles, Felicia MG Magpantay, and Pejman Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. *Proc. R. Soc. B*, 282(1806):20150347, 2015.

[37] Roger Koenker and Zhijie Xiao. Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990, 2006.

[38] Robert Kyeyagalire, Stefano Tempia, Adam L Cohen, Adrian D Smith, Johanna M McAnerney, Veerle Dermaux-Msimang, and Cheryl Cohen. Hospitalizations associated with influenza and respiratory syncytial virus among patients attending a network of private hospitals in south africa, 2007–2012. *BMC infectious diseases*, 14(1):694, 2014.

[39] Vasileios Lampos, Andrew C. Miller, Steve Crossan, and Christian Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, 5:12760, August 2015.

[40] Erik Lindström, Edward Ionides, Jan Frydendall, and Henrik Madsen. Efficient iterated filtering. *IFAC Proceedings Volumes*, 45(16):1785–1790, 2012.

[41] Rachel Lowe, Trevor C Bailey, David B Stephenson, Tim E Jupp, Richard J Graham, Christovam Barcellos, and Marilia Sá Carvalho. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in southeast brazil. *Statistics in medicine*, 32(5):864–883, 2013.

[42] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2095–2128, 2012.

[43] N Gregory Mankiw and Matthew D Shapiro. News or noise? an analysis of gnp revisions, 1986.

[44] Edson Zangiacomi Martinez, Elisângela Aparecida Soares da Silva, and Amaury Lelis Dal Fabbro. A sarima forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 44(4):436–440, 2011.

[45] Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. Disability-adjusted life years

(dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2197–2223, 2012.

[46] Richard W Niska and Iris Shimizu. Hospital preparedness for emergency response: United states, 2008. 2011.

[47] Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3):309–316, 2014.

[48] Philip D O'Neill. Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in medicine*, 29(20):2069–2077, 2010.

[49] Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y. Del Valle. Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy. August 2017.

[50] Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, 2014.

[51] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, page 201708856, 2018.

[52] Sen Pei and Jeffrey Shaman. Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nature communications*, 8(1):925, 2017.

[53] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria, 2003.

[54] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[55] Tamer Rabie and Valerie Curtis. Handwashing and risk of respiratory infections: a quantitative systematic review. *Tropical medicine & international health*, 11(3):258–267, 2006.

[56] Evan L Ray, Krzysztof Sakrejda, Stephen A Lauer, Michael A Johansson, and Nicholas G Reich. Infectious disease prediction with kernel conditional density estimation. *Statistics in medicine*, 36(30):4908–4929, 2017.

[57] MA Rolfes, IM Foppa, S Garg, B Flannery, L Brammer, JA Singleton, and others. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. `https://www.cdc.gov/flu/about/disease/2015-16.htm`, 2016.

[58] Robert E Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494, 1963.

[59] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[60] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4, 2013.

[61] James M Simmerman, Piyarat Suntarattiwong, Jens Levy, Richard G Jarman, Suchada Kaewchana, Robert V Gibbons, Ben J Cowling, Wiwan Sanasuttipun, Susan A Maloney, Timothy M Uyeki, et al. Findings from a household randomized controlled trial of hand washing and face masks to reduce influenza transmission in bangkok, thailand. *Influenza and other respiratory viruses*, 5(4):256–267, 2011.

[62] Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex Smola. Hilbert space embeddings of hidden markov models.(2010). 2010.

[63] Thorsten Suess, Cornelius Remschmidt, Susanne B Schink, Brunhilde Schweiger, Andreas Nitsche, Kati Schroeder, Joerg Doellinger, Jeanette Milde, Walter Haas, Irina Koehler, et al. The role of facemasks and hand hygiene in the prevention of influenza transmission in households: results from a cluster randomised trial; berlin, germany, 2009-2011. *BMC infectious diseases*, 12(1):26, 2012.

[64] Maha Talaat, Salma Afifi, Erica Dueger, Nagwa El-Ashry, Anthony Marfin, Amr Kandeel, Emad Mohareb, and Nasr El-Sayed. Effects of hand hygiene campaigns on incidence of laboratory-confirmed influenza and absenteeism in schoolchildren, cairo, egypt. *Emerging infectious diseases*, 17(4):619, 2011.

[65] MG Thompson, DK Shay, H Zhou, CB Bridges, PY Cheng, E Burns, JS Bresee, and NJ Cox. Estimates of deaths associated with seasonal influenza — united states, 1976-2007. *Morbidity and Mortality Weekly Report*, 59(33):1057, 2010.

[66] William W Thompson, David K Shay, Eric Weintraub, Lynnette Brammer, Nancy Cox, Larry J Anderson, and Keiji Fukuda. Mortality associated with influenza and respiratory syncytial virus in the united states. *Jama*, 289(2):179–186, 2003.

[67] Steffen Unkel, C Farrington, Paul H Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82, 2012.

[68] Cecile Viboud, Mark Miller, Donald R Olson, Michael Osterholm, and Lone Simonsen. Preliminary estimates of mortality and years of life lost associated with the 2009 a/h1n1 pandemic in the us and comparison with past influenza seasons. *PLoS currents*, 2, 2010.

[69] Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.

[70] Daren Wang. Predicting seasonal influenza epidemics, 2016. Carnegie Mellon University Department of Statistics Advanced Data Analysis project. Advisors: Ryan Tibshirani, Wilbert Van Panhuis.

[71] World Health Organization. The top 10 causes of death.

[72] World Health Organization. WHO | Influenza (Seasonal), 2016.

[73] Shihao Yang, Mauricio Santillana, John S. Brownstein, Josh Gray, Stewart Richardson, and S. C. Kou. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infectious Diseases*, 17:332, May 2017.

[74] Shihao Yang, Mauricio Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, November 2015.

[75] Wan Yang, Alicia Karspeck, and Jeffrey Shaman. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS computational biology*, 10(4):e1003583, 2014.

[76] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*, pages 311–319. International World Wide Web Conferences Steering Committee, 2017.

[77] Hong Zhou, William W Thompson, Cecile G Viboud, Corinne M Ringholz, Po-Yung Cheng, Claudia Steiner, Glen R Abedi, Larry J Anderson, Lynnette Brammer, and David K Shay. Hospitalizations associated with influenza and respiratory syncytial virus in the united states, 1993–2008. *Clinical infectious diseases*, 54(10):1427–1436, 2012.

[78] Christoph Zimmer, Sequoia I Leuba, Ted Cohen, and Reza Yaesoubi. Accurate quantification of uncertainty in epidemic parameter estimates and predictions using stochastic compartmental models. *Statistical methods in medical research*, page 0962280218805780, 2018.

[79] Christoph Zimmer, Reza Yaesoubi, and Ted Cohen. A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLoS computational biology*, 13(1):e1005257, 2017.