Predicting Spatio–Temporal Propagation of Seasonal Influenza Using Variational Gaussian Process Regression

Ransalu Senanayake[1,2], Simon O'Callaghan[2] and Fabio Ramos[1,2]
[1]School of Information Technologies, The University of Sydney, Australia
[2]National ICT Australia (NICTA)

Abstract

Understanding and predicting how influenza propagates is vital to reduce its impact. In this paper we develop a nonparametric model based on Gaussian process (GP) regression to capture the complex spatial and temporal dependencies present in the data. A stochastic variational inference approach was adopted to address scalability. Rather than modeling the problem as a time-series as in many studies, we capture the space-time dependencies by combining different kernels. A kernel averaging technique which converts spatially-diffused point processes to an area process is proposed to model geographical distribution. Additionally, to accurately model the variable behavior of the time-series, the GP kernel is further modified to account for non-stationarity and seasonality. Experimental results on two datasets of state-wide US weekly flu-counts consisting of 19,698 and 89,474 data points, ranging over several years, illustrate the robustness of the model as a tool for further epidemiological investigations.

Introduction

Influenza is an infectious disease caused by a virus whose activity is typically peaked during winter. Influenza is responsible for more than 2% of all deaths in the US which is the highest mortality due to any infectious disease (CDC). Moreover, deaths can be in the order of millions during a pandemic: 0.4 million deaths during 2009 swine flu, 1 million during 1968 Hong Kong flu, 2 million during 1959 Asian flu and 40–50 million during 1918 Spanish flu (WHO 2005).

Annihilating all sources of virus (types A, B and C) are impossible. Although vaccines can be designed to abate the overrun of influenza virus, they cannot be used over consecutive years as influenza virus change its biological form frequently due to its high mutation rate. Additionally, modification and manufacturing of vaccines on a large scale takes time. A plausible solution to reduce the annual death rate and to prevent transforming a seasonal epidemic to a pandemic is to subside transmission. Consequently, envisaging how the virus spreads across geographical areas with time, i.e. spatio-temporal dynamics, is vital.

Influenza Prediction

In most studies, modeling influenza or influenza-like illnesses (ILI) has been considered as a time series problem. Therefore, autoregressive models have been the popular choice of many researchers (Dugas et al. 2013) (Viboud et al. 2003). While many studies model seasonal effects, Wang et al. (2015) have focused on improving the short-term prediction accuracy. Similarly, variations of particle filters and ensemble filters have been used to predict influenza activity. Yang et al. (2014) compared six state-of-the-art filters and concluded that their results are comparable. Additionally, ensemble of other simple regressions such as matrix factorized based regression, nearest neighbor based regression, etc. (Chakraborty et al. 2014) have been tested . Although these autoregressive, filter-based and ensemble models are convenient to use, they ignore the disease's strong geographical dependencies. Crucially, they do not provide an uncertainty measure about the prediction which prohibits any risk-based decision-making process.

Although attempts to map retrospective spatio-temporal effects of influenza and other diseases are not rare, correlating space and time is not widely studied. A hierarchical Bayesian parametric model has been proposed for modeling the spatio-temporal interaction of generic disease mapping (Waller et al. 1997) however it lacks the ability to forecast future outbreaks. Unlike influenza whose case count in a given place can suddenly increase or decrease, lung cancer data that they have used in experiments are smooth in both space and time. Moreover, Markov chain Monte Carlo (McMC) calculations in their solution ultimately limits the maximum size of the dataset that can be considered.

Bayesian Nonparametric Models

For any dataset, it is less desirable to utilize a parametric model unless all of its dependencies are known. Although few studies show the correlation between influenza and external factors (Charland et al. 2009) (Viboud et al. 2006), all major factors that affect influenza activity remain elusive. This is mainly due to the complexity of the problem, amount of data and the number of plausible factors. As a result, this eventually becomes a big data analytics problem (Chakraborty et al. 2014) (Davidson, Haim, and Radin 2015) with unknown dependencies and hence many aforementioned parametric methods are less useful.

Considering the disadvantages of exploiting parametric models in influenza or ILI prediction and the increasing popularity of non-parametric models for problems with complex interactions, we use a Gaussian process (GP) — a non-parametric Bayesian method — for prediction. Importantly, such methods always provide the uncertainty of prediction (typically as standard deviation) which is essential for any meaningful forecast. GP regression is best known for its superior performance in multi-dimensional spatial models. In contrast to the conventional use of GP regression as an interpolation model for smooth and small-scale datasets, we demonstrate Big Data GP (Hensman, Fusi, and Lawrence 2013) for prediction purposes with highly variable influenza data. Moreover, rather than considering time as merely another dimension of the GP model, we carefully combine kernels to model non-stationary and periodic behavior of the time series.

Although Carrat and Valleron (1992) have attempted to use GP for modeling influenza, their results were inaccurate mostly due to the infancy of GP methods a couple of decades ago. The model could neither capture spatio-temporal dependencies nor could it make forecasts. In our study, we capture the space-time interaction by combining kernels with addition and multiplication. More recently, Saavedra et al. (2015) used a mixture of GP-priors with thin plate splines for modeling influenza cases in Western Australia. Since splines are typically used for interpolation in the space domain, it is not clear how the model can be extended for predicting into the future.

This paper presents a novel spatio–temporal model for influenza prediction trained from a large amount of data and able to capture long and short-term patterns. The model is a powerful platform for epidemiologists to identify factors that affect influenza transmission. Further, our contribution extends to the Gaussian process community by demonstrating that i) GPs are a valid choice for prediction of spatial-temporal processes and; ii) stochastic variational inference can be used to handle large amount of data which would otherwise be impossible with conventional GP learning techniques (Hensman, Fusi, and Lawrence 2013).

Large Scale Gaussian Process Regression

Gaussian Process Regression

The objective is to predict the output distribution $p(y_* | \mathbf{x}_*, \mathcal{D}; \theta) = \mathcal{N}(\mu_*, \sigma_*^2)$ for an arbitrary D-dimensional input $\mathbf{x}_*$, given training data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$. Consider the model $y_i = f(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The joint Gaussian distribution for the latent process $f(\mathbf{x})$ having a zero mean and covaraince function $k(\mathbf{x}_i, \mathbf{x}_j)$ is given by the Gaussian process $f(\mathbf{x}) \sim \mathcal{GP}\big(0, k(\mathbf{x}_i, \mathbf{x}_j)\big)$. The popular choice for $k(\mathbf{x}_i, \mathbf{x}_j)$ is $\sigma^2 \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / 2l^2)$ where hyperparameters $\theta = (\sigma^2, l)$ are typically learned using a gradient-based optimization technique by maximizing the log-marginal likelihood (Rasmussen 2006).

For $i, j = 1, 2, \dots, N$, $k(\mathbf{x}_i, \mathbf{x}_j)$ generates the covariance matrix K_{NN} where the subscripts indicate the size of the matrix. The equations (Rasmussen 2006) of log-marginal likelihood and inference contain the term K_{NN}^{-1}, meaning that kernel inversion is essential in both training and prediction phases. Since this inversion has a computational complexity of $\mathcal{O}(N^3)$, the number of data points is typically limited to around 2000 to perform learning and inference in a feasible time using a standard PC.

Nyström Approximation for GP

Although it is possible to "greedily" use a subset of data (SoD) points, $M < N$ to make approximations through subsampling, previous methods attempted to make use of all data yet keeping the complexity manageable. Quiñonero-Candela and Rasmussen (2005) have adopted Nyström method, which was originally used as a device for numerical integration, to approximate the covariance matrix as,

$$K_{NN} \approx \widetilde{K}_{NN} = K_{NM} K_{MM}^{-1} K_{NM}^{\top}. \qquad (1)$$

K_{MM} is generated from a set of M *inducing inputs* $\check{\mathbf{x}}$ s.t. $M \ll N$. Although it has several forms with slight modifications (Quiñonero-Candela and Rasmussen 2005), projected process (PP) approach is popular and intuitive.

All of these methods have complexity $\mathcal{O}(M^2 N)$. However, if weekly influenza cases in all states of the US are considered, there are around 2500 data points per year. Since it is required to have training data of past few years, even this sparse approximation is not useful. Further, training and inference becomes clearly infeasible when using a multitude of hyperparameters to model the spatio-temporal dynamics of the past few decades.

Stochastic Variational Inference for GP

Following a different philosophy – variational inference – Titsias (2009) obtained the same predictive distribution as PP. As a successor, (Hensman, Fusi, and Lawrence 2013) proposed the Big Data GP model which has a computational complexity of $\mathcal{O}(M^3)$. Unlike the SoD model it considers all data points (at least a majority) and hence more representative.

For notational simplicity, let the latent function of inputs be $\mathbf{f} := \mathbf{f}(\mathbf{x})$ and latent function of inducing inputs be $\check{\mathbf{f}} := \mathbf{f}(\check{\mathbf{x}})$. Therefore, $p(\check{\mathbf{f}}) = \mathcal{N}(\check{\mathbf{f}}; \mathbf{0}, K_{MM})$. The novelty of Hensman's work is that they introduced an explicit variational distribution $\hat{p}(\check{\mathbf{f}}) = \mathcal{N}(\check{\mathbf{f}}; \mathbf{m}, \mathbf{S})$ as an approximating distribution and derived a decomposable lower bound,

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\log \mathcal{N}(y_i | K_{MN,:i}^{\top} K_{MM}^{-1} \mathbf{m}, \sigma^2) - \frac{1}{2\sigma^2} \tilde{K}_{NN,ii} - \frac{1}{2} \mathrm{tr}(\mathbf{S} \mathbf{\Lambda}_i) \Big) - \mathbb{KL}\big(\hat{p}(\check{\mathbf{f}}) \| p(\check{\mathbf{f}})\big), \qquad (2)$$

where $\mathbf{\Lambda}_i = \sigma^2 K_{MM}^{-1} K_{MN,:i} K_{MN,:i}^{\top} K_{MM}^{-1}$, : represents all elements, and $\mathbb{KL}$ is the Kullback–Leibler divergence.

Optimization under this framework is taken place using *natural gradients* of $\frac{\partial \mathcal{L}}{\partial \mathbf{m}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{S}}$ to approximate $\hat{p}(\check{\mathbf{f}})$. In each natural gradient descent (NGD) step, hyperparameters θ are optimized using stochastic gradient descent (SGD) with a learning rate α: $\theta \leftarrow \theta + \alpha \frac{\partial \mathcal{L}}{\partial \theta}$. The requirements for SGD are satisfied as $\mathcal{L}$ is represented as a summation.

Furthermore, it is possible to consider mini-batches of size $R \in \{1, 2, \cdots, N\}$ in each gradient descent step. In our experiments, mini-batches were chosen s.t. $R \leq M$ to maintain $\mathcal{O}(M^3)$.

Having approximated $\hat{p}(\breve{\mathbf{f}})$ and θ in the training phase, the predictive distribution $\mathbf{y}_* \sim \mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$ for $\mathbf{x}_*$ can be inferred from (3) and (4),

$$\mu_* = K_{*M} K_{MM}^{-1} \mathbf{m}, \tag{3}$$

$$\mathbf{\Sigma}_* = K_{**} - K_{*M}(K_{MM}^{-1}, -K_{MM}^{-1} \mathbf{S} K_{MM}^{-1}) K_{*M}^{\top} \tag{4}$$

$$\sigma_* = \mathrm{diag}(\mathbf{\Sigma}_*). \tag{5}$$

Constructing the Spatio-Temporal Kernel

The kernel employed in our model consists of three separate components (time, space, and cross-covariance) whose weights are learned during an optimization phase. This dichotomy allows us to explicitly incorporate domain knowledge about the behavior of the disease in each dimension.

Let the independent variables in the training dataset be $\{\mathbf{t}, X\}$ having N samples where $\mathbf{t} = \{t\}_{i=1}^{N} \in \mathbb{R}^{N \times 1}$ is the time component and $X = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times 2}$ is the space component (longitude and latitude).

Time Component

Periodicity: Given the fact that flu activity is high during winter and low during summer, a periodic kernel (Rasmussen 2006) (Guizilini and Ramos 2015) was used. The (t_i, t_j) element of the covariance matrix is given by,

$$k_{sin}(t_i, t_j; \theta) = \sigma_{sin}^2 \exp\left(- \frac{2 \sin^2(\pi f \Delta_t)}{l_{sin}^2} \right), \tag{6}$$

where $\Delta_t = |t_i - t_j|$ is the distance metric and, $\theta_{sin} = (\sigma_{sin}^2, l_{sin}, f)$ are scaler hyperparameters. The frequency f is expected to be around 1 year.

Non-stationarity: Typically, GPs assume a constant length-scale l across the entire input space and hence the covariance only depends on the distance Δ between two input locations t_i and t_j. Intuitively, length-scale represents the realm of correlation. For instance, if the output variable decreases sharply, farther points should not be affected requiring a short length-scale only in the local region. To account for this non-stationarity, an input dependent length-scale was used, following the method of Paciorek and Schervish (2004). Paciorek's kernel was used successfully for 3-D digital terrain modeling (Lang, Plagemann, and Burgard 2007) and environmental monitoring (Garg, Singh, and Ramos 2012). Other popular treatments to deal with non-stationarity such as mixture of GPs (Tresp 2000) are less suitable for predictive analytics as regions of mixtures should be pre-defined. For multidimensional input $\mathbf{z}$, Paciorek's squared-exponential kernel is defined as,

$$k_{pac}(\mathbf{z_i}, \mathbf{z_j}) = \sigma_{pac}^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^{-\frac{1}{2}} \exp\left(- (\mathbf{z_i} - \mathbf{z_j})^{\top} \left(\frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\mathbf{z_i} - \mathbf{z_j}) \right), \tag{7}$$

where $\Sigma_{\bullet} := \Sigma_{pac}(\mathbf{z}_{\bullet})$ is the local squared-exponential kernel at input location $\mathbf{z}_{\bullet}$. Although the derivation of (7) is based on convolutional kernel $\int_{\mathbb{R}^2} k_{\mathbf{z}_i}(\mathbf{u}) k_{\mathbf{z}_j}(\mathbf{u}) d\mathbf{u}$ which is positive semi-definite, intuitively it can be thought as the average between two.

Since the spatial distribution of influenza activity is observed to be smooth and only the time series is jagged, Paciorek's kernel is used only for the time component. Since the time component is one-dimensional, $\Sigma_{\bullet}$ reduces to $l_{\bullet} := l_{pac}(t_{\bullet})$ and hence the modified Paciorek's kernel is defined as,

$$k_{pac}(t_i, t_j) = \sigma_{pac}^2 \left(\frac{2 l_i l_j}{(l_i^2 + l_j^2)} \right)^{\frac{1}{2}} \exp\left(- \frac{2 \Delta_t^2}{(l_i^2 + l_j^2)} \right). \tag{8}$$

Length-scale hyperparameter $l_{\bullet}$ has to be evaluated for every $t_{\bullet}$. The original framework used a separate GP to model the length scale in a hierarchical formulation and used Markov chain Monte Carlo (McMC) sampling to learn the model. Obviously, this framework is not computationally efficient, as noted by authors, due to i) hierarchical nature and ii) use of McMC. As shown in Figure 1, we propose to place length-scales $\mathbf{l}_{pac}$ for summer and winter each year. We call the length-scale bases $\bar{\mathbf{t}}$, following the notation of Plagemann, Kersting, and Burgard (2008), and define another internal GP for length-scale $\mathbf{l}_{pac} \sim \mathcal{GP}_{\bar{\mathbf{t}}}(\sigma_{\bar{\mathbf{t}}}^2, l_{\bar{\mathbf{t}}})$. In the learning process, as an approximation, hyperparameters of the length-scale GP which is based on a squared-exponential kernel, are optimized using SGD together with all other hyperparameters $\theta_{pac} = (\sigma_{pac}^2, \mathbf{l}_{pac}, \sigma_{\bar{\mathbf{t}}}^2, l_{\bar{\mathbf{t}}})$ thus breaking the hierarchy and eliminating the requirement of McMC.

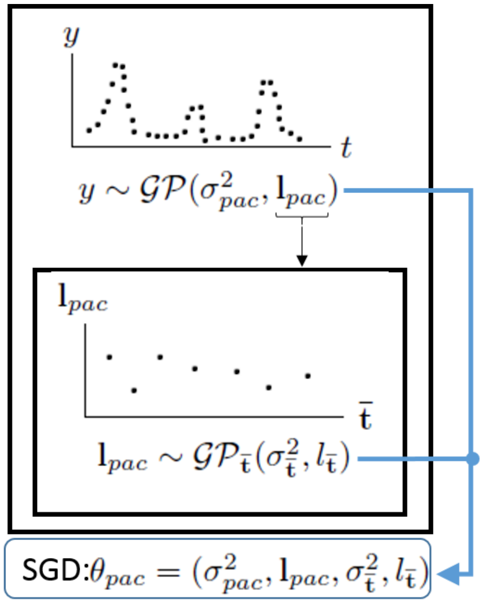
y
t
$y \sim \mathcal{GP}(\sigma_{pac}^2, \mathbf{l}_{pac})$
$\mathbf{l}_{pac}$
$\bar{\mathbf{t}}$
$\mathbf{l}_{pac} \sim \mathcal{GP}_{\bar{\mathbf{t}}}(\sigma_{\bar{\mathbf{t}}}^2, l_{\bar{\mathbf{t}}})$
SGD:$\theta_{pac} = (\sigma_{pac}^2, \mathbf{l}_{pac}, \sigma_{\bar{\mathbf{t}}}^2, l_{\bar{\mathbf{t}}})$

Figure 1: Schematic diagram of the non-stationary model. All hyperparameters of Paciorek's non-stationary kernel are learned together.

Short-term and long-term trends: Although Paciorek's kernel is used to capture in-season variations, long-term variations will not be captured as length-scales are relatively small. Therefore, another squared-exponential kernel (9) having a long length-scale l_{exp1} is superimposed for this purpose. This is done by initializing the length-scale to a proportionately large value which is further optimized using SGD,

$$k_{exp1}(t_i, t_j) = \sigma^2_{exp1} \exp\Big(-\frac{\Delta_t^2}{l^2_{exp1}}\Big). \tag{9}$$

Combining the three kernels yields the final kernel for time,

$$k_{time} = \underbrace{k_{sin}(t_i, t_j)}_{\text{periodic - (6)}} + \underbrace{k_{pac}(t_i, t_j)}_{\text{short-term - (8)}} + \underbrace{k_{exp1}(t_i, t_j)}_{\text{long-term - (9)}}. \tag{10}$$

Space Component

A squared-exponential kernel can be used to model the spatial variation $\mathbf{x}_\bullet$ which has two sub-components, latitude and longitude as indicated below,

$$k_{exp2}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2_{exp2} \exp\Big(-\frac{\Delta_x^2}{l^2_{exp2}}\Big), \tag{11}$$

where $\Delta_x = \|\mathbf{x}_i - \mathbf{x}_j\|_2$.

Averaging kernel: Influenza case count is given for a geographical region (e.g. a state in the US), not the count of a specific position. When using (11) to represent regions, the question where to place $\mathbf{x}_\bullet$ naturally arises as the equation is valid only for point processes. Centroid is a good choice if regions are symmetric or small in area. Otherwise, the morphology of the region should be taken into account for accurate representation. Inspired by the integral kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \int\int_{\mathcal{A}_{\mathbf{x}i}} \int\int_{\mathcal{A}_{\mathbf{x}j}} k\big(\mathbf{x}_i(\mathbf{u}), \mathbf{x}_j(\mathbf{v})\big) d\mathbf{u} d\mathbf{v}$ for areas $\mathcal{A}_{\mathbf{x}i}$ and $\mathcal{A}_{\mathbf{x}j}$ which requires expensive quadrature calculations (O'Callaghan and Ramos 2011), (Reid 2011), we propose an averaging kernel (12) illustrated in Figure. 2. It is straightforward to prove that the kernel is positive semi-definite,

$$k_{space}(S_{\mathbf{x}i}, S_{\mathbf{x}j}) = \frac{1}{|S_{\mathbf{x}i}||S_{\mathbf{x}j}|} \sum_{\mathbf{u} \in S_{\mathbf{x}i}} \sum_{\mathbf{v} \in S_{\mathbf{x}j}} k_{exp2}(\mathbf{u}, \mathbf{v}), \tag{12}$$

where $|S_\bullet|$ is the cardinality of set $S_\bullet$ which consists of a finite number of points in the area of interest as illustrated in Figure 2. For instance, a region can be better represented using multiple pairs of equi-spaced longitude–latitude coordinates rather than a pair of longitude–latitude coordinates.

For a query point $\mathbf{x}_*$, output can be calculated based on $k_{space}(S_{\mathbf{x}_*}, S_{\mathbf{x}j})$ as in (12) to obtain coarse boundaries and $k_{space}(\mathbf{x}_*, S_{\mathbf{x}j})$ as in (13) to obtain smooth boundaries,

$$k_{space*}(\mathbf{x}_*, S_{\mathbf{x}j}) = \frac{1}{|S_{\mathbf{x}j}|} \sum_{\mathbf{v} \in S_{\mathbf{x}j}} k_{exp2}(\mathbf{x}_*, \mathbf{v}). \tag{13}$$

The averaging kernel may be unnecessary if the area of $S_\bullet$ is small compared to the entire distribution.

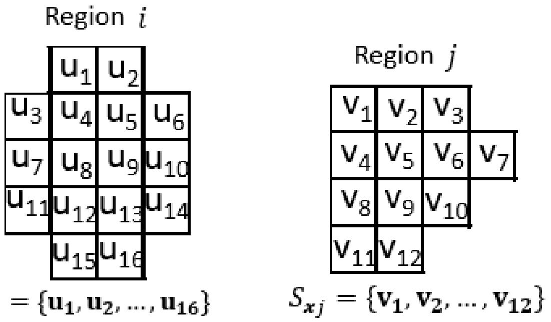
Region i
$\mathbf{u}_1$ $\mathbf{u}_2$
$\mathbf{u}_3$ $\mathbf{u}_4$ $\mathbf{u}_5$ $\mathbf{u}_6$
$\mathbf{u}_7$ $\mathbf{u}_8$ $\mathbf{u}_9$ $\mathbf{u}_{10}$
$\mathbf{u}_{11}$ $\mathbf{u}_{12}$ $\mathbf{u}_{13}$ $\mathbf{u}_{14}$
$\mathbf{u}_{15}$ $\mathbf{u}_{16}$
$S_{\boldsymbol{x}i} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{16}\}$
Region j
$\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$
$\mathbf{v}_4$ $\mathbf{v}_5$ $\mathbf{v}_6$ $\mathbf{v}_7$
$\mathbf{v}_8$ $\mathbf{v}_9$ $\mathbf{v}_{10}$
$\mathbf{v}_{11}$ $\mathbf{v}_{12}$
$S_{\boldsymbol{x}j} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{12}\}$

Figure 2: Illustration representing the averaging kernel idea.

Spatio-Temporal Covariance

To capture relationships between space and time, the time kernel (10) is multiplied by another squared-exponential kernel $k_{exp3}(\mathbf{x}_i, \mathbf{x}_j)$ which has a similar form of (11) with different hyperparameters $\theta_{space\text{-}time} = (\sigma^2_{exp3}, l_{exp3})$,

$$k_{space\text{-}time} = k_{exp3}(\mathbf{x}_i, \mathbf{x}_j) \times k_{time}. \tag{14}$$

By combining (10), (12) and (14), the final kernel is obtained,

$$k = k_{time} + k_{space} + k_{space\text{-}time}. \tag{15}$$

Experiments

To demonstrate the predictive power of the method for the propagation of influenza, two datasets were used:

1) Google flu trend (GFT): Recent studies show an increasing interest in using web query data to predict flu (Chakraborty et al. 2014). The US state-wide GFT (Ginsberg et al. 2009) data (flu count/population) of 402 weeks from 02/Dec/2007 to 09/Aug/2015 were used in the experiments. Alaska and Hawaii were excluded from the analysis as they are not geographically connected to the mainland, resulting in 49 states including District of Columbia (19,698 data points).

2) 1972-2006 dataset (CDC): In order to analyze long-term trends, exact ILI counts (CDC) of 1826 weeks from 1972 to 2006 (Viboud et al. 2006) were used. The same states as in the GFT dataset were used resulting in 89,474 data points. Compared to this dataset, GFT contains more recent data, however the two datasets were not merged due to the different sources and sampling methods.

Experiment 1: Effect of each time component. In order to verify the effect of each component in k_{time} (Eq. 10), total US flu count in the CDC dataset was used with $M = N/10$ to run the regression model separately with kernel combinations $(k_{sin} + noise)$, $(k_{sin} + k_{exp1} + noise)$ and $(k_{sin} + k_{exp1} + k_{pac} + noise)$. As illustrated in Figure 3 (a) and (b), although periodic kernel alone has constant peaks, as expected, adding a squared–exponential kernel clearly captures the long-term variation, decreasing the

MSE to 0.0433 from 0.0552. Superimposing the short-term Paciorek kernel with the aforementioned combination, as illustrated in Figure 3 (c), further decreases MSE to 0.0167, with 70% overall decrease in MSE.

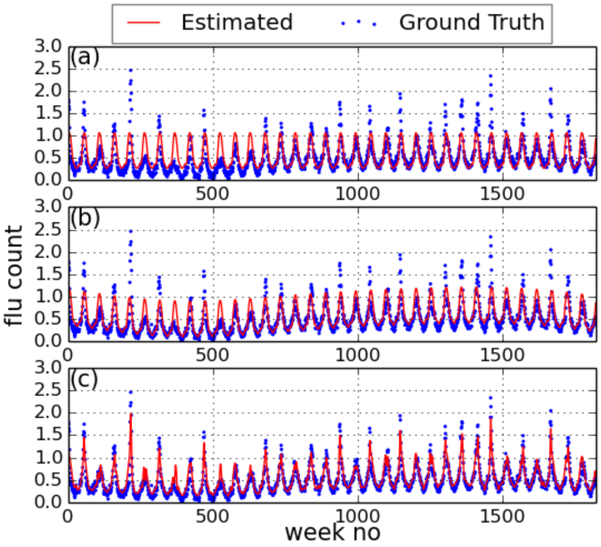

Figure 3: Experiment 1. (a) Periodic kernel with noise. (b) Adding long-term kernel to a. (c) Adding short-term non-stationary kernel to b.

Experiment 2: The averaging kernel. Flu-counts in both datasets are state-wide composites and hence it is difficult to select an approximate center-point to each state. Since all states are not the same shape and size, each state was represented by a set of points with 1° latitude and longitude accuracy. Then (Eq. 12) was used to calculate the covariance. Figure 4 (a) shows state-wise flu-count at approximate center-points at a given time (week of 02/Dec/2007) while (b) shows the mean estimations calculated based on smooth querying (Eq. 13).

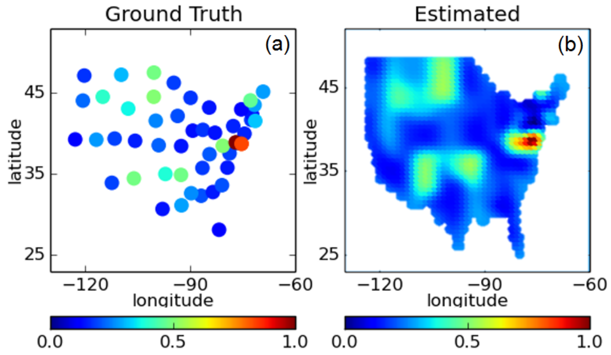

Figure 4: Experiment 2. (a) Ground truth state-wise flu-count marked at center points of states. (b) Estimated predictive mean.

Experiment 3: Space-time modeling. In order to further investigate the effects of the averaging kernel, the US was divided into 4 regions as in other major flu propagation related studies (Viboud et al. 2006): East includes the following states CT, DE, ME, MA, NH, NJ, NY, PA, RI and VT; West includes AZ, CA, CO, ID, MT, NV, NM, OR, UT, WA and WY; South includes AL, AR, DC, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA and WV; and Midwest includes IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD and WI. Figure 5 shows mean estimations and standard deviations given by (3) and (5) for East region using the CDC dataset.

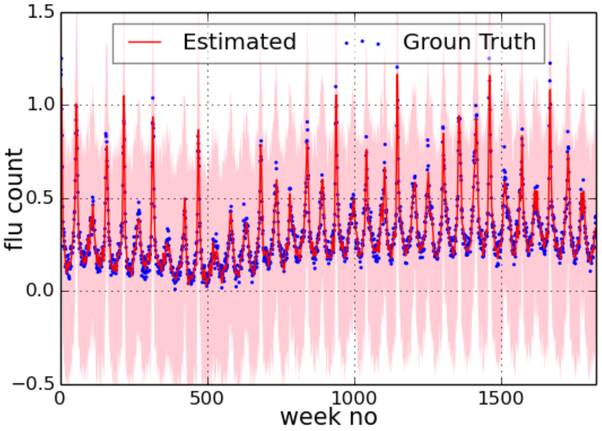

Figure 5: Experiment 3. Prediction and associated standard deviation for the East region.

Experiment 4: Scalability. The space-time model was run on both datasets demonstrating the use of stochastic variational inference in GPs for predictions with thousands of datapoints. Choosing only $M = N/4$ of data as inducing points average of MSE was found to be 0.0014 for GFT. [1]

Experiment 5: Prediction. One, two and fifty-two weeks ahead predictions were made for years 2013, 2014 and 2015 based on GFT data with $M = N/10$. Figure 6 shows one-week ahead prediction results for six states. It was observed that the model sometimes attempts to overestimate in the non-flu season due to the arrangement of inducing points. Nevertheless, more accurate results can be obtained by increasing the number of inducing points with a higher computational cost or using a distributed Gaussian process.

Table 1: Experiment 5. Mean squared error (MSE) averaged over states for spatio-temporal prediction of 150 weeks.

Method	**1 week**	**2 weeks**	**1 year**
Our approach	0.0043	0.0128	0.2900
GP - subset of data (SoD)	0.0156	0.0157	0.8642
k-NN regression	0.0068	0.0105	0.3586
LS polynomial regression	0.0169	0.0211	1.6557

[1] Supplementary materials - https://goo.gl/VCuVW3

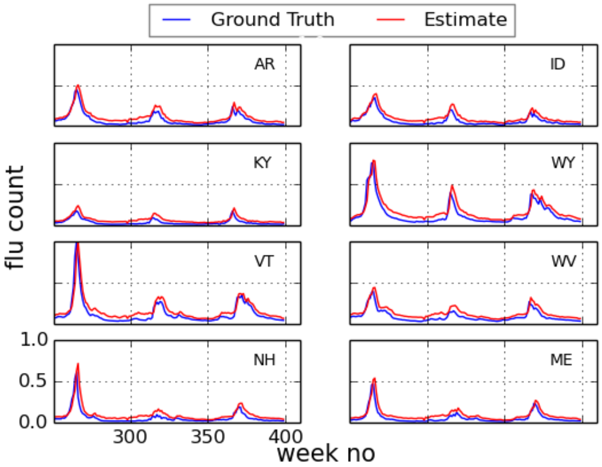

Figure 6: Experiment 5. One-week-ahead prediction results for 6 states.

Our method was compared to other conventional methods for ILI prediction - the results of which are shown in Table 1. The MSE over all states of our method generally outperforms other methods, although k-NN regression (Chakraborty et al. 2014) is comparable. For instance, n-step-ahead predictions of k-NN method (a non-probabilistic method) approximately follows the last training data point. Hence its predictions are less accurate in peaks. MSE of k-NN is relatively high because peak periods are smaller compared to off season. In contrast, our method has an explicit sinusoidal kernel combined with several squared exponential kernels to adjust for locality. Future comparisons will include recent methods such as (Davidson, Haim, and Radin 2015) and alternative filtering techniques. Predictive performance during a flu outbreak was further analyzed as this is the most vital period. Figure 7 illustrates the first peak of Figure 6 (year 2013) for two states.

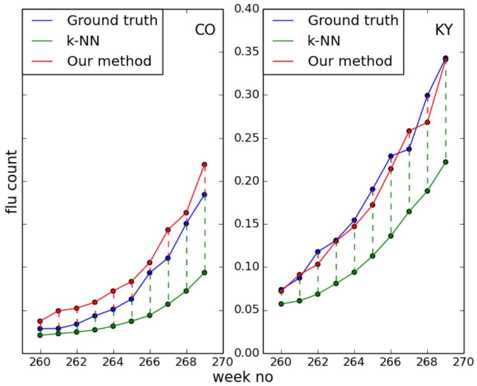

Figure 7: Experiment 5. Influenza prediction in the vital period.

In autoregressive moving average (ARMA) model and its variations such as STARMA and VARMA, a considerable amount of preprocessing has to be performed before model fitting. For instance, any trends and seasonality should be removed from data as AR models are valid only for stationary data. Flu count is clearly non-stationary and removing the seasonality is difficult as the dates of outbreaks vary from year to year, though typically occurring in winter. The most popular preprocessing technique is differencing (differentiation in discrete domain) which severally deforms the dataset. In our approach, rather than removing valuable features such as seasonality from data, different kernels were used with associated hyper-parameters. Non-linear transformations have not been applied to data to preserve the original shape.

Practical Considerations. When learning variational distribution and hyperparameters, the learning rate of natural gradient descent (NGD) should be higher than that of SGD. For instance, in our experiments empirical rates of 10^{-2} and 10^{-5} were used as the learning rates of NGD and SGD respectively. Though rates of natural gradients are less studied, it is possible to use time-varying rate or averaged SGD (ASGD) for the best and smooth convergence (Bottou 2012). Regarding natural gradients, it was observed that $\mathbf{S}$ converges several iterations after $\mathbf{m}$.
Deciding the number of inducing points and placing them appropriately is crucial for convergence time and accuracy. Naïve approaches to choose inducing points includes greedily, equidistantly and clustering such as k-means. Nevertheless, it is not essential to satisfy $\check{\mathbf{x}} \in \mathbf{x}$ as inducing points can be placed anywhere in the space. Hence, alternatively with a greater cost, it is also possible to learn the position of inducing points along with all other hyperparameters as part of SGD. Intuitively, placing some inducing points around regions with the highest rate of change, i.e. $\operatorname{argmax}_{\mathbf{x}} \frac{\partial y}{\partial \mathbf{x}}$, increases accuracy, especially when the surface is not smooth.

Conclusions

We presented a variational Gaussian process regression technique to model and predict spatial and temporal variation of influenza cases. Stochastic variational inference allowed the use of Gaussian process for tens of thousands of data points. Since the approach is non-parametric, explicit knowledge of the underlying function's complexity is not required, making it suitable for this application. To address seasonality, non-stationarity, short and long-term variations, different kernels were combined. Future work will consider more complicated dependencies, for example of weather and influenza activity. We intend to determine whether there is improvement in prediction accuracy when using meteorological variables such as point-wise humidity and temperature, by extending variational Gaussian process regression to its multi-output form. Furthermore, a Poisson likelihood can better reflect the nature of the output variable (non-negative, integer counts). Several recent works have suggested methods for incorporating a GP framework with a non-Gaussian likelihoods (Nguyen and Bonilla 2014), (Steinberg and Bonilla 2014) and these will form the basis for the algorithm's next iteration.

Acknowledgements

The authors would like to thank Dr. Aldo Saavedra, Dr. Lionel Ott and NICTA's machine learning team for insightful discussions.

References

Bottou, L. 2012. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. Springer. 421–436.

Carrat, F., and Valleron, A.-J. 1992. Epidemiologic mapping using the kriging method: application to an influenza-like epidemic in france. *American journal of epidemiology* 135(11):1293–1300.

CDC. Estimating seasonal influenza-associated deaths in the united states: Cdc study confirms variability of flu. http://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm Accessed: 2015-08-01.

Chakraborty, P.; Khadivi, P.; Lewis, B.; Mahendiran, A.; Chen, J.; Butler, P.; Nsoesie, E. O.; Mekaru, S. R.; Brownstein, J. S.; Marathe, M.; et al. 2014. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the 2014 SIAM International Conference on Data Mining. Proceedings. Society for Industrial and Applied Mathematics*, 262–270.

Charland, K.; Buckeridge, D.; Sturtevant, J.; Melton, F.; Reis, B.; Mandl, K.; and Brownstein, J. 2009. Effect of environmental factors on the spatio-temporal patterns of influenza spread. *Epidemiology and infection* 137(10):1377–1387.

Davidson, M. W.; Haim, D. A.; and Radin, J. M. 2015. Using networks to combine big data and traditional surveillance to improve influenza predictions. *Scientific reports* 5.

Dugas, A. F.; Jalalpour, M.; Gel, Y.; Levin, S.; Torcaso, F.; Igusa, T.; and Rothman, R. E. 2013. Influenza forecasting with google flu trends. *PloS one* 8(2):e56176.

Garg, S.; Singh, A.; and Ramos, F. 2012. Learning non-stationary space-time models for environmental monitoring. *Proceedings of the AAAI Conference on Artificial Intelligence* 25(35):45.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Guizilini, V. C., and Ramos, F. T. 2015. A nonparametric online model for air quality prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 651–657.

Hensman, J.; Fusi, N.; and Lawrence, N. D. 2013. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Lang, T.; Plagemann, C.; and Burgard, W. 2007. Adaptive non-stationary kernel regression for terrain modeling. In *Robotics: Science and Systems*.

Nguyen, T. V., and Bonilla, E. V. 2014. Automated variational inference for gaussian process models. In *Advances in Neural Information Processing Systems*, 1404–1412.

O'Callaghan, S. T., and Ramos, F. T. 2011. Continuous occupancy mapping with integral kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1494–1500.

Paciorek, C., and Schervish, M. 2004. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems* 16:273–280.

Plagemann, C.; Kersting, K.; and Burgard, W. 2008. Non-stationary gaussian process regression using point estimates of local smoothness. In *Machine learning and knowledge discovery in databases*. Springer. 204–219.

Quiñonero-Candela, J., and Rasmussen, C. E. 2005. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research* 6:1939–1959.

Rasmussen, C. E. 2006. Gaussian processes for machine learning.

Reid, A. 2011. *Gaussian Process Models for Analysis of Remotely Sensed Geo-Spatial Data*. Ph.D. Dissertation, Australian Centre for Field Robotics, University of Sydney.

Saavedra, A. F.; Wood, S.; Geoghegan, J. L.; Holmes, E.; and Durrant-Whyte, H. 2015. Modelling the spread of influenza in western australia. In *Proceedings of the 2015 SIG KDD Workshop on Population Informatics for Big Data*.

Steinberg, D. M., and Bonilla, E. V. 2014. Extended and unscented gaussian processes. In *Advances in Neural Information Processing Systems*, 1251–1259.

Titsias, M. K. 2009. Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 567–574.

Tresp, V. 2000. Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems*, 654–660.

Viboud, C.; Boëlle, P.-Y.; Carrat, F.; Valleron, A.-J.; and Flahault, A. 2003. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology* 158(10):996–1006.

Viboud, C.; Bjørnstad, O. N.; Smith, D. L.; Simonsen, L.; Miller, M. A.; and Grenfell, B. T. 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* 312(5772):447–451.

Waller, L. A.; Carlin, B. P.; Xia, H.; and Gelfand, A. E. 1997. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association* 92(438):607–617.

Wang, Z.; Chakraborty, P.; Mekaru, S. R.; Brownstein, J. S.; Ye, J.; and Ramakrishnan, N. 2015. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294. ACM.

WHO. 2005. Ten things you need to know about pandemic influenza. http://web.archive.org/web/20090923231756/http://www.who.int/csr/disease/influenza/pandemic10things/en/index.html Accessed: 2015-08-01.

Yang, W.; Karspeck, A.; Shaman, J.; and Ferguson, N. M. 2014. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS computational biology* 10(4):e1003583.