UK-Citadel-Team Documentation

Release 0.0.1

Harrison Zhu, Taketomo Isazawa, Benjamin Hillyard, Parley Yang

Mar 31, 2019

CONTENTS:

Please navigate our submission via the contents side bar!

CHAPTER

ONE

TEAM 18 MEMBERS:

Taketomo Isazawa, University of Cambridge, 1st year PhD in Natural Language Processing.

Harrison Zhu, Imperial College London, 4th year Undergraduate in Mathematics.

Benjamin Hillyard, University of Oxford, MSc in Statistics.

Parley Yang, University of Cambridge, MPhil in Economic Research.

1.1 How can we better allocate resources to contain influenza?

1.1.1 Introduction

It is estimated that every year, more than 291,000 people die from seasonal flu-related illnesses[1]. While there are diseases that cause more deaths, we know how to virtually completely prevent them. For example, tuberclosis claimed 1.6 million deaths in 2017[2], but only around 500 of those deaths were in the United States[3]. Diseases like tuberclosis, cholera, or measles have been effectively 'solved' in developed countries. All that we need to do is to 'port' these solutions to the developing world. It may cost money, it may take time, but we know what works.

Influenza is different. Approximately 80,000 people died in the United States in 2017 alone[4], and as recently as 2009, the World Health Organisation declared an influenza pandemic[5]. Especially with influenza's extremely fast rate of mutation, it is unlikely that we will be able to find a fundamental solution to influenza any time in the near future, making it vital that we find ways to contain the spread of influenza.

As can be seen from the constant budget crises facing the British National Health Services[6], our society seems to have priorities other than human life. It is therefore of paramount importance that we allocate the limited resources available to most effectively contain influenza and minimise its impact. We will be focussing on the developed world to ensure that the hypotheses we explore and the policies we propose go beyond the state-of-the-art.

To answer our question of how to better allocate resources to contain influenza, we must first think about what information we need to make these decisions. We believe that to make the best decisions, what we need to be able to do is

[1] https://www.cdc.gov/media/releases/2017/p1213-flu-death-estimate.html

[2] https://www.who.int/tb/publications/global_report/en/

[3] https://www.cdc.gov/tb/publications/factsheets/statistics/tbtrends.htm

[4] https://www.cdc.gov/flu/about/burden/2017-2018.htm

[5] https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

[6] https://www.theguardian.com/society/2018/may/22/hospitals-struggling-to-afford-new-equipment-after-nhs-budget-cuts

predict, ahead of time, when, and with what severity a region will be impacted by influenza. Once we know that, we can simply apply procedures that would usually be applied after we find an influenza outbreak, but apply it earlier, to contain the disease before it can spread.

1.1.2 Models

To obtain these pieces of information, we created two models, one to predict in the short term when and where influenza will strike, and another long-term model which predicts the severity of the influenza season.

The short-term model uses, to the best of our knowledge, a newly proposed Gaussian process mixture model with an XGBoost mean function, taking into account of geographical and spatiotemporal factors to identify, with high precision, when and where an outbreak will occur. During evaluation, we found that it performed well on predicting outbreaks in 2018, with an AUC of 0.762 and a false negative rate lying in a credible interval of (10.2%,13.1%). The surveillance system was also able to capture how influenza spreads spatiotemporally, as explained in the Models section.

We also have a Bayesian model that accurately predicts the severity of the next year's influenza season given the data from the current year. We find this model to also perform exceptionally well with PERFORMANCE DATA TO BE ADDED BY BENJAMIN.

1.1.3 Applications

Based off this, we can make the following policy suggestions:

1.2 Introduction

It is estimated that every year, more than 291,000 people die from seasonal flu-related illnesses[1]. While there are diseases that cause more deaths, we know how to almost entirely prevent them. For example, tuberclosis claimed 1.6 million deaths in 2017[2], but only around 500 of those deaths were in the United States[3]. Diseases like tuberclosis, cholera, or measles have been effectively 'solved' in developed countries. All that we need to do is to 'port' these solutions to the developing world. It may cost money, it may take time, but we know what works.

Influenza is different. Approximately 80,000 people died in the United States in 2017 alone[4], and as recently as 2009, the World Health Organisation declared an influenza pandemic[5]. Especially with influenza's extremely fast rate of mutation, it is unlikely that we will be able to find a fundamental solution to influenza any time in the near future, making it vital that we find ways to contain the spread of influenza.

As can be seen from the constant budget crises facing the British National Health Services[6], our society seems to have priorities other than human life. It is therefore of paramount importance that we allocate the limited resources available to most effectively contain influenza and minimise its impact. We will be focussing on the developed world to ensure that the hypotheses we explore and the policies we propose go beyond the state-of-the-art.

We first start by cleaning the data we have and adding other datasets we believe may be informative in INSERT SECTION LINK HERE. We then perform some EDA on this collected data to check its validity in INSERT SECTION LINK HERE. Models are created based on this refined data in INSERT SECTION LINK HERE, and the conclusions are presented in INSERT SECTION LINK HERE.

[1] https://www.cdc.gov/media/releases/2017/p1213-flu-death-estimate.html

[2] https://www.who.int/tb/publications/global_report/en/

[3] https://www.cdc.gov/tb/publications/factsheets/statistics/tbtrends.htm

[4] https://www.cdc.gov/flu/about/burden/2017-2018.htm

[5] https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

[6] https://www.theguardian.com/society/2018/may/22/hospitals-struggling-to-afford-new-equipment-after-nhs-budget-cuts

1.3 Datasets

While most of the data we required was provided, and already in a fairly structured form, we had to take care to process data to account for missing data, at times interpolating between data points. This section will outline the procedures undertaken to obtain/clean the data for each dataset used. Missing data in general was marked as `N/A` and the time steps where such data occurred were disregarded during modelling.

1.3.1 Influenza Data

WHO FluNet

As the project was focussed on the spread of influenza, the logical choice was to use the WHO FluNet database, provided in the `influenza_activity` dataset. As the different types of influenza have similar levels of symptoms, we decided that the field of most interest was the number of total detected influenza viruses as opposed to the data for each subtype of influenza. This also had the advantage of giving us more data to work with; many countries recorded total influenza activity while not necessarily recording data for each type of influenza.

The dataset was relatively clean and didn't require further cleaning, and any dates with missing values were ignored. To create predictive models, a large number of other datasets were also created and joined with this dataset using the `country_code` as the primary key.

The dataset was used as-is but a version was also made where the data was collected such that it was at a monthly timescale as opposed to a weekly one, due to many of the other variables having a monthly or yearly timescale.

As analysis continued, we found that this data was at times difficult to model due to its gradual increase over time. In particular, after the 2009 influenza pandemic, far more influenza samples were detected as governments around the world took the threat of another influenza outbreak increasingly seriously. To alleviate this issue, alternative sources of data were investigated.

CDC ILINet

We found the Centers for Disease Control and Prevention (CDC), a government agency of the United States of America, monitored the spread of influenza-like illnesses (ILI)[1] via ILINet[2], and this data was freely available online. As the data is restricted exclusively to the US, it was not suitable for some of our usecases, but unlike the FluNet data, did not have a constantly increasing background.

1.3.2 Physicians

A factor that we believed was likely to affect the spread of influenza was the number of physicians per unit population, which was provided in the `health_indicators` dataset. While the data was clean when provided, there were numerous periods with missing data, resulting in the need for interpolation.

With the assumption that the measurements were made in the January of every year, monthly data was created using quadratic spline interpolation. This interpolation method was chosen for a number of reasons; we wanted to ensure that the interpolation equalled the actual measurements at the points the measurements were made, that the interpolation did not overfit the data, and that the interpolation was smooth, as the data would often go up and down. Quadratic spline interpolation satisfies all of this due to it being a smooth interpolation method with very few parameters.

[1] A patient is defined to have an influenza-like illness when they have a fever of 37.8 °C or greater and a cough and/or sore throat in the absence of a known cause other than influenza. (https://gis.cdc.gov/grasp/fluview/FluViewPhase2QuickReferenceGuide.pdf)

[2] ILINet collects information on patient visits to healthcare providers for influenza-like illnesses, with data available online here

There were also cases when there were only two measurements in total, in which case quadratic spline interpolation was not sensible so linear interpolation was used instead. Finally, in the case there was only one datapoint for a country, that value was set for the January of that year, and all other values were recorded as not available.

1.3.3 Healthcare Expenditure

While it is well-known that life expectancy correlates with total healthcare expenditure[3], we also wanted to investigate its effects on controlling influenza. This was found by combining the domestic government healthcare expenditure per capita and domestic private healthcare expenditure per capita adjusted for purchasing power parity in current international dollars. This data was not interpolated as these budgets are usually set on an annual basis, and so the value for any time in each year was taken to be the value measured for that year.

1.3.4 Smoking prevalence

Although not commonly recognised as a risk factor for influenza, there have been small-scale studies that have indicated that it increases both the *risk* for contracting influenza and *severity* of such infections[4]. The data was extracted from the `health_indicators` dataset. As smoking rates were linearly decreasing with time around the world, as can be seen in the case of the United States

INSERT FIGURE HERE,

> linear interpolation was used between the given measurements to find the smoking rate at any given time.

1.3.5 Number of hours worked

So-called presenteeism, when ill workers come into work due to societal pressure and spread disease, can contribute to the spread of disease, with a study estimating that presenteeism costing the U.S. economy a staggering $150 billion a year[5]. We wanted to factor in presenteeism culture in different countries into our models; presumably, the higher the degree of presenteeism, the faster the spread of influenza. However, without expensive primary research, it is near impossible to estimate the degree of presenteeism and even then, it is not possible to extrapolate this data to the past.

Instead, we looked at the number of hours worked as a proxy for this. If there is a high degree of presenteeism, this should manifest in the number of hours that people work. This data was found for OECD countries in the form of number of hours worked per year[6]. The value was processed so that the number of hours worked was constant through the calendar year as the measurements given were in the form of hours worked per year; it didn't make sense to divide the data any further as in reality there is seasonality to the number of hours worked per month.

1.3.6 Remote Sensing Data

Influenza viruses can survive much longer at low humidity and low temperatures, partially contributing to the seasonality of flu outbreaks[7].

We obtained the coordinates of the capitals of each country and performed an SQL left join of `influenza` on the coordinates. We picked the coordinates of the capitals because these would usually indicate the regions with most of the population. We can make the following observations.

[3] https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure

[4] A study of an outbreak of A(H1N1) influenza in an Israeli military unit with 336 healthy young men found that the smokers were ~1.4x more likely to contract influenza, and ~1.6x as likely to lose work days. (https://www.nejm.org/doi/full/10.1056/NEJM198210213071702)

[5] https://www.forbes.com/sites/karenhigginbottom/2018/04/20/the-price-of-presenteeism-2/#4742f0f37f9c

[6] https://stats.oecd.org/index.aspx?DataSetCode=ANHRS

[7] http://sitn.hms.harvard.edu/flash/2014/the-reason-for-the-season-why-flu-strikes-in-winter/

- Influenza outbreaks seems to appear in clusters of regions. Especially for Europe and Central + South America. One of our goals could be to identity how the spread occurs over space and time.
- There are more outbreak reports in Europe and fewer in South America. This may be due to better surveying and medical infrastructure in Europe. Another subject of study for us would be to use the existing data to interpolate what could happen in countries where there is little or no observation, using a spatiotemporal model.

To use the dragging cursor, click on the play icon and select the second icon.

Figure link. Our previous visualisation and studies view that there is a yearly seasonality. Many recent studies have been on studying the relationship of spatiotemporal spread of influenza and diseases over a particular regional clusters. For example, Bhatt et al., 2017[10] looked at mapping disease over space-time using a GP in sub-Saharan Africa, Chen et al., 2019[9] looked at seasonal influenza spread in Shenzhen, China and Senanayake et al., 2016[12] on weekly flu occurrence in the USA.

Motivated by Bhatt et al., 2017[10], we use live satellite imagery (NOAA, TerraClimate) to obtain aggregated remote sensing data of temperature, precipitation, humidity etc… to augment our existing feature space. The data can be found from Google Earth Engine API[11] newly-developed by Google. An extraction pipeline is illustrated below.

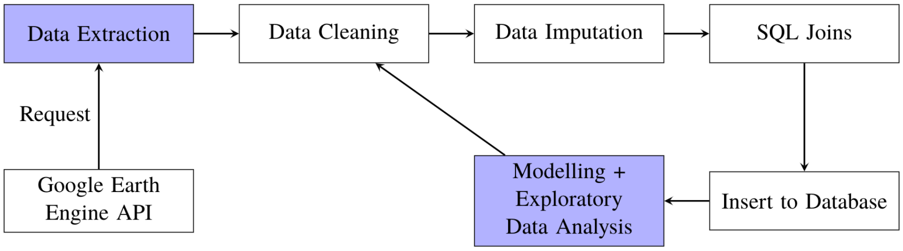


Figure 1: Data pipeline concept. * Minimum working example provided in GitHub repository and can be modified for deploying at scale.

The extraction procedure is complicated, as all the computations for extracting the final `csv` is done in the Google Cloud Server, which has specific data structures for everything, through requests using the Python API. The satellite images are stored in the `ImageCollection` data structure as a collection of images. We first obtain the `ImageCollection`, select the range of dates we are interested in and then reduce the collection to a single image by taking the mean of all the features and pixels. We then obtain the coordinates of every country through a nations `FeatureCollection`, perform a mean `reduction` over 4000x4000m squares over each country to obtain the average feature values of each country during the specified range of dates. Finally, we make a request to export to a single `csv` and save it into Google Drive.

For our study, we extract monthly and weekly remote sensing data from NOAA and TerraClimate respectively[11]. We then merge all the monthly or weekly data together and then perform SQL joins with the coordinates of the capitals of each country and the `influenza_activity.csv` dataset.

Using Lasso regularised regression and ElasticNet, we select the following features for use in spatiotemporal modelling later on:

- Capital city latitude
- Capital city longitude

[10] Bhatt, S., Cameron, E., Flaxman, S.R., Weiss, D.J., Smith, D.L. and Gething, P.W., 2017. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. Journal of The Royal Society Interface, 14(134), p.20170520.

[9] Chen, S., Xu, J., Wu, Y., Wang, X., Fang, S., Cheng, J., Liu, X. 2019. Predicting temporal propagation of seasonal influenza using improved gaussian process model. Journal of Biomedical Informatics, 93, 103144. https://doi.org/https://doi.org/10.1016/j.jbi.2019.103144

[12] Ransalu Senanayake, Simon O'Callaghan, and Fabio Ramos. 2016. Predicting spatio–temporal propagation of seasonal influenza using variational Gaussian process regression. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press 3901-3907.

[11] N.Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine:Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 2017. doi: 10.1016/j.rse.2017.06.031. URLhttps://doi.org/10.1016/j.rse.2017.06.031.

- Temperature
- Evapotranspiration, derived using a one-dimensional soil water balance model
- Surface pressure
- Surface Height
- Year
- Week

1.3.7 Google Trends

There have been a number of attempts to use Google search data to predict influenza prevalence, the most famous being Google Flu Trends[8]. We decided to scrape all available data from Google Trends at a weekly resolution going back to 2004 to add as an input to our models. Google only allows querying 5 years at a time for weekly resolution data and normalises the data within that time range such that the most number of queries in the requested time period is 100, so we had to apply a scaling factor to normalise the data, which was calculated by getting a year overlap between queries and looking at the corresponding values. Furthermore, the Google Trends API accepts geographical codes in two-letter codes as opposed to the three-letter codes provided, so a short script was written to transform between the two.

We used the query terms of 'fever' and 'cough' as indications that people have the flu. The obvious terms 'influenza' and 'flu' were omitted as they scaled more with interest in the disease from media coverage than with the actual number of people infected. A problem with this dataset was that as the number of people using Google has been steadily increasing, the search count has been constantly increasing with time as well, as can be seen in the graph below (TO BE ADDED). To get around this, WHAT CAN WE DO?

1.4 Models

Please navigate the content via the contents side bar!

1.4.1 When and where will influenza strike?

Motivation

It is of paramount importance that policy makers know ahead of time when and where influenza outbreaks will occur. Without this, it would be impossible to have optimal allocation of resources for disaster prevention and containment. We will be exploring the hypothesis that spatiotemporal and geographical factors are enough to predict influenza outbreaks. While this may seem straightforward, the study of the spatiotemporal and geographical factors and its relation to disease outbreaks is a relatively computationally expensive and novel approach that has only recently become a source of research interest[1][2].

In this section, we explore 3 classes of models to provide a model for influenza outbreaks: Gaussian process regression, Deep Gaussian processes and a Gaussian process mixture with an XGBoost mean function. To the best of our knowledge, this third model represents a completely novel approach.

[8] http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf , https://www.mitpressjournals.org/doi/full/10.1162/NECO_a_00756#.Vu5zr0eAY4A

[1] Bhatt, S., Cameron, E., Flaxman, S.R., Weiss, D.J., Smith, D.L. and Gething, P.W., 2017. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. Journal of The Royal Society Interface, 14(134), p.20170520.

[2] Chen, S., Xu, J., Wu, Y., Wang, X., Fang, S., Cheng, J., Liu, X. 2019. Predicting temporal propagation of seasonal influenza using improved gaussian process model. Journal of Biomedical Informatics, 93, 103144. https://doi.org/https://doi.org/10.1016/j.jbi.2019.103144

Note that while in theory these models could model everything, we focussed on the *when* and *where* of the outbreaks, not the *severity*. This is because the values for severity were found to be so varied that they were difficult to model well at the same time as modelling the location and time. Instead, we created a separate *Bayesian model* that specialises in predicting the severity of outbreaks.

Defining $\{x_i, y_i\}_{i=1}^N$ as our features and response (number of positive influenza cases) respectively, we assume the following underlying relationship:

$$y_i = f(x_i) + \epsilon_i,$$

where $x_i \in \mathbb{R}^p$ is the feature, $\epsilon_i \sim N(0, \sigma^2)$ and f is the underlying function. Our features x_i contain both spatial and temporal features, including latitudes, longitudes, and temperatures, as described in the datasets section. Due to this, the standard regression methods of generalised additive models (GAMs)[3], gradient boosting and regression trees[4] are not suited to this problem, and do not help us understand the underlying causality/correlation. In addition, pure time series models such as Long-short term memory (LSTM) recurrent neural networks[5], SARIMA and ARMA-GARCH models[3] are also unsuitable as they would not take the spatial variation into account. On the other hand, stochastic processes such as Gaussian processes (GP)[6] or solutions to stochastic partial differential equations (SPDE)[7] are well adapted to what we would like to accomplish.

Influenza outbreaks often contain complicated causal relationships between many different social, geographical and political factors. SPDEs are models of the form

$$Lu = f + \xi \circ dW,$$

where L is a differential operator, f is some function and $\xi \circ dW$ is driven white noise. It is perhaps the most natural approach to modelling spatiotemporal phenomena, adding a degree of noise to a partial differential equation (PDE). However, there are limited software packages that provide solutions to these SPDEs. `R-INLA`[8] is a library that uses the Bayesian method integrated nested Laplace approximation (INLA) to construct weak solutions to linear fractional SPDEs, but this places too much restriction on the underlying SPDE and would result in black-box modelling.

Gaussian Processes Review

Due to the limitations of other, perhaps more obvious models mentioned above, we will isntead use GPs as a spatiotemporal framework to study spatiotemporal variations in this study. We let f have a Gaussian process prior, giving

$$f \sim GP(\mu(\cdot), k(\cdot, \cdot)),$$

where $\mu(\cdot) := E[f(\cdot)]$ and $k(\cdot, \cdot) := E[f(\cdot), f(\cdot)]$ are our chosen mean and covariance functions. Following the convention from literature, we will be calling $k(\cdot, \cdot)$ a kernel. The mean is usually chosen to be either zero, a constant value, a polynomial function, or splines. This helps us capture the trend of y_i. However, more important than the mean is capturing the covariance between different features. It is worth mentioning that GP regression is a form of non-parametric (infinite-dimensional) regression, making it perfect for modelling complex relations like this where the features can interact in complex, unforeseen ways.

Suppose we observe some data X, y with N observations and we are given a test set X_* with M observations, and we would like to predict y_*. Using the Sherman-Morrison-Woodbury identity on the joint posterior of the GP, we obtain the posterior predictive distribution of f_* as[6]

$$f_*|X, y, f \sim N_M(K_*(K + \sigma^2 I_N)^{-1}[f + \mu], K_{**} - K_*(K + \sigma^2 I_N)^{-1} K_*^T),$$

[3] A.C. Davison. Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. CambridgeUniversity Press, 2003. doi: 10.1017/CBO9780511815850.

[4] Bishop, C.M., 2006. Pattern recognition and machine learning. Springer.

[5] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[6] Williams, C.K. and Rasmussen, C.E., 2006. Gaussian processes for machine learning (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT Press.

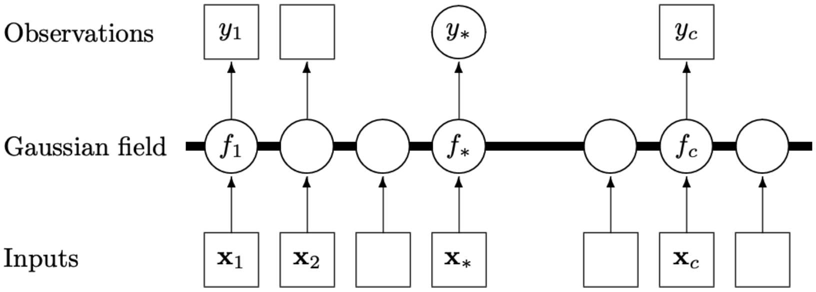
[7] Hairer, M., 2009. An introduction to stochastic PDEs. arXiv preprint arXiv:0907.4178.

[8] Lindgren, F. and Rue, H., 2015. Bayesian spatial modelling with R-INLA. Journal of Statistical Software, 63(19), pp.1-25.

where f, f_* denote the value of the function for the training and test set, K, K_*, K_{**} are the covariance matrices of the training, test-training and test sets for the GP function, and μ is short hand for the mean of the training set. In addition, using the expected square error as the loss function, one can derive the optimal prediction to be[6]

$$E[f_*|X, y, f] = K_*(K + \sigma^2 I_N)^{-1}[f + \mu].$$

Another way to understand Gaussian processes is to think of it as a graphical model with p dimensions (p types of inputs). Below is a 1-dimensional example visualisation of a Gaussian process random field[6].



While we will also deploy a Gaussian process classification model, we will omit the details as the theory is more complicated and requires variational inference. If we were to go through the derivation, however, we would find that the optimal predictor for a classification problem with the 0-1 loss is the Bayes' classifier.

Model 1: Gaussian process

To treat the seasonal effects, we choose a kernel

$$k(t', t) = \exp\left(\frac{2\sin^2(\pi||t - t'||_1 f)}{l^2}\right)$$

for years t, t', where f and l' are the kernel frequency and length scale respectively. We encode a prior distribution for the frequency to favour the value 1, as we believe that influenza outbreak occurs annually during winter.

Our exploratory data analysis indicated a relatively smooth trend for the weekly effects, justifying the use of a radial basis kernel:

$$k(x', x) = \exp\left(- \frac{(x_1 - x_2)^T(x_1 - x_2)}{l}\right),$$

where l is the length scale. Our claim is supported by the theory of reproducing Hilbert spaces[9], since if the underlying functional relationship of the weekly effect is sufficiently regular (Holder-Sobolev of certain exponents), then a GP will provide a good estimate.

For the spatial and remote sensing features, we use Matérn covariance kernels. This is defined as

$$k(x', x) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}d)K_\nu(\sqrt{2\nu}d),$$

where K_ν is the modified Bessel function, v' the smoothness parameter, and d is defined to be $||x_1 - x_2||_\Theta$, where Θ is a lengthscale parameter in matrix form.

Although more complicated kernels were tried, we found that the most straightforward kernel, consisting of a sum of all four kernels, was the most effective.

We will use by a zero mean by default for simplicity.

[9] http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf

Model 2: Deep Gaussian processes

Suppose now that we have a latent feature extractor. We perform GP classification with variational inference to approximate the posterior and marginal likelihood, and use 3 layers of linear regressor-ReLU as the feature extractor. This forms a Deep Gaussian process with linear layers in between.

Model 3: Gaussian process mixture with XGBoost mean function

To account for the extreme values that occur during outbreaks, we construct what is, to the best of our knowledge, a new type of GP mixture model by replacing the mean function with a pre-trained XGBoost regressor. Through this transfer learning procedure, we are able to provide uncertainty quantification for the previously purely black-box XGBoost model and augment the mean function of the GP with a more sophisticated feature regressor. We could also understand the replacement as encoding our prior belief of the true underlying function f. Finally, this also allows us to understand the spatiotemporal and climatic relationship in our data.

Experimental Results

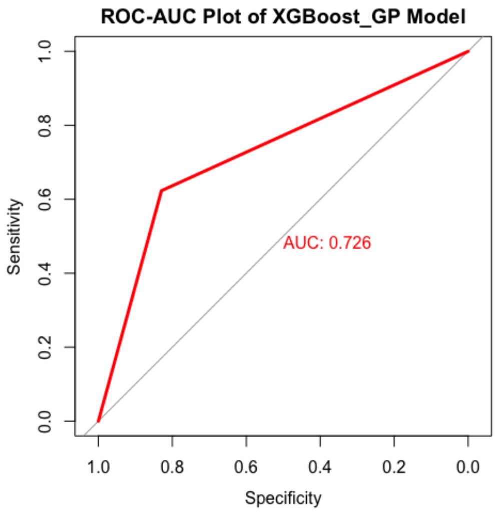
To conduct prediction, we first learn the underlying function f and then obtain a prediction of the number of influenza cases. It is clear from the nature of the data that outbreaks are often extreme, and therefore without extreme value or SPDE theory it unfeasible to make predictions of the peaks with Gaussian processes. However, the GP is very good at capturing the trend, and therefore we take 5% of the maximum value of the number of influenza cases for each country as the threshold for classifying an outbreak there respectively.

We conducted hyperparameter tuning and training using the PyTorch framework on the Imperial College GPU Cluster with two 31GB RAM Tesla K40c GPUs on Ubuntu 16.04.5. In particular, we made heavy use of the `gpytorch`[10] library to model the GPs. We found that our newly proposed model was most suitable for policy-making purposes, as it provides accurate predictions with quantifiable uncertainties. The pure Gaussian process model was good at estimating the trend but performed poorly when looking at the magnitude. The Deep Gaussian process similarly had the same issue, which justified the use of transfer learning with the XGBoost prior function. The below figure illustrates an optimal prediction of whether there is an outbreak or not in space-time. The dataset is explained in the datasets section.

Note: To use the dragging cursor, click on the play icon and select the second icon.

Figure source. The Gaussian Process mixture with XGBoost mean function results in an AUC (area under curve) on the ROC(Receiver Operating Characteristic) curve of 0.762, as shown below. From a policy perspective, within reasonable bounds, the proportion of false negatives is more important, as a false positive will only strengthen the prevention of an outbreak. We find that out of 382 test points in 2018, we have a 95% credible interval of (10.2%,13.1%) on the percentage of false negatives, with the optimal prediction yielding 11.8%.

[10] https://gpytorch.readthedocs.io/en/latest/index.html



We also observe exactly what we hypothesised - the spread of influenza in space - in the above diagram. We can see that when an outbreak is observed in 1 country, it spreads very rapidly to neighbouring countries (especially visible during the transition from week 49 to 50).

Shortcomings

As mentioned in the analysis, we have mainly focused on predicting the occurrence of outbreaks, rather than the exact number of cases. To predict the latter, there have been multiple recent studies on stochastic partial differential equations and INLA[8]. Our team decided to implement a Bayesian model to do this, which is detailed in *the next section*.

Finally, there is also an existing framework for extreme value statistics that would be a more suitable model for predicting either the extreme events or looking at the probability of threshold exceedances. Moreover, the current remote sensing data focusses on capital cities, while a finer grain data source would improve the quality of the fit.

1.4.2 How big will an influenza epidemic be?

Motivation

Before we start modelling the severity of influenza seasons, we investigated the state-of-the-art for this field. As the leading national public health institute in the United States and one of the world's premier infectious disease surveillance bodies, the Centre for Disease Control (CDC) represents the status quo in influenza modelling. The CDC currently deploy an adaptation of Serfling's method[1], which uses cyclic regression to model the weekly proportion of deaths from pneumonia and influenza. Adaptations to the basic model have incorporated indicators such as counts of patient visits for influenza like illness (ILI)[2][3]. However, regardless of modern modifications, the methodology is limited by its unfounded assumption that observations are independent and identically distributed.

In this contribution, we attempt to shift the methodology towards the Bayesian framework in order to provide better epidemic thresholds that are adjusted for seasonal effects. In doing so, we build prior and observation models for the number of individuals infected by influenza within a specific region. After building the models, we simulate from the prior to test the model's performance and ensure the prior model is sufficiently grounded in reality to produce justified

[1] Robert E. Serfling. (1963). Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. Public Health Reports (1896-1970), 78(6), 494-506. doi:10.2307/4591848

[2] L.Simenson, K. Fukuda, L. B. Schonberg, and N. J. Cox. The impact of influenza epidemics on hospitalizations. The Journal of Infectious Diseases, 181:831–837, 2000.

[3] F.C. Tsui, M. M. Wagner, V. Dato, and C. C. H. Chang. Value ICD-9-Coded chief complaints for detection of epidemics. In Proceedings of the Annual AMIA Fall Symposium, 2001.

posterior inference. Once we are satisfied with the model we deploy Approximate Bayesian Computation (ABC) to generate approximate posterior samples and proceed to make probabilistic statements to inform policy makers.

Prior Elicitation

We begin by outlining the prior and observation models in this setting. Whilst the systematic use of parametrised distributions is not always justifiable, when building the prior we arbitrarily restrict ourselves to a parametrised density where we can make subjective evaluations of the parameters in line with our knowledge of the world.

Prior Model

We build a model for the number of individuals infected by influenza in a given week for a two year period. Whilst the model is agnostic to geographical location, we focus on specifying the prior distribution in line with European influenza cycles. The parameters are $\Theta = (X_{1:104}, \mu, \theta, \alpha, \rho, \ell)$. The notation is defined as follows: $(X_{1:104}) := (X_1, \ldots, X_{104})$, where X_i is the number of positive influenza samples recorded in week i, and the others are parameters for the model, as defined below. Using these parameters, we model the weekly flu process $(X_i)_{i=1}^{104}$ over a 2 year period (for example 2018 to 2019) as a weekly mean with an autoregressive (AR) process. By considering the seasonality of infection count we use a single AR process for each winter, since each winter has a number of hidden variables. For example, the strands of influenza active, health care spending, temperature and so on differ depending on the year. By only considering short time periods, we lower the risk of being affected by the change of these hidden variables.

Putting this all together, we have $X_t|X_{t-1}, \phi = m_t + y_t$ with $y_t \sim AR(\rho, 50)$ and $\phi = (\mu, \theta, \rho, \ell, \alpha)$. The mean in week t is given by

$$m_t = \mu + \theta t + \alpha sin^8\Big(\frac{\pi}{52}t - \ell\pi\Big).$$

That is to say, the weekly mean is a baseline infection count μ, with θt to describe the secular trend, and a suitably scaled and lagged sine function to capture seasonality.

The prior $\pi(\Theta)$ is composed of the following:

$$\begin{aligned}
X_1|\phi &\sim \mathcal{N}\Big(m_1, \frac{50^2}{1-\rho^2}\Big) & \mu &\sim \text{Unif}(0, 1000)\\
X_{27}|\phi &\sim \mathcal{N}\Big(m_1, \frac{50^2}{1-\rho^2}\Big) & \theta &\sim \text{Unif}(0, 0.5)\\
X_{79}|\phi &\sim \mathcal{N}\Big(m_1, \frac{50^2}{1-\rho^2}\Big) & \rho &\sim \text{Unif}(0.6, 0.9)\\
X_t|X_{t-1}, \phi &\sim \mathcal{N}\Big(m_t + \rho(X_{t-1} - m_{t-1}), 50^2\Big) & \ell &\sim \text{Unif}(0.7, 1)\\
& & \alpha &\sim \text{Unif}(3000, 25000)
\end{aligned}$$

for $t \in \{2, ..., 26, 28, ..., 78, 80, ..., 104\}$.

Given this prior model we have the following decomposition:

$$\begin{aligned}
\pi(\Theta) &= \pi(X_{1:104}|\phi)\pi(\phi)\\
&= \pi(X_{104}|X_{1:103}, \phi)\pi(X_{1:103}|\phi)\pi(\phi)\\
&= \pi(X_{104}|X_{103}, \phi)\pi(X_{1:103}|\phi)\pi(\alpha)\pi(\rho)\pi(\ell)\pi(\theta)\pi(\mu)\\
&= \Bigg[\prod_{i=2}^{26}\pi(X_i|X_{i-1}, \phi)\Bigg]\pi(X_1|\phi)\Bigg[\prod_{i=28}^{78}\pi(X_i|X_{i-1}, \phi)\Bigg]\pi(X_{27}|\phi)\cdot\\
&\times\Bigg[\prod_{i=80}^{104}\pi(X_i|X_{i-1}, \phi)\Bigg]\pi(X_{79}|\phi)\pi(\alpha)\pi(\rho)\pi(\ell)\pi(\theta)\pi(\mu)
\end{aligned}$$

Observation Model

We model a two year period above and make predictions for the second year given the observations in the first. The data from the first year involves some noise due to poor data collection and miss-classification of illnesses, that is to say that the observed data Y_i has the form

$$Y_{1:52} = X_{1:52} + (\epsilon_i)_{i=1}^{52},$$

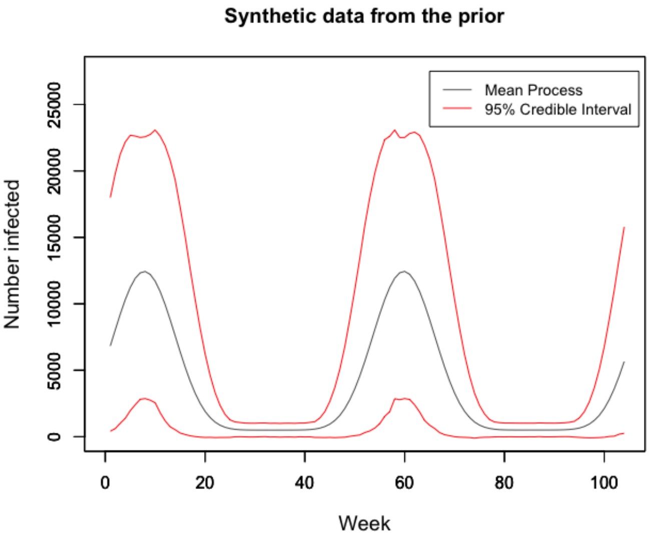
where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Thus we have

$$\pi(X_{1:52}|\Theta) = \prod_{i=1}^{52} \mathcal{N}(Y_i, 1).$$

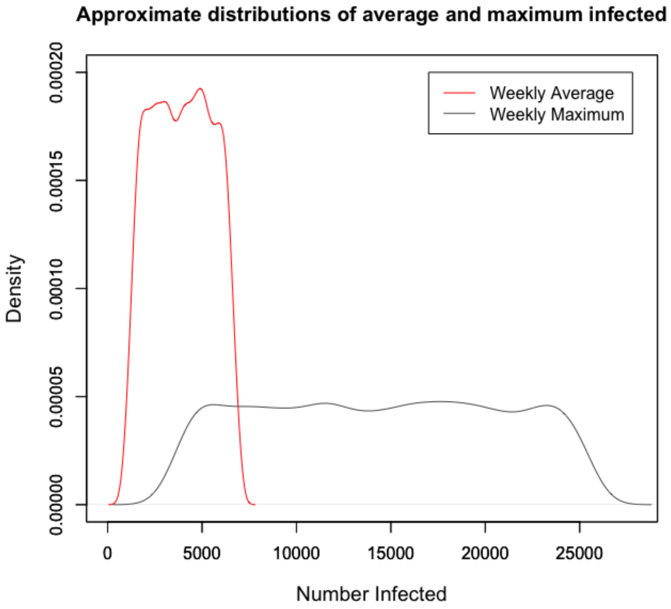
Simulating our prior model

As our prior model is a reductive representation of a complex random phenomena, it is vital to evaluate the model for likeness to the real world to ensure our posterior inference is justified.

We first consider 100,000 samples from the prior model in the below figure. This graph demonstrates likeness to real observed data for Europe over the past 5 years. Additionally, the credible intervals plotted show a sufficiently large range of realisations. The mean weekly flu count is 3934 (Credible Interval: 1313, 6629) which provides a reasonable fit to the reality of the weekly average of 4611 patients in 2018 in Europe.



It is important to scrutinise the prior for informativeness with respect to the quantities we are particularly interested in. In the below figure, the approximate distribution of average and maximum counts for 100,000 samples are given. Both are satisfactory since they fall roughly uniform across wide intervals. The weekly average of 4611 in 2018 falls in the range of high density for the average, and the European 2018 maximum of 19,074 patients infected also sits in the high density region of the approximate maximum. Both distributions reflect reality well and do not over-inform.



A Quick Remark

When choosing a prior it is important to consider alternatives. In this project a range of distributions for each of the parameters $(\alpha, \rho, \ell, \mu, \theta)$ were considered in order to represent different states of knowledge. We verified that the results of our analysis were not sensitive to this range of priors. For example in our choice of μ, which provides the base-level for the weekly mean m_t, we considered variants of uniform, normal and triangle distributions, including $\mathcal{N}(10000, 3), \text{Unif}(3000, 25000)$ and $\text{Tri}(3000, 25000, 10000)$. We observed reasonable similarity between the distributions and ultimately decided to work with the uniform distribution since it best represented our prior beliefs.

Model Choice

We are interested in understanding whether or not our current model, $\mathcal{M}_1$, is adequate. In doing so, we compare its performance with alternative models whose difference with our current model is the sine function raised to a high power. That is, for alternative models $\mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$ and $\mathcal{M}_6$, we alter the weekly mean number of influenza positive virus as:

$$\begin{aligned}
\mathcal{M}_2 &: m_t = \mu + \theta t + \alpha \sin^{10}\left(\frac{\pi}{52}t - \ell\pi\right) \\
\mathcal{M}_3 &: m_t = \mu + \theta t + \alpha \sin^{12}\left(\frac{\pi}{52}t - \ell\pi\right) \\
\mathcal{M}_4 &: m_t = \mu + \theta t + \alpha \sin^{16}\left(\frac{\pi}{52}t - \ell\pi\right) \\
\mathcal{M}_5 &: m_t = \mu + \theta t + \alpha \sin^{20}\left(\frac{\pi}{52}t - \ell\pi\right) \\
\mathcal{M}_6 &: m_t = \mu + \theta t + \alpha \sin^{30}\left(\frac{\pi}{52}t - \ell\pi\right).
\end{aligned}$$

Note that all models considered are even powers of sine as we know the weekly mean number of influenza positive virus to be always positive. Here a finite number of model comparisons is made. If one wants to consider an infinite number of models a more delicate construction of the unconditional probabilities $(p_i : i \in \mathbb{N})$ is required (for example adhering to notions of coherence). Assuming an equal prior weighting, we progress to consider Bayes factors.

Bayes factors depend on estimates of the marginal likelihood for the observation in question, that is, the first year falling in line with recorded data. We make use of the following consistent estimator:

```
Algorithm 1 Naive Approximation
procedure
    for $t \in 1, ..., n$ do $\Theta^t \sim \pi(\Theta|\mathcal{M}_k)$
Calculate:
    $\hat{p} = n^{-1} \sum_{i=1}^{n} \pi(Y_{1:52}|\Theta^i, \mathcal{M}_k)$
```

When implemented using $n = 100,000$ the approximation produced unstable results despite efforts to reduce computational underflow. To assess the evidence for accepting $\mathcal{M}_k$, $k \neq 1$, over $\mathcal{M}_1$ we compute the Bayes factor for the best performing of $\mathcal{M}_2, ...\mathcal{M}_6$ against $\mathcal{M}_1$. In 10 runs we realised a range of $(0.004, 12.656)$ with the Naive approximation. However, the particular $\mathcal{M}_k$ with the best performance was consistently $\mathcal{M}_1$, so we proceed with $\mathcal{M}_1$.

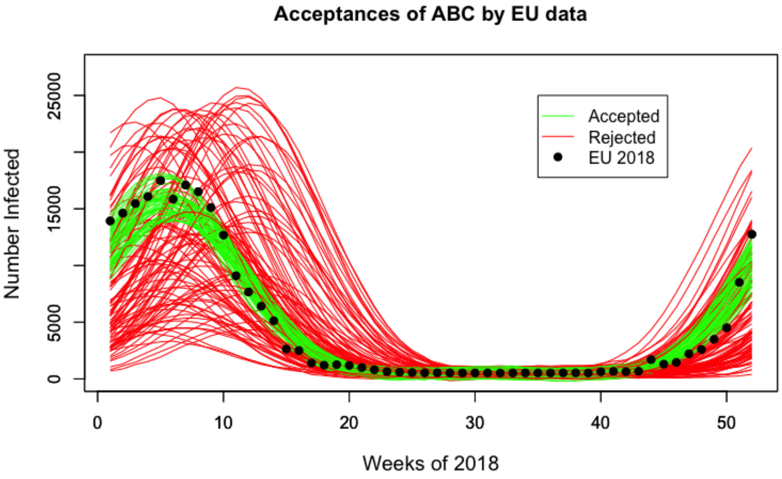
Posterior sampling

Now that we are confident with the prior model, we proceed to generate approximate samples of the posterior distribution given observed European data. Whilst it would be possible to generate true posterior samples, for example by using Metropolis-Hastings and assessing the quality of fit with ACFs, trace plots, and checking that marginal distributions agree, we instead deploy ABC to generate approximate uncorrelated samples.

Approximate Bayesian Computation

With the aim to make probabilistic statements about 2019 we deploy approximate Bayesian computation to target the posterior. In doing so, we generate samples from $\pi(\Theta|Y_{1:52})$ where $Y_{1:52}$ are given in `influenza_activity.csv`.

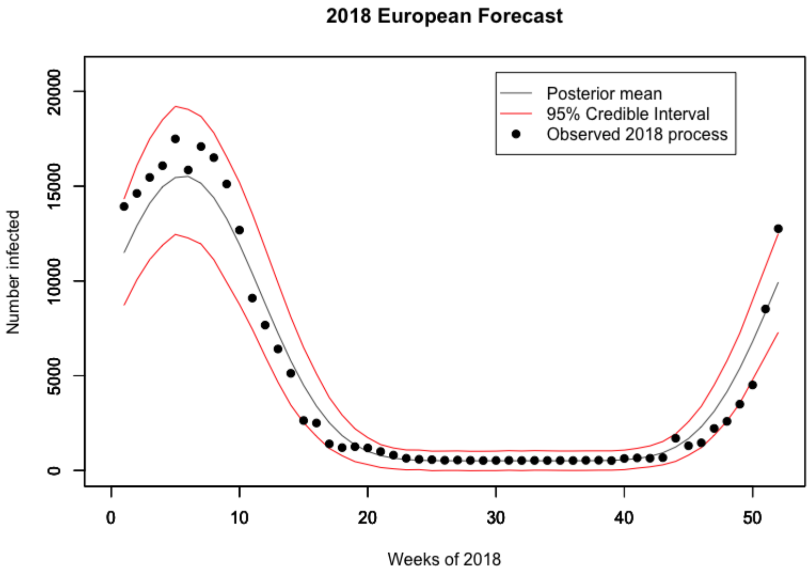
Below we observe the first year of some synthetic data, with samples accepted by ABC in green. These samples provide a satisfactory fit to the observed process.



Results

Using the posterior distribution we can inform policy makers about the probable magnitudes of the outbreaks, allowing for improved emergency planning and resource allocation. This methodology also provides an opportunity to look at the posterior for different regions of a country. Medical professionals can then strategically allocate their resources to regions with a higher probability of outbreak.

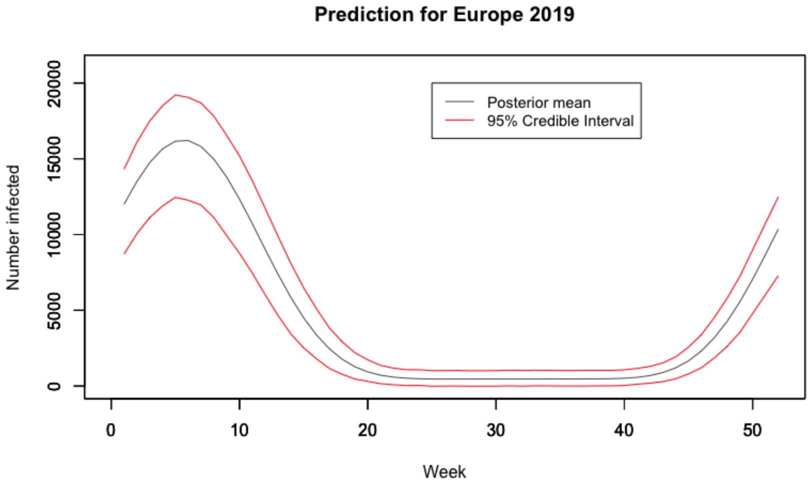
To demonstrate its usefulness, we consider 2018 model predictions given 2017 cycle observations. While we observe the peaks to be consistently above the mean prediction, we find that the observations for 2018 consistently fell within our tight Highest Posterior Density (HPD) interval. Given that the European Centre for Disease Prevention and Control recognised 2018 as a reasonably large season, we are encouraged by the fact the observations still fell within our bounds[4].



These results demonstrate our model's usefulness; with a traditional model based off Serfling's method, we could not have estimated the amount of extra resources required to manage large seasons such as this one, but our model gives a credible interval which can even account for these. We envision that policy makers could use our model to make sure that they are well prepared for large events whilst also ensuring that they do not overbudget.

Using 2018 observations for 2019 predictions, we observe an expected maximum number of viruses testing positive for influenza at 14,487 with a 95% credible interval of (3882,24675) in the prior. This expected maximum shifts to 19,413 in the posterior with a 95% credible interval at (14507,20085), putting the 2019 flu season on track to be about as large if not larger than the 2018 one. Below we also produce the expected flu cycle for 2019 with 95% HPD intervals.

[4] https://ecdc.europa.eu/en/seasonal-influenza/season-2017-18



We also believe that this could provide an alternative epidemic threshold to that currently used by the Centre for Disease Control; if we were to find that the number of infections lies outside the HPD region, this would be an indication that we are failing to control the outbreaks and on the verge of an epidemic, and suitable measures should be taken.

Shortcomings

Whilst we achieved success in developing a model that reframed and extended the existing approach, there are a few shortcomings to be mentioned. Firstly, it is generally difficult to assess whether arbitrary features of the prior predominate our posterior analysis. The question of robustness has been tackled in the literature and we could extend our models by considering the prior belonging to a class of distributions as proposed by Berger's classification[5]. Attempts could then be made to derive bounds on posterior quantities and hence produce analysis that is less sensitive to the choice of prior.

Beyond criticism of the arbitrariness and importance of the prior, we must also consider the use of ABC. The applications of ABC are often based on improved versions of the basic rejection scheme[6], and have already yielded valuable insights into questions concerning the rate of spread of pathogens[7],[8], although we go beyond past applications that have typically focused on parameter estimation rather than posterior prediction. In our case, ABC provides the benefit of independent samples. However, true posterior samples could be found by the implementation of Hamiltonian Monte Carlo[9].

Finally, the Naive approximation of Bayes factors in this setting proved unstable. Future work could focus on deploying more stable estimators for the marginal likelihood, such as a Harmonic approximation.

1.4.3 How long will the epidemic last?

Introduction

Our other two models tell us when and where the epidemics will strike, and how hard. The final factor we need to better allocate resources is to know how long such an episode will last. We hypothesise that we may be able to explain

[5] (Berger's (1990a))

[6] Beaumont, M.A. et al. (2002) Approximate Bayesian Computation in population genetics. Genetics 162, 2025–2035

[7] Tanaka, M. et al. (2006) Estimating tuberculosis transmission parameters from genotype data using approximate Bayesian computation. Genetics 173, 1511–1520

[8] Shriner, D. et al. (2006) Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution 60, 1165–1176

[9] https://arxiv.org/abs/1701.02434

which regions have long-lasting influenza seasons by looking at variables that we haven't investigated yet that may increase the 'risk factor' of a nation to staying in a state of high influenza infection rates.

We first collect a number of variables, both social and physical, as described in *the datasets section*. We then construct a simple elastic net model[1] , i.e. we minimise the objective function

$$\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

to perform variable selection. This model will have a set of coefficients that represent the way that each variable correlates with the output. To ensure that we could properly see which variables were most significant, we normalise all the input features to be between 0 and 1. We train the model on developed countries to control for underdeveloped countries having other unaccounted factors that will skew our results. We observe the following coefficients as a result of running this process:

```
Minimum temperature :  -35.75207412089624
Maximum temperature :  -27.572710323493112
Precipitation :  -1.1936198807223342
Climate water deficit :  0.9117976671368961
Actual evapotranspiration :  -25.67468313304453
Downward surface shortwave radiation :  -15.056164498685526
Vapour pressure :  -8.59360337613683
Hours worked per year :  48.18539588156608
Total healthcare expenditure per capita, adjusted for PPP :  21.214423216995893
Number of physicians per capita :  -36.074258171081844
```

As we know and expect, the higher the temperature, the fewer the positive instances of influenza[2]. What is surprising is the high positive coefficient for the hours worked per year. While we would expect for there to be some type of

Theory

Results

Performance

Application to 2018

Limitations

[1] https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.124.4696

[2] https://jvi.asm.org/content/88/14/7692s