Multi-Task Learning Improves Disease Models from Web Search

Bin Zou
Department of Computer Science
University College London
United Kingdom
bin.zou.14@ucl.ac.uk

Vasileios Lampos
Department of Computer Science
University College London
United Kingdom
v.lampos@ucl.ac.uk

Ingemar Cox*
Department of Computer Science
University College London
United Kingdom
i.cox@ucl.ac.uk

ABSTRACT

We investigate the utility of multi-task learning to disease surveillance using Web search data. Our motivation is two-fold. Firstly, we assess whether concurrently training models for various geographies — inside a country or across different countries — can improve accuracy. We also test the ability of such models to assist health systems that are producing sporadic disease surveillance reports that reduce the quantity of available training data. We explore both linear and nonlinear models, specifically a multi-task expansion of elastic net and a multi-task Gaussian Process, and compare them to their respective single task formulations. We use influenza-like illness as a case study and conduct experiments on the United States (US) as well as England, where both health and Google search data were obtained. Our empirical results indicate that multi-task learning improves regional as well as national models for the US. The percentage of improvement on mean absolute error increases up to 14.8% as the historical training data is reduced from 5 to 1 year(s), illustrating that accurate models can be obtained, even by training on relatively short time intervals. Furthermore, in simulated scenarios, where only a few health reports (training data) are available, we show that multi-task learning helps to maintain a stable performance across all the affected locations. Finally, we present results from a cross-country experiment, where data from the US improves the estimates for England. As the historical training data for England is reduced, the benefits of multi-task learning increase, reducing mean absolute error by up to 40%.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Applied computing** → **Health informatics**; • **Computing methodologies** → **Supervised learning by regression**; **Multi-task learning**; • **Theory of computation** → **Gaussian processes**;

KEYWORDS

Web Search; User-Generated Content; Disease Surveillance; Multi-Task Learning; Regularized Regression; Gaussian Processes

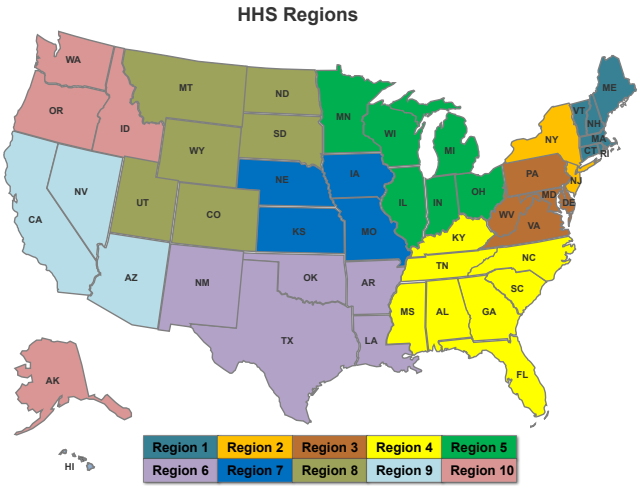

Figure 1: The 10 US regions as specified by the Department of Health & Human Services (HHS).

1 INTRODUCTION

Online user-generated content contains a significant amount of information about the *offline* behavior or state of users. For the past decade, user-generated content has been used in a variety of scientific areas, ranging from the social sciences [5, 21, 25] to psychology [26, 39, 46] and health [14, 22, 29]. Focusing on the health aspect, user-generated content has the advantage of being a real-time and inexpensive resource, covering parts of the population that may not be accessible to established healthcare systems. Thus, it can facilitate novel approaches that may offer complementary insights to traditional disease surveillance schemes.

Existing algorithms for disease surveillance from user-generated content are predominantly based on supervised learning paradigms [17, 22, 31, 42]. These frameworks propose single task learning solutions that do not consider the correlations of data across different geographies. They are also not accounting for situations, where significantly fewer health reports are available for training a model. In this paper, we investigate the utility of multi-task learning to exploit these correlations to both improve overall performance and to compensate for a lack of training data in one or more geographic locations.

Multi-task learning can train a number of disease models jointly. Compared to single task learning, it has the potential to improve the generalization of a model by taking advantage of shared structures in the data. Previous work has shown that this may result in significant performance gains [2, 4, 6, 8, 13, 20, 32]. In the context

*Also with Department of Computer Science, University of Copenhagen, Denmark.

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186050

of disease modeling, we investigate whether it can provide an improved estimate of disease rates when (a) training data is available for multiple geographic locations, specifically geographic regions of the United States (US), and (b) when ground truth training data (health reports) is sporadic. In addition, we investigate its utility in estimating disease rates in a different country by exploiting a denser health reporting scheme of a reference country. We explore both linear and nonlinear regression models, namely multi-task elastic net [35] and multi-task Gaussian Processes [11], comparing them to their respective single task formulations.

We use influenza-like illness (ILI) as a case study and conduct experiments, where ILI rates are estimated, on the US (nationally and regionally) and England. Our experiments show that multi-task learning models improve regional as well as national ILI rate estimates from Google search data for the US. The percentage of improvement increases up to 14.8%, in terms of mean absolute error, as the historical training data is reduced, indicating that multi-task learning can facilitate the derivation of accurate models using significantly less training data. We also simulate situations, where partial ground truth data are available, perhaps due to unexpected reasons (natural disasters, a spreading epidemic, technical problems) or due to limitations of a public health system. Our experimental results indicate that multi-task learning models can mitigate such effects. Finally, we apply multi-task learning to a cross-country setting, where complete data for one country could improve the models of another country with insufficient health reports. In that case, it is shown to improve ILI rate estimates for England (up to 40% of mean absolute error decrease) under the assumption that increasingly limited historical data exist, when training models jointly with data from the US.

Here is a summary of the main contributions of the paper:

(1) This is the first work to assess the utility of multi-task learning in infectious disease surveillance from Web search data.
(2) We use ILI as a case study and show that multi-task learning models improve:
   (a) regional as well as national disease models for the US,
   (b) regional US disease models, under the assumption of increasingly limited historical health reports (simulated by applying three different sampling methods), and
   (c) country-level disease models for England, when training is performed jointly with data from a different, but culturally similar, country (the US).

2 METHODS

We first provide a description for the disease modeling task, under both single and multi-task learning settings. Then, we present the linear and nonlinear techniques for performing single and multi-task regression used in our experiments.

2.1 Task Description

Our aim is to infer disease rates as reported by an established health surveillance system using the frequencies of Web search queries. We formulate this as a regression task, where we learn a function $f\colon \mathbf{X} \rightarrow \mathbf{y}$ that maps the input space $\mathbf{X} \in \mathbb{R}^{n\times p}$ to the target variable $\mathbf{y} \in \mathbb{R}^n$; n denotes the number of samples and p is the size of our feature space, i.e. the number of unique search queries we consider. $\mathbf{X}$ contains time series of normalized frequencies of search queries and $\mathbf{y}$ represents the disease rates at the same time points as reported by the health agency. A normalized query frequency is defined as the count of a query divided by the total number of searches during a fixed time interval, e.g. one week.

In multi-task disease rate inference, we are modeling disease rates simultaneously for a number of different geographical locations (tasks). A tensor $\mathbf{Q} \in \mathbb{R}^{n\times p\times m}$ is used to represent our input data for the m tasks.[1] $\mathbf{Q}$ can simply be interpreted as m versions of $\mathbf{X}$; in the remainder of the script, we denote them using $\mathbf{Q}_j$, where j refers to the j^{th} task or geographical location. An element of $\mathbf{Q}$, $\mathbf{Q}_{tij}$, represents the normalized frequency of a query i for the location j during the time interval t. The corresponding target variables, i.e. the disease rates for the m locations are denoted by $\mathbf{Y} \in \mathbb{R}^{n\times m}$. Similarly, we use $\mathbf{Y}_j$ to refer to the disease rates at the location j. Based on the aforementioned formulations, our task now becomes to learn a function f, such that $f\colon \mathbf{Q} \rightarrow \mathbf{Y}$.

2.2 Linear Regularized Regression

Linear regressors have been successfully applied for conducting disease surveillance from web search and social media data [17, 22, 28–30]. We use *elastic net* [56] to train a linear regression model. It can be seen as an extension of the ℓ_1-norm regularization, known as *lasso* [48], that incorporates an ℓ_2-norm, or *ridge* [24], regularizer on the inferred weight vector. Elastic net encourages sparse solutions, thereby performing feature selection. At the same time, it addresses model consistency problems that arise when collinear predictors exist in the input space [23].

Elastic Net (EN). Given the input matrix $\mathbf{X}$ and the observations $\mathbf{y}$, linear regression has the form of $\mathbf{y} = \mathbf{X}\mathbf{w}+\beta$, where β is an intercept term and $\mathbf{w} \in \mathbb{R}^p$ is a weight vector. Elastic net [56] estimates $\mathbf{w}$ and β by minimizing

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left(\|\mathbf{y} - \beta - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \|\mathbf{w}\|_1 \right), \qquad (1)$$

where λ_1 and λ_2 are the regularization parameters, and $\|\cdot\|_1$, $\|\cdot\|_2$ denote the ℓ_1-norm and ℓ_2-norm, respectively.

Multi-Task Elastic Net (MTEN). We extend the standard elastic net model to a multi-task version [53]. It is specified by the following optimization task

$$\underset{\mathbf{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\mathbf{Y} - \boldsymbol{\beta} - \mathbf{Q}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{W}\|_{2,1} + \lambda_2 \|\mathbf{W}\|_F^2 \right), \qquad (2)$$

where $\mathbf{W} \in \mathbb{R}^{p\times m}$, $\boldsymbol{\beta} \in \mathbb{R}^m$ are the weight matrix and intercept vector for all the m tasks, and the norms — $\ell_{2,1}$ and Frobenius (F) — are given by

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{m} W_{ij}^2} \quad \text{and} \qquad (3)$$

$$\|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{m} W_{ij}^2}. \qquad (4)$$

[1] Note that the number of samples n may be different for different locations (tasks).

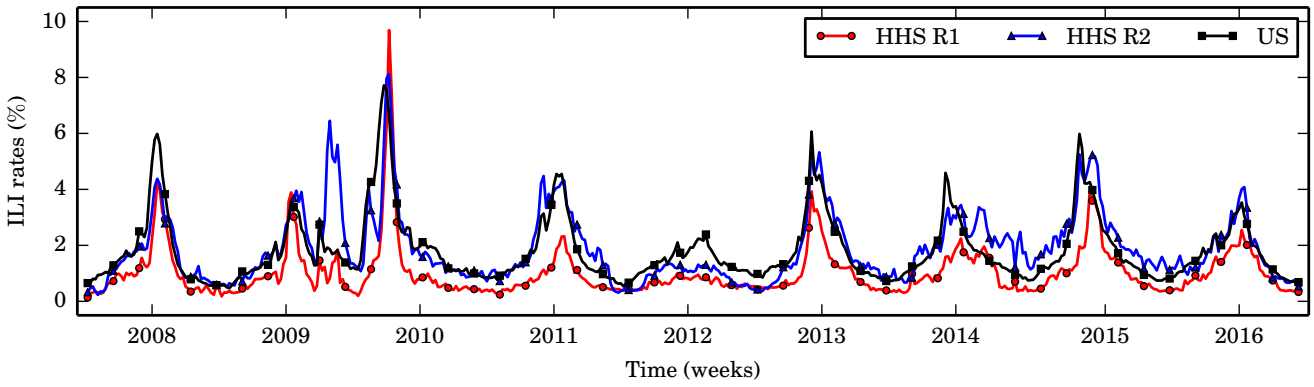

Figure 2: Weekly ILI rates (from CDC) for the US (national level) as well as the US Regions 1 and 2.

2.3 Nonlinear Regression

We also deploy nonlinear regression models using Gaussian Processes as previous works have shown that the relationship between query frequencies and disease rates is significantly better captured by a nonlinear function [31, 33, 34, 50].

Gaussian Process (GP). GP models [45] assume that the function $f\colon \mathbf{X} \rightarrow \mathbf{y}$ is a probability distribution over functions denoted as

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))\,, \tag{5}$$

where $\mathbf{x}, \mathbf{x}'$ are rows of the input matrix $\mathbf{X}$, $\mu(\mathbf{x})$ is the mean function of the process, and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function (or kernel) that captures a relationship between input observations. We assume that $\mu(\mathbf{x}) = 0$, and use the Squared Exponential kernel plus noise as our covariance function. It is defined by

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right) + \sigma_n^2 \cdot \delta(\mathbf{x}, \mathbf{x}')\,, \tag{6}$$

where ℓ is the length-scale parameter, δ is a Kronecker delta function, and σ^2, σ_n^2 are scaling constants that represent the overall variance. In GPs, predictions $(\mathbf{y}_*)$ can be made by using the conditional distribution $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) \sim \mathcal{N}(\mu_*, \sigma_*^2)$, where $\mathbf{x}_*$ denotes a new observation. Following the assumption that $\mu(\mathbf{x}) = 0$, μ_* and σ_*^2 are given by

$$\mu_* = \mathbf{K}(\mathbf{x}_*, \mathbf{X})^\top \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{y}\,, \text{ and} \tag{7}$$

$$\sigma_*^2 = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}(\mathbf{x}_*, \mathbf{X})^\top \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{K}(\mathbf{X}, \mathbf{x}_*)\,, \tag{8}$$

where $\mathbf{K}$ is a covariance matrix derived by applying Eq. 6 elementwise. The hyperparameters of the GP model, $\boldsymbol{\theta} = \{\sigma, \ell, \sigma_n\}$, are learned by minimizing the negative log-marginal likelihood [45], given by

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(-\frac{1}{2}\mathbf{y}_j^\top (\mathbf{K}(\mathbf{X}, \mathbf{X}))^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}(\mathbf{X}, \mathbf{X})| - \frac{n}{2}\log 2\pi\right). \tag{9}$$

Multi-Task Gaussian Process (MTGP). GPs models were extended to a multi-task version (MTGP) by Bonilla et al. [11] and have been used in various tasks, including natural language processing applications [7, 15]. The MTGP model incorporates all m tasks into a single GP that is defined by

$$f(\mathbf{Q}) \sim \mathcal{GP}(\mu_{\mathrm{M}}(\mathbf{x}), k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}'))\,, \tag{10}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are inputs from tasks j and j', respectively. As with the single-task GP, we assume $\mu_{\mathrm{M}}(\mathbf{x}) = 0$. MTGP's covariance function, $k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}')$, is formed by placing a GP prior over the kernel function in Eq. 6, so that we directly induce correlations between the tasks [11]. It is given by

$$k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}') = k^{\mathrm{c}}(j, j') \times k^{\mathrm{x}}(\mathbf{x}, \mathbf{x}')\,, \tag{11}$$

where k^{c} is a correlation kernel that explains the relation between tasks j and j', and k^{x} is the covariance that explains the relation of inputs $\mathbf{x}$ and $\mathbf{x}'$. This approach is also known as the *intrinsic correlation model* [49].

Let $\mathbf{K}_{\mathrm{M}}$ be the covariance matrix of $\mathbf{Q}$, $\mathbf{K}^{\mathrm{c}}$ the task correlation matrix, and $\mathbf{K}^{\mathrm{x}}$ the covariance matrix of inputs. We define $\mathbf{K}_{\mathrm{M}}$ as

$$\mathbf{K}_{\mathrm{M}} = \mathbf{K}^{\mathrm{c}} \otimes \mathbf{K}^{\mathrm{x}}\,, \tag{12}$$

where $\otimes$ denotes a Kronecker product. $\mathbf{K}^{\mathrm{c}}$ is assumed to be a valid covariance matrix (satisfying Mercer's theorem). Its diagonal elements describe the correlation of the tasks with themselves and the non-diagonal elements correspond to the correlation between tasks. It can be constructed using the Cholesky decomposition and is parameterized by the elements of the lower triangular matrix of

$$\mathbf{K}^{\mathrm{c}}(j, j') = \mathbf{J}\mathbf{J}^\top,\ \mathbf{J} = \begin{pmatrix} \theta_1^{\mathrm{c}} & 0 & \dots & 0 \\ \theta_2^{\mathrm{c}} & \theta_3^{\mathrm{c}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{\zeta-m+1}^{\mathrm{c}} & \theta_{\zeta-m+2}^{\mathrm{c}} & \dots & \theta_{\zeta}^{\mathrm{c}} \end{pmatrix}, \tag{13}$$

where $\theta^{\mathrm{c}} = \{\theta_u^{\mathrm{c}}\}$, $u \in \{1, 2, \dots, \zeta\}$ is the set of $\mathbf{K}^{\mathrm{c}}$'s hyperparameters, with $\zeta = m(m+1)/2$.

Inference and hyperparameter learning in MTGPs is conducted similarly to the single task GPs [11, 18]. Given a new data point $\mathbf{x}_*$, for task j, the predictions (y_*) can be made by using the conditional distribution $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{Q}, \mathbf{Y}) \sim \mathcal{N}(\mu_{j*}, \sigma_{j*}^2)$, where

$$\mu_{j*} = \left(\mathbf{k}_j^{\mathrm{c}} \otimes \mathbf{k}_*^{\mathrm{x}}\right)^\top \mathbf{K}_{\mathrm{M}}^{-1}\mathbf{Y}\,, \text{ and} \tag{14}$$

$$\sigma_{j*}^2 = \mathbf{K}_{\mathrm{M}} + \mathbf{D} \otimes \mathbf{I}\,. \tag{15}$$

In the above equations, $\mathbf{k}_j^c$ is the j^{th} column of $\mathbf{K}^c$, $\mathbf{k}_*^x$ is the vector of covariances between $\mathbf{x}_*$ and the training points, and $\mathbf{D}$ is an $m \times m$ matrix in which the $(j,j)^{\text{th}}$ element is the noise variance (σ_j^2) for the j^{th} task.

3 EXPERIMENTS

Our experiments assess a number of different disease modeling scenarios, where we expect that multi-task learning will have a positive impact. We focus on the estimation of ILI rates, which is a well-studied task [22, 31, 43]. The locations of interest are the US at the national level, US regions as defined by the Department of Health and Human Services (HHS), and England.

3.1 Data Sets and Experiment Settings

ILI rates from health agencies. For the US, we use weekly ILI rates from the Centers for Disease Control and Prevention (CDC). These rates represent the average percentage of all outpatient visits to health care providers normalized by the respective regional population figures and are recorded by CDC's ILI surveillance network, ILINet.[2] The 10 HHS US regions considered by the CDC are shown in Fig. 1. Our data spans from September 1, 2007 to August 31, 2016 (both inclusive), which includes 9 consecutive influenza seasons as defined by the CDC. Each (expanded) flu season begins on September 1 and ends on August 31 of the next year. To provide further insight, we have plotted the ILI rates of US regions 1, 2, and the US as a whole in Fig. 2. As expected, we see that the time series are strongly correlated, but each signal may be peaking at different moments throughout a flu season. For England, we obtain weekly ILI rates from Public Health England (PHE) through the syndromic surveillance network developed by the Royal College of General Practitioners. We focus on the same time period as for the US.

Search query frequencies. We iteratively used Google Correlate[3] starting with flu-related query seeds (such as the word 'flu') to obtain a set of 1,641 candidate search queries. However, due to the existing seasonal confounders, many of the candidate queries we ended up with, such as 'college basketball' or 'spring break', were not related to flu. To remove these unrelated queries in a principled fashion, we applied a topic filter specified using word embeddings. The filtering process was similar to the one we proposed in [34], but without the notion of a negative context. Embeddings were trained using `word2vec` on Google news [40].[4] We consider a query q as a set of z textual tokens, $\{\varepsilon_1, \ldots, \varepsilon_z\}$. The embedding of q, $\mathbf{e}_q$, is computed by averaging across the embeddings of its tokens,

$$\mathbf{e}_q = \frac{1}{z} \sum_{i=1}^{z} \mathbf{e}_{\varepsilon_i} \,. \tag{16}$$

We define a topic about flu, $\mathcal{T}$, as a set of two flu-related terms, specifically the name of the disease and one of its main symptoms, $\mathcal{T} = \{$'flu', 'fever'$\}$. For each of the queries, we calculate a similarity score defined as the product of the cosine similarities between the embeddings of the terms in $\mathcal{T}$ and $\mathbf{e}_q$, i.e.

$$S(\mathbf{q}, \mathcal{T}) = \prod_{i=1}^{2} \cos(\mathbf{e}_\mathbf{q}, \mathbf{e}_{\mathbf{T}_i}) \,, \tag{17}$$

where each cosine similarity component is mapped to $[0, 1]$ via $(\cos(\cdot,\cdot) + 1) / 2$.[5] Queries with $S \leq 0.5$ are filtered out and are not considered in our experiments. The 0.5 threshold guarantees that even in the extreme case, where a candidate query has a perfect cosine similarity (equal to 1) with one of the two concept queries, it also needs to have a non-negative cosine similarity (prior to the $[0, 1]$ mapping) with the other concept query. The semantic filter succeeds in eliminating some confounding features, i.e. queries that may be highly correlated with ILI rates, but are referring to different topics.[6]

We retain 128 search queries after applying the word embedding filter described above.[7] The frequencies of these queries are retrieved through a private Google Health Trends API, provided for academic research with a health-oriented focus. The query frequency expresses the probability of a short search session[8] conducted within a geographic region and during a specified time period. The probability is estimated based on a 10-15% sample of all Google searches. We obtained daily frequencies at the state-level (for the US) and the national-level (for the US and England) from September 1, 2007 to August 31, 2016 (both inclusive). Weekly frequencies were estimated by averaging the daily frequencies. Similarly, regional US frequencies were computed by averaging the state-level frequencies.

Baselines, evaluation and parameter learning. To demonstrate the effectiveness of multi-task learning models, we compare MTEN and MTGP with their single-task formulations, EN and GP, respectively. We use Pearson correlation (r) and the Mean Absolute Error (MAE) between inferred and target ILI rates as our evaluation metrics. For reporting the performance of multi-task learning models, we use the average MAE and correlation of the different test periods across all tasks (locations). The statistical significance of a performance improvement is tested via a paired-sample t-test by using the mean MAEs across all locations for the applied test periods (for the two methods under comparison). In our results, we use an asterisk (*) to indicate that a difference in performance is **not** statistically significant at the .05 level ($p\text{-}value \geq .05$). For learning the regularization parameters of the linear models, we perform grid search on 20% of the training data; all models are trained on the remaining 80% subset of the training data. We begin by training a model on data from the first ϕ flu seasons, and test the model in the following season ($\phi + 1$). Then, we increase our training data by including one more flu season ($\phi + 1$) and test in the following season ($\phi + 2$); we repeat this process until we have tested on the last flu season in our data set. Before training a model, we only retain search queries that have a Pearson correlation higher than

[2] See `gis.cdc.gov/grasp/fluview/fluportaldashboard.html`

[3] Google Correlate, `google.com/trends/correlate`

[4] The embeddings were downloaded from `code.google.com/archive/p/word2vec`. The specific training settings are detailed in [40].

[5] This resolves misleading similarity scores based on different sign combinations.

[6] All candidate queries together with their similarity scores are listed at `github.com/binzou-ucl/google-flu-mtl`.

[7] For the experiments on England, two queries referring to medication available in the US are replaced by England-based equivalent medication (see Section 3.4).

[8] A search session can be seen as a time window that may include more than one consecutive search queries from a user account. Therefore, a target search query is identified as a part of a potentially larger query set within a search session.

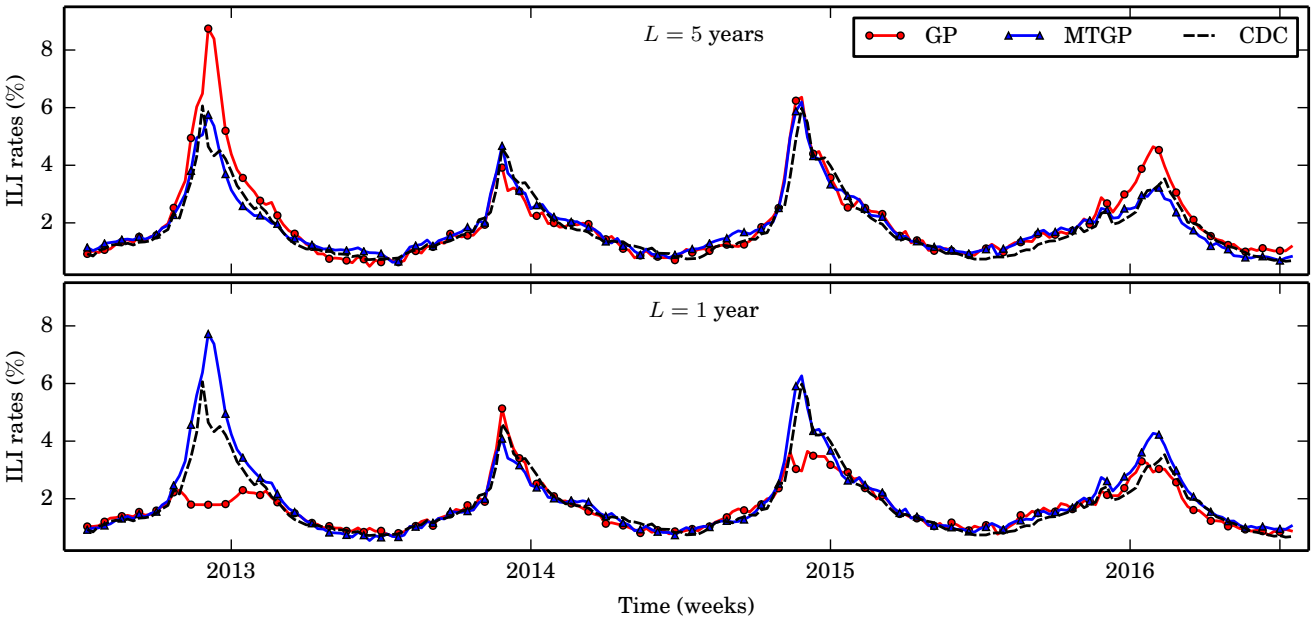

Figure 3: Comparing GP (red) and MTGP (blue) ILI estimates for the US using $L = 5$ years and $L = 1$ year of training data.

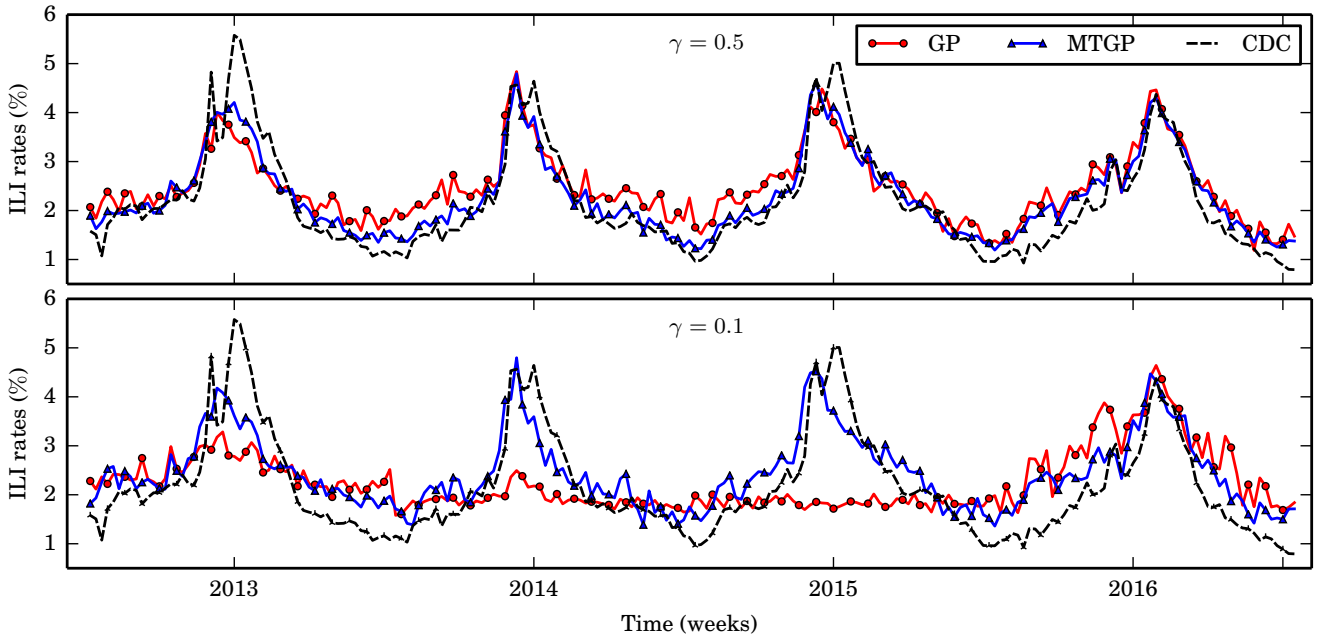

Figure 4: Comparing GP (red) and MTGP (blue) ILI estimates for US Region 9 for two burst error sampling (type C) rates (γ).

.3 with the respective disease rates (per location). This correlation threshold choice was motivated by the extensive experiments we conducted in [34] (see Table 3 in that paper). Note that the correlation filter is applied to each training data set separately and it may result in retaining different features for each task. Whenever this is the case, we maintain the intersection of features among the tasks. In addition, the GP and MTGP models are trained on the features that received a nonzero weight by the respective elastic net model, similarly to the methodology proposed in [31].

3.2 Multi-Task Learning on US Regional and National ILI Surveillance Tasks

First, we investigate whether multi-task learning can improve the accuracy of regional US models for the estimation of ILI rates. We

Table 1: Performance of single and multi-task learning models for estimating ILI rates on US HHS regions. L denotes the length of the training period in years.

	EN		MTEN		GP		MTGP	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.928	.347	.935	.344*	.936	.335	.944	.330*
4	.919	.379	.927	.371*	.926	.355	.938	.346*
3	.912	.398	.921	.385*	.916	.382	.929	.369*
2	.901	.438	.913	.414	.906	.424	.924	.398
1	.845	.531	.858	.491	.844	.535	.867	.467

The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

test this hypothesis under a decreasing amount of training samples, where L varies from 5 to 1 year(s) of historical data. By doing this we can additionally assess whether multi-task learning models can have a positive impact when the historical training data are limited. The multi-task learning models are trained on data from the 10 US HSS regions jointly and their performance is compared to the performance obtained by learning these models separately.

Table 1 enumerates the performance for the aforementioned comparison.[9] We observe that in general multi-task learning models perform better than their single-task alternatives both in terms of MAE and correlation. In addition, the nonlinear models tend to outperform the linear ones. However, performance gains from multi-task learning (in MAE) only become statistically significant when $L \leq 2$ years of historical training data are used. The greatest improvement occurs for $L = 1$; for this case MTEN reduces EN's MAE by 7.5%, whereas the MTGP reduces GP's MAE by 12.7%.

We next expand our observations by adding data for the US at a national level. Hence, we are now considering 11 tasks (US plus the 10 US regions). The aim is to test whether we can obtain a better model at the national level by training it together with regional data in a multi-task learning fashion. The results enumerated in Table 2 confirm that this is the case. The impact of multi-task learning is greater and statistically significant (in terms of MAE), when $L \leq 3$ years. The greatest improvement happens for $L = 1$; for this case MTEN reduces EN's MAE by 12.6%, whereas the MTGP reduces GP's MAE by 14.8%. In Fig. 3, we compare the estimates from the GP and MTGP models for the ILI rates in the US during the test periods from 2012 to 2016 (4 flu seasons) under two different training data lengths (5 vs. 1 year of historical data) and against the rates reported by CDC. Even under the 5-year training period, where the difference in average performance between the models is small, we see that the GP makes a significant over-prediction of the peak during the 2012/13 flu season, something that the MTGP does not. The bottom sub-figure, where $L = 1$ year, showcases more clearly the level of improvement obtained by applying a multi-task learning scheme; MTGP delivers a quite accurate model despite being trained on a few samples. This is an important characteristic as it suggests that we can develop accurate disease prevalence models with much less historical data than previously considered [22, 29, 31].

[9]Numbers in the table represent the average performance across the 10 US regions and the 4 test periods. For additional clarity, all individual performance estimates (for $L = 1$) are enumerated at `github.com/binzou-ucl/google-flu-mtl`.

Table 2: Performance of single and multi-task learning (including regional data) models for estimating US ILI rates; notational conventions as in Table 1.

	EN		MTEN		GP		MTGP	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.960	.353	.962*	.351*	.965	.253	.966*	.245*
4	.951	.356	.954*	.353*	.947	.265	.949*	.251*
3	.939	.398	.945	.374	.942	.286	.947*	.268
2	.930	.408	.936	.362	.933	.351	.941	.323
1	.854	.531	.868	.464	.854	.513	.875	.437

The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

3.3 Mitigating the Effect of Sporadic ILI Health Reports with Multi-Task Learning

In many real-world scenarios, health surveillance reports are or can become temporally and/or geographically sporadic. For instance, syndromic surveillance networks, especially in developing countries, may focus on a few regions rather than an entire country due to infrastructure and economic constraints. Furthermore, established health surveillance schemes may be exposed to data loss due to unprecedented events, such as technical faults, natural disasters or a spreading epidemic during which doctor visits are discouraged. In the following experiments, we assess whether multi-task learning can help us establish more accurate disease models under various scenarios of sporadic health reporting. To assess this, we have performed several forms of down-sampling on the training data of several US HHS regions. All experiments were conducted by setting $L = 1$, i.e. based on 1-year long training periods, and results represent the average performance after 50 sampling trials.

We have applied the following sampling techniques: (**A**) *random weekly sampling*, (**B**) *random monthly sampling*, and (**C**) *random burst-error sampling*. In (A), we simply take random samples from our data, thereby simulating scenarios where reports for a specific week may be missing. In (B), we first partition our data into non-overlapping monthly periods and then randomly sample over these periods, thereby simulating situations where health systems may be affected for longer time periods. Finally, in (C) we randomly discard a block of temporally contiguous data points, and use the remaining points only. We apply a sampling rate $\gamma = \{0.1, 0.2, \ldots, 1\}$, where $\gamma = 1$ means that all data are used (no sampling), and $\gamma = 0.1$ that 10% of the weekly data (for A) or monthly periods (for B) are maintained. In C, γ determines the size of the error block B, $B = (1 - \gamma)\tau$, where τ is equal to the size of the training data. In all experiments, we are sampling per location, meaning that the time points in the training data can vary across locations.[10]

We begin by assessing the added value of multi-task learning in situations, where progressively less health reports are obtained for half of the regions of a country. To simulate this, we partition the 10

[10]We have also conducted experiments where sampling is temporally synchronized across regions, but we did not observe a significant difference in the performance outcomes. Due to space constraints, we only report the non-synchronized results.

Table 3: Performance of single and multi-task learning models for estimating ILI rates on US HHS regions belonging to $\mathcal{R}$-odd under three sampling methods (A, B and C). Training data in $\mathcal{R}$-odd regions is down-sampled using a sampling rate (γ).

		EN		MTEN		GP		MTGP	
	γ	r	MAE	r	MAE	r	MAE	r	MAE
	1.0	.825	.492	.843	.488*	.828	.502	.856	.460
A	0.9	.823	.504	.840	.494*	.825	.503	.852	.465
	0.8	.806	.512	.839	.498*	.817	.505	.850	.465
	0.7	.805	.523	.834	.499*	.811	.506	.849	.467
	0.6	.800	.528	.824	.501*	.804	.512	.835	.468
	0.5	.798	.541	.823	.502*	.804	.513	.835	.469
	0.4	.789	.550	.822	.508	.801	.534	.829	.469
	0.3	.768	.555	.817	.511	.801	.545	.825	.474
	0.2	.758	.567	.803	.520	.789	.564	.824	.476
	0.1	.698	.694	.793	.554	.700	.686	.824	.482
B	0.9	.813	.516	.835	.495*	.814	.519	.851	.463
	0.8	.806	.531	.827	.505*	.805	.528	.843	.468
	0.7	.793	.549	.823	.511*	.792	.540	.834	.475
	0.6	.775	.555	.821	.516	.776	.565	.825	.476
	0.5	.752	.574	.820	.523	.756	.570	.823	.478
	0.4	.702	.598	.818	.534	.751	.594	.819	.485
	0.3	.621	.751	.815	.544	.650	.748	.817	.491
	0.2	.510	.781	.814	.547	.516	.776	.814	.497
	0.1	.425	.942	.806	.583	.433	.930	.809	.503
C	0.9	.817	.524	.836	.497*	.818	.525	.848	.466
	0.8	.805	.539	.829	.506*	.810	.532	.839	.470
	0.7	.796	.554	.817	.513	.801	.552	.832	.471
	0.6	.784	.576	.814	.528	.788	.569	.825	.473
	0.5	.756	.606	.807	.535	.766	.588	.819	.477
	0.4	.689	.637	.799	.543	.713	.626	.818	.480
	0.3	.621	.739	.794	.557	.632	.711	.804	.492
	0.2	.483	.792	.781	.561	.506	.791	.800	.498
	0.1	.414	.934	.780	.571	.424	.906	.796	.505

The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

US HSS regions into two sub-groups, $\mathcal{R}$-odd and $\mathcal{R}$-even consisting of the odd and even regions respectively (following the numbering of Fig. 1). For the regions in $\mathcal{R}$-odd, we have increasingly down-sampled their training data; regions in $\mathcal{R}$-even were not subject to down-sampling.

Table 3 enumerates the results of this experiment. The numbers in the table represent the average MAE of all test periods over the $\mathcal{R}$-odd regions. Generally, the performance of the multi-task learning models degrades less as down-sampling increases, i.e. there are less training data. MTGP always offers a statistically significant improvement over GP, whereas MTEN, in the worst case (for sampling type A), requires a $\gamma \leq 0.4$ to achieve this. Type A sampling, which can be seen as having missing weekly reports in various regions at random time points, affects single task learning models much more than multi-task learning models. For example, for the EN model, the MAE increased from .492 for $\gamma = 1$ (no down-sampling), to .694

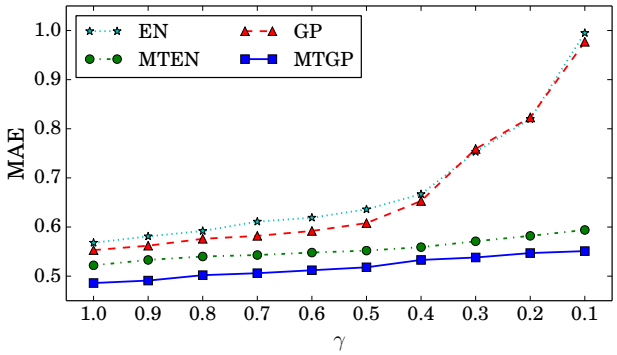

Figure 5: Comparing the performance of EN (dotted), GP (dashed), MTEN (dash dot) and MTGP (solid) on estimating the ILI rates for US HHS Regions (except Regions 4 and 9) for varying burst error sampling (type C) rates (γ).

for $\gamma = 0.1$, a degradation of 41.1%. In contrast, the MTEN model degrades by 13.5%. The effect is more pronounced for the nonlinear models, with GP degrading by 36.7% while MTGP degrades by only 4.8%. Note that MTGP's MAE is equal to .482 when the fewest data points are used (10% for $\gamma = 0.1$), which is smaller than EN's or GP's MAEs, when no sampling is taking place (.492 and .502 respectively).

All models degrade worse for B and C sampling methods, which drop blocks of data points from the training set. However, the degradation in performance of the multi-task learning models is much less than for the comparative EN or GP models. For example, when $\gamma = 0.1$, MTGP improves GP's MAE by 45.9% and 44.3% for B and C sampling types, respectively. Fig. 4 illustrates this performance difference by comparing the ILI estimates from the GP and MTGP models for US region 9 under burst error sampling, for $\gamma = 0.5$ (top) and $\gamma = 0.1$ (bottom).[11] Clearly, for low sampling rates ($\gamma = 0.1$) the MTGP model is still able to provide acceptable performance.

In a subsequent experiment, we performed burst-error sampling on all but two US regions with the highest population figures (Regions 4 and 9). The rational behind this setting is that in many occasions health reports are available for central locations in a country

[11] Region 9 includes the states of California, Nevada and Arizona and one of the largest in terms of population ($\approx$ 49.1 million).

Table 4: Performance of single and multi-task learning models for estimating ILI rates in England; notational conventions as in Table 1.

	EN		MTEN		GP		MTGP	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.885	.696	.896	.491	.891	.599	.903	.474
4	.873	.734	.887	.504	.880	.664	.894	.491
3	.860	.788	.876	.530	.868	.742	.883	.517
2	.854	.842	.871	.554	.859	.815	.875	.528
1	.836	.999	.857	.603	.846	.977	.860	.586

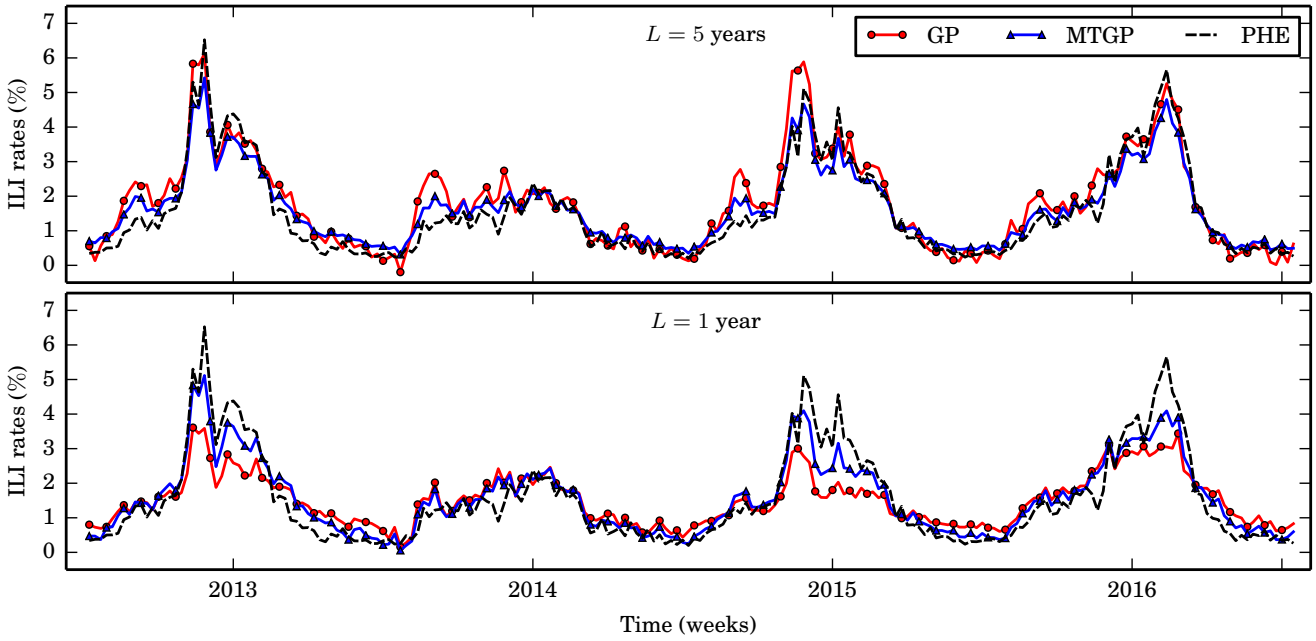

Figure 6: Comparing GP (red) and MTGP (blue) ILI estimates for England under varying training data sizes.

(i.e. two big cities), but are limited anywhere else. Fig. 5 compares the performance of all regression models under this scenario. It confirms that the pattern observed in the previous experiment still holds, i.e. that the multi-task models are much less affected by down-sampling. We can also see that MAE in single task learning models increases at an exponential rate as γ decreases.

3.4 Multi-Task Learning Across Countries

We expand on the previous results to test whether a stable data stream for a country could be used to enhance a disease model for a different, but culturally similar, country. The underlying assumption here is that countries that share a common language and have cultural similarities may also share common patterns of user search behavior.

For this purpose, we use data from the US and England and assume that there are increasingly less historical health reports for England only, in a similar fashion as in the experiments described in Section 3.2 (L from 5 to 1 year). For the US data, we always assume that the training window is based on the past $L = 5$ years. The search queries used in both countries are the same, with the following exception. Two of the US search queries about medication were changed to their British equivalent because their search frequencies in England are low; we changed "tussin" to "robitussin" and "z pak" to "azithromycin".

Table 4 shows a similar pattern of results as in the previous experiments. All multi-task learning models register statistically significant improvements compared to the single task learning ones. As the length of the training period is reduced, the improvements are greater; MTGP reduces MAE by 20.9% and 40.0% for $L = 5$ and $L = 1$ year, respectively. Fig. 6 presents the estimates for the GP and MTGP models for these extreme cases. Whereas both models seem to be inferring the trends of the time series correctly, the multi-task estimates are more close to the actual values of the signal's peaks.

The results confirm our original hypothesis that data from one country could improve a disease model for another country with similar characteristics. This motivates the development of more advanced transfer learning schemes [41], capable of operating between countries with different languages by overcoming language barrier problems, using variants of machine translation.

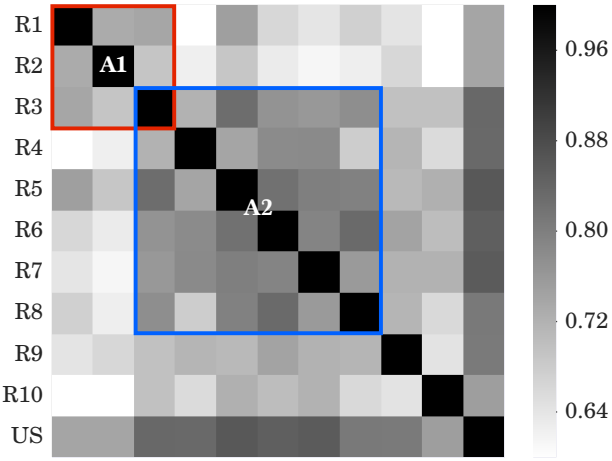

Figure 7: A heat map depicting MTGP's correlation matrix (K^c) for modeling ILI rates based on all US data (regional and national).

3.5 Qualitative Insights from the MTGP Model

The main motivation for using multi-task learning models within the context of disease surveillance from Web search data was our assumption that relations between *correlated* locations will be identified and accounted for. According to our results, the best performing model under the vast majority of experimental settings was the MTGP model. Given this empirical result, we assume that the hyperparameters of the MTGP model could provide further insight about the inner-workings of the model. After training an MTGP model on all the US data available (10 HHS regions and US nationally), we examined the inferred correlation matrix ($\mathbf{K}^c$ in Eq. 13), which we depict using a gray-scale heat map in Fig. 7. We can identify two areas of the heat map that are characterized by increased correlation values, denoted as A1 and A2. A1 and A2 are "clusters", representing groups of north-east and central states/regions of the US, respectively. Using the region numbering as a crude proxy for the distance between regions, we also observe that correlations are generally higher for neighboring areas. The same holds for smaller internal sub-clusters of the area A2 (e.g. regions R6-R8). Both observations provide further evidence that the MTGP model is probably capturing existing geographical relations.

4 RELATED WORK

The fundamentals of multi-task learning have been thoroughly presented in [12, 13]. Compared to single task learning that attempts training on isolated tasks, multi-task learning performs this jointly using a shared representation. The tasks can be used as valuable sources of inductive bias for each other, leading to a more accurate model [12, 13]. This may also allow more difficult problems, such as target variables with partial observations, to be modeled successfully [4, 8, 12, 13]. The majority of multi-task regression models were developed by extending their single-task formulations. Some examples for linear regression are the multi-task ℓ_1-norm regularization [3] and the $\ell_{2,1}$-norm regularization [36]. Nonlinear multi-task regression models have also been explored, extending Support Vector Machines [20], Gaussian Processes [11], Convolutional or Recurrent Neural Networks [1, 37].

In this work, we study the utility of multi-task learning in disease surveillance from Web search data. Existing approaches have routinely used single task models such as regularized regression [17, 22, 31, 43], Gaussian Processes [31, 34], and autoregressive frameworks [31, 42, 47]. Here, we have chosen to apply MTEN [35] and MTGP [11] for the following reasons: (a) EN and GPs have been applied in many text regression [27, 44] and disease modeling approaches [31, 33, 34, 55], and (b) the sample sizes we are operating on are limited and no performance gain would have been achieved by deploying neural network structures (such as [16, 51, 52]).

Multi-task learning has been applied in the context of user-generated data modeling [32, 38] and computational health [9, 10, 19, 53, 54]. Given various tasks and objectives, multi-task learning frameworks can be different. Zhou *et al.* and Emrani *et al.* formulated a fused sparse group lasso [54] and a graph regularization approach [19], respectively, aiming to model disease progression. Both models focused on the temporal relation between the various tasks and utilized image data from patients. However, our work focuses on textual user-generated content and the spatial relation among tasks. In [9], Benton *et al.* used online multimodal user-generated content to train a multi-task feedforward neural network for classifying the mental health condition of online users. This model tries to capture shared structures of user attributes in relation to mental conditions. Our work, however, focuses on a collective regression task, aiming to exploit relationships at a higher level, determined by geography, rather than specific user characteristics. Finally, Zhao *et al.* proposed a linear regularized multi-task regression model to detect civil unrest events in various locations using Twitter data [53]. In our work, apart from a different thematic focus, we also deploy nonlinear multi-task learning frameworks.

5 CONCLUSIONS

We have investigated the utility of multi-task learning to disease surveillance from Web search data. Disease surveillance models for various geographies — inside a country and across different countries — were trained jointly such that knowledge between different tasks could be shared. We explored both linear and nonlinear models (MTEN and MTGP) and used ILI surveillance as a case study. Experiments were conducted on the US and England. Our empirical results indicate that multi-task learning improves regional as well as national models for the US. The percentage of improvement increases as we reduce the historical training data. For a 1-year training period, the MTGP model improved MAE by 14.8% at the regional level. Furthermore, in simulated scenarios, where health reports (training data) are limited, we showed that multi-task learning helps to maintain a stable inference performance across all the affected locations. Experiments, where data for England were modeled in conjunction with US data, indicated that more accurate estimates were obtained for England, maxed at a 40% of MAE reduction when using 1-year long training periods. This suggests that multi-task learning can benefit models across different countries as well. Finally, our assumption that correlations in the search behavior of users across similar geographies and cultures will significantly assist this type of disease modeling is also supported by empirical evidence.

Future work will aim to extend this type of modeling by developing appropriate frameworks for transfer learning, e.g. between countries with different languages, and apply it in real-world situations. These should include locations (regions or countries) with underdeveloped disease surveillance schemes as well as different disease types for which fewer historical health reports are available.

ACKNOWLEDGMENTS

This work has been supported by the grant EP/K031953/1 (EPSRC). The authors would like to acknowledge PHE and RCGP for providing syndromic surveillance data, and Google for providing access to the Google Health Trends API. We also thank the anonymous reviewers for their constructive feedback.

REFERENCES

[1] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia. 2015. Multi-Task CNN Model for Attribute Prediction. *IEEE Transactions on Multimedia* 17, 11 (2015), 1949–1959.
[2] A. Argyriou, T. Evgeniou, and M. Pontil. 2006. Multi-Task Feature Learning. In *Proceedings of Advances in Neural Information Processing Systems 19.*
[3] A. Argyriou, T. Evgeniou, and M. Pontil. 2008. Convex Multi-Task Feature Learning. *Machine Learning* 73, 3 (2008), 243–272.

[4] B. Bakker and T. Heskes. 2003. Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research* 4 (2003), 83–99.

[5] E. Bakshy, S. Messing, and L. A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[6] J. Baxter. 2000. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research* 12, 1 (2000), 149–198.

[7] D. Beck, T. Cohn, and L. Specia. 2014. Joint Emotion Analysis via Multi-task Gaussian Processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1798–1803.

[8] S. Ben-David and R. Schuller. 2003. Exploiting Task Relatedness for Multiple Task Learning. In *Proceedings of the 16th Annual Conference on Learning Theory and 7th Kernel Workshop*. 567–580.

[9] A. Benton, M. Mitchell, and D. Hovy. 2017. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 152–162.

[10] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. 2008. Multi-Task Learning for HIV Therapy Screening. In *Proceedings of the 25th International Conference on Machine Learning*. 56–63.

[11] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. 2007. Multi-task Gaussian Process Prediction. In *Proceedings of Advances in Neural Information Processing Systems 20*. 153–160.

[12] R. Caruana. 1993. Multitask Learning: A Knowledge-based Source of Inductive Bias. In *Proceedings of the 10th International Conference on Machine learning*. 41–48.

[13] R. Caruana. 1998. Multitask Learning. In *Learning to Learn*. Springer, 95–133.

[14] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. 128–137.

[15] T. Cohn and L. Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 32–42.

[16] R. Collobert and J. Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*. 160–167.

[17] A. Culotta. 2010. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proceedings of the 1st Workshop on Social Media Analytics*. 115–122.

[18] R. Durichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton. 2014. Multi-task Gaussian process Models for Biomedical Applications. In *Proceedings of the 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics*. 492–495.

[19] S. Emrani, A. McGuirk, and W. Xiao. 2017. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1457–1466.

[20] T. Evgeniou and M. Pontil. 2004. Regularized Multi-Task Learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 109–117.

[21] H. Gil de Zúñiga, N. Jung, and S. Valenzuela. 2012. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *Journal of Computer-Mediated Communication* 17, 3 (2012), 319–336.

[22] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting Influenza Epidemics using Search Engine Query Data. *Nature* 457, 7232 (2009), 1012–1014.

[23] T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition.* Springer.

[24] A. E. Hoerl and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 1 (1970), 55–67.

[25] M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[26] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[27] V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 405–413.

[28] V. Lampos and N. Cristianini. 2010. Tracking the flu pandemic by monitoring the Social Web. In *Proceedings of the 2nd International Workshop on Cognitive Information Processing*. 411–416.

[29] V. Lampos and N. Cristianini. 2012. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology* 3, 4 (2012), 1–22.

[30] V. Lampos, T. De Bie, and N. Cristianini. 2010. Flu Detector - Tracking Epidemics on Twitter. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. 599–602.

[31] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports* 5, 12760 (2015).

[32] V. Lampos, D. Preoţiuc-Pietro, and T. Cohn. 2013. A user-centric model of voting intention from social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 993–1003.

[33] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. 2015. Assessing the Impact of a Health Intervention via User-generated Internet Content. *Data Mining and Knowledge Discovery* 29, 5 (2015), 1434–1457.

[34] V. Lampos, B. Zou, and I. J. Cox. 2017. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proceedings of the 26th International Conference on World Wide Web*. 695–704.

[35] S. Lee, J. Zhu, and E. P. Xing. 2010. Adaptive Multi-task Lasso: With Application to eQTL Detection. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*. 1306–1314.

[36] J. Liu, S. Ji, and J. Ye. 2009. Multi-task Feature Learning via Efficient $\ell_{2,1}$-Norm Minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 339–348.

[37] P. Liu, X. Qiu, and X. Huang. 2016. Recurrent Neural Network for Text Classification with Multi-task Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 2873–2879.

[38] M. Lukasik, T. Cohn, and K. Bontcheva. 2015. Classifying Tweet Level Judgements of Rumours in Social Media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2590–2595.

[39] A. M. Manago, T. Taylor, and P. M. Greenfield. 2012. Me and my 400 friends: The anatomy of college students' Facebook networks, their communication patterns, and well-being. *Developmental Psychology* 48, 2 (2012), 369–380.

[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26*. 3111–3119.

[41] S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.

[42] M. J. Paul, M. Dredze, and D. Broniatowski. 2014. Twitter Improves Influenza Forecasting. *PLOS Currents Outbreaks* (2014).

[43] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. 2008. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases* 47, 11 (2008), 1443–1448.

[44] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. 1754–1764.

[45] C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

[46] H. A. Schwartz et al. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8, 9 (2013), e73791.

[47] J. Shaman and A. Karspeck. 2012. Forecasting Seasonal Outbreaks of Influenza. *Proceedings of the National Academy of Sciences* 109, 50 (2012), 20425–20430.

[48] R. Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* 58, 1 (1996), 267–288.

[49] H. Wackernagel. 2013. *Multivariate Geostatistics: An Introduction with Applications*. Springer.

[50] M. Wagner, V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. 2017. Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content. *Journal of Medical Internet Research* 19, 12 (2017), e416.

[51] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji. 2015. Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1475–1484.

[52] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. 2014. *Facial Landmark Detection by Deep Multi-task Learning*. 94–108.

[53] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. 2015. Multi-Task Learning for Spatio-Temporal Event Forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1503–1512.

[54] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. 2012. Modeling Disease Progression via Fused Sparse Group Lasso. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1095–1103.

[55] B. Zou, V. Lampos, R. Gorton, and I. J. Cox. 2016. On Infectious Intestinal Disease Surveillance using Social Media Content. In *Proceedings of the 6th International Conference on Digital Health*. 157–161.

[56] H. Zou and T. Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.