# Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions

Prithwish Chakraborty [*] Pejman Khadivi [*] Bryan Lewis [†] Aravindan Mahendiran [*]
Jiangzhuo Chen [†] Patrick Butler [*] Elaine O. Nsoesie [†‡§] Sumiko R Mekaru[‡]
John S Brownstein[‡§] Madhav Marathe [†] Naren Ramakrishnan [*]

**Abstract**

Modern epidemiological forecasts of common illnesses such as the flu rely on both traditional surveillance sources as well as digital surveillance data such as social network activity and search queries. However, most published studies have been retrospective. For a real-time prediction system, we posit that one of the key challenges is to effectively handle the uncertainty associated with reports of flu activity. Such reports are in general lagged by several weeks and typically revised for several weeks after they are first reported. In this paper, we present a detailed prospective analysis of the generation of robust quantitative predictions of temporal trends of flu activity using several *surrogate* data sources for 15 Latin American countries. We present our findings about the limitations and advantages of correcting the uncertainty associated with official flu estimates. We also compare the prediction accuracy between model-level fusion of different surrogate data sources against data-level fusion. Finally, we present a novel matrix factorization approach using neighborhood embedding to predict flu case counts. Comparing our proposed ensemble method against several baseline methods helps us demarcate the importance of different data sources for the countries under consideration.

## 1 Introduction

Surveillance reports published by health organizations are one of the primary resources for monitoring influenza like illness (ILI) cases and, for years, have been the primary source of information used by healthcare officials for policy decisions. However traditional surveillance reports are published with a considerable delay and thus recent research has focused on mining social signals from search engine query volume [1, 2] and social media chatter [3, 4, 5, 6, 7].

One of the pioneering works in this space is the work of Ginsberg et al. [2] where ILI case counts are predicted from the volume of search engine queries. This work inspired significant follow-on work, e.g., [1], where Yuan et al. used search query data from Baidu (a popular search engine in China) to detect influenza outbreaks. More real-time ILI detection [4] systems have been proposed by modeling Twitter streams.

Apart from such social media sources, there has also been considerable research on exploiting physical indicators such as climate data. The primary advantage of such data sources is that the effects are much more causal and less noisy. Shaman et. al. [8, 9, 10] explored this area in detail and found absolute humidity to be a good indicator of influezna outbreaks.

While the above works have made important strides, there are important areas that have been relatively less studied. First, only a few works have focused on combining mutiple data sources [11, 3] to aid in forecasting. In particular, to the best of our knowledge there has been no work that investigates the combination of social indicators and physical indicators to forecast ILI incidence. Second, and more improtantly, official estimates as reported by health organizations (e.g., WHO, PAHO) are often lagged by several weeks and even when reported are typically revised for several weeks before the case counts are finalized. Real-time prediction systems must be designed to handle the forecasting of such a 'moving target'. Finally, most existing works have been retrospective and not set in the context of a formal data mining validation framework. To overcome these deficiences, we propose a novel approach to ILI case count forecasting. Our contributions are:

- Our approach integrates both social indicators and physical indicators and thus leverages the selective superiorities of both types of feature sets. We systematize such integration using a novel matrix factorization-based regression approach using neighborhood embedding, thus helping account for

[*]Dept. of Computer Science, Virginia Tech, Blacksburg, VA, USA

[†]Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

[‡]Childrens Hospital Informatics Program, Boston Childrens Hospital, Boston, MA, USA

[§]Department of Pediatrics, Harvard Medical School, Boston, MA, USA

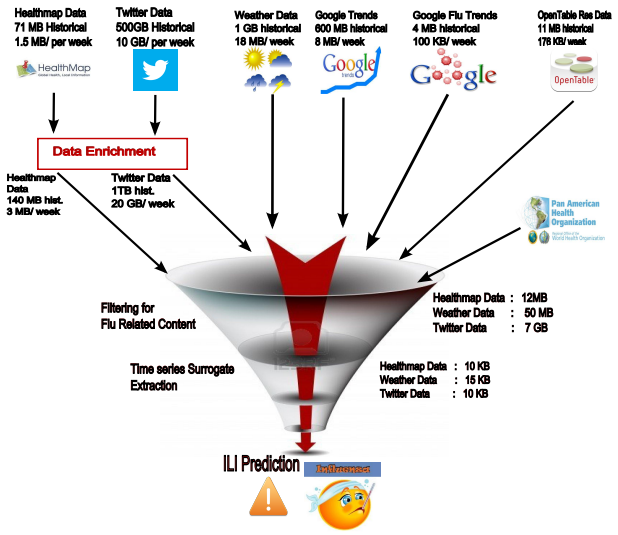



Figure 1: Our ILI data pipeline, depicting six different data sources used in this paper to forecast ILI case counts.

non-linear relationships between the surrogates and the official ILI estimates.

- We investigate the efficacy of combining diverse different sources at two levels: data fusion level, and model level, and discuss the relative (de)merits.
- We propose different ways of handling uncertainties in the official estimates and factor these uncertainities into our prediction models.
- Finally, we present a detailed and prospective analysis of our proposed methods by comparing predictions from a near-horizon real time prediction system to official estimates of ILI case counts in 15 countries of Latin America.

## 2 Related Works

Related work naturally falls into the categories of social media analytics, physical indicators, and event dynamics modeling.

**Social media analytics:** Most work in social media analytics focuses on Twitter, specifically tracking a dictionary of ILI-related keywords in the data stream. Some investigations have focused on the importance of diversity in keyword lists, e.g., [5, 6]. In [5], Kanhabua and Nejdl used clustering methods to determine important topics in Twitter data, construct time series for matched keywords, and used Jaccards coefficient to characterize the temporal diversity of tweets. They note that such temporal diversity may be correlated with real-world ILI outbreaks. In [6] the authors study the dynamics of changes in tweets related to the H1N1 virus. We present our approach towards creating the keyword dictionary in Section 6.3.1.

**Physical indicators for detecting ILI incidence levels:** Tamerius et al. [8] investigated the existence of seasonal cycles of influenza epidemics in different climate regions by considering climatic information from 78 globally distributed sites. Through logistic regression they found that, in cold-dry and humid-rainy environments, strong correlation exists between influenza epidemics and weather conditions. Similary exciting results were found by Shaman et. al. [9, 10] where they discovered absolute humidity to be a key indicator of flu. To uncover these relationships they used non-linear regressors such as Kalman filters, and this was a key inspiration to us to find a uniform model for the vaired data sources as explained in Section 3.1.

**Event dynamics modeling:** Denecke et al. [3] have proposed an event-based approach for early prediction of ILI threats [3]. Their method (M-Eco) considers multiple resources such as Twitter, TV reports, online news articles, and blogs and uses clustering to identify signals for event detection. Network dynamic solutions have also been used [12] to study the behavior of an epidemic in a society.

## 3 Problem Formulation

Let $\mathcal{P} = \langle P_1, P_2, \ldots, P_T \rangle$ denote the known total weekly ILI case count for the country under consideration, where $P_t$ denotes the case count for time point $t$ and $T$ denotes the time point till which the ILI case count is known. Corresponding to the ILI case count data, let us denote the available surrogate information for the same country by $\mathcal{X} = \langle \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{T1} \rangle$, where $T1$ is the time point till which the surrogate information is available and $\mathcal{X}_t$ denotes the surrogate attributes for time point $t$. The problem we desire to solve is to find a predictive model ($f$) for the case count data, as presented formally in Eqn 3.1.

$$f : \mathcal{P}_t = f(\mathcal{P}, \mathcal{X}) \tag{3.1}$$

**3.1 Methods** We employ non-linear temporal regressions over the surrogate attributes to forecast the case count using three broad models: (a) Matrix Factorizaiton Based Regression (MF), (b) Nearest Neighbor Based Regression (NN), and (c) Matrix Factorization Regression using Nearest Neighbor embedding (MFN). For each of the methods, we define two parameters: $\beta$ and $\alpha$. $\alpha$ is the *lookahead window length*, denoting distance of the time point for prediction from $T$; $\beta$ is the *lookback window length* denoting the number of time points to look back in order to find the regression relation between the case count and the surrogate data.

We define regression vectors $V_t$ and labels $L_t, \forall t = 1, \ldots, T$ as below:.

$$\begin{aligned} V_t &\equiv \langle P_{t-\beta-\alpha}, \mathcal{X}_{t-\beta-\alpha}, P_{t+1-\beta-\alpha}, \mathcal{X}_{t+1-\beta-\alpha}, \ldots, \\ &\quad P_{t-\alpha}, \mathcal{X}_{t-\alpha} \rangle \\ L_t &\equiv P_t \end{aligned}$$

The regression vector for predicting the case count at time point $T'(T+\alpha > T' > T)$ is given by equation 3.2.

$$V_{T'} \equiv \langle P_{T'-\beta-\alpha}, \mathcal{X}_{T'-\beta-\alpha}, P_{t+1-\beta-\alpha}, \mathcal{X}_{t+1-\beta-\alpha}, \ldots, P_{T'-\alpha}, \mathcal{X}_{T'-\alpha} \rangle \tag{3.2}$$

Under these definitions we describe the models as follows:

**3.1.1 Matrix Factorization Based Regression (MF):** Matrix Factorization is a well accepted technique in the recommender systems literature, to predict user preferences from incomplete user ratings/informations. Typically [13] a user-preference matrix is factored into an user-factor and factor-preference matrix. However, such factorizations are incognizant of any temporal continuity. As such to enforce temporal continuity, to predict for the time point $T'(T+\alpha > T' > T)$ we use the regression vectors and labels as defined earlier, to define a $m \times n$ *prediction matrix* $\mathcal{M}$, as given in equation 3.3:

$$\mathcal{M} = \begin{bmatrix} V_{\alpha+\beta+1} & L_{\alpha+\beta+1} \\ \vdots & \vdots \\ V_T & L_T \\ V_{T'} & L_{T'} \end{bmatrix} \tag{3.3}$$

The prediction matrix is factorized into a $f \times m$ factor-feature matrix $U$ and a $f \times n$ factor-prediction matrix as:

$$\widehat{\mathcal{M}}_{i,j} = b_{u,i} + U_i^T \times F_j$$

Here, $b_{i,j}$ is the baseline estimate given by:

$$b_{i,j} = \bar{\mathcal{M}} + b_j \tag{3.4}$$

where $\bar{\mathcal{M}}$ represents the all-element average and $b_j$ represents the column wise deviations from the average and is generally a free-parameter, i.e., it is fitted as part of the optimization problem. $U$ and $F$ matrix are estimated by minimizing the error function:

$$b_*, F, U = \text{argmin}(\sum_{i=1}^{m-1} \left(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n}\right)^2 + \lambda_1 \times (\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2)) \tag{3.5}$$

where $\lambda_1$ is a regularization parameter. An important design criteria in the error function of Eqn 3.5 is the fact that we only compute the error between the predicted label values and the actual label values i.e., the $n^{th}$ column of the prediction matrix $\mathcal{M}$. The rationale behind this choice is the fact that unlike traditional recommender systems we are only concerned with the label column and can sacrifice reconstruction accuracies for other columns.

The lookback window $\beta$, the factor size $f$ and the regularization parameter $\lambda_1$ are estimated using cross-validation and the final prediction for time point $T'$ is given by:

$$\widehat{P}_{T'} = b_{m,n} + U_m^T \times F_{m,n}$$

**3.1.2 Nearest Neighbor Based Regression (NN):** For nearest neighbor models, we define a training set $\Gamma_{NN} = \{V_t, L_t\}$, where $V_t$ represents the regression attributes and $L_t$ denote the corresponding labels. Also, let us define the set $\mathcal{N}(i) = \{k : V_k \text{ is one of the top K nearest neighbors of } V_i\}$ where $K$ indicates the maximum number of nearest neighbors considered. The predicted count $\widehat{P}_{T'}$ for the time point $T'$ is given as:

$$\widehat{P}_{T'} = \frac{\sum_{k \in \mathcal{N}(T')} \theta_k L_{k,T-\alpha}}{\sum_{k=1}^{K} \theta_k} \tag{3.6}$$

Here $\theta_k$ indicates the weight assigned to the $k^{th}$ nearest neighbor. Typically the inverse Euclidean distances to $V_{T'}$ are chosen as the weights.

**3.1.3 Matrix Factorization Based Regression using Nearest Neighbor Embedding (MFN):** It has been shown in [14] that matrix factorization using nearest neighbor constraints can outperform classical matrix factorization approach as well as traditional nearest neighbor approaches towards recommender systems. Here, we modify the method to suit the temporal nature of our problem in similar ways as described in section 3.1.1. We again define a similar prediction matrix $\mathcal{M}$ (see equation 3.3). Following [14], we define the matrix decomposition rule as

$$\widehat{\mathcal{M}}_{i,j} = b_{i,j} + U_i^T \times F_j + F_j \times |\mathcal{N}(i)|^{-\frac{1}{2}} \sum_{k \in N(i)} (\mathcal{M}_{i,k} - b_{i,k}) x_k \tag{3.7}$$

The key difference between equation 3.7 and the one proposed in [14] is that we don't have any term for implicit feedback and, further, only the top $K$ neighbors as found through Euclidean distance are used. The model is fitted using Eqn 3.8 as given below:

$$b_*, F, U, x_* = \text{argmin}(\sum_{i=1}^{m-1} \left(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n}\right)^2 + \lambda_2 \times (\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2 + \textstyle\sum_k ||x_k||^2)) \tag{3.8}$$

## 4 Ensemble Approaches

In the last section, we described different strategies to correlate a specific source with the ILI case count of a specfic country and predict future ILI counts. In practice, we desire to work with a multitude of data

sources and there are two broad ways to accomplish this objective. In data level fusion, a single regressor is constructed from different data sources to the ILI case count, while in model level fusion, we build one regressor for each data source and subsequently combine the predictions from the models. In this section, we describe these fusion methods. Experimental results with both methods are presented in Section 7.

**4.1 Data level fusion:** Here we express the feature vector $\mathcal{X}$ as a tuple over all the different data sources and then proceed with either of the regression methods as outlined in Section 3.1. For example, while combining Twitter and weather data sources (see Fig. 1), the feature vector $\mathcal{X}$ is given by:

$$\mathcal{X}_t = \langle \mathcal{T}_t, \mathcal{W}_t \rangle$$

where $T_t$ and $W_t$ denote attributes derived from Twitter and weather, respectively.

**4.2 Model level fusion:** In this approach, the models are combined using matrix factorization regression with nearest neighbor embedding by comparing the prediction estimates from each model with the actual estimate (since the ground truth can change as well) and the average ILI case count for the month for the particular country (to help organize a baseline). Let us denote the average ILI case count for a particular calendar month $I$ for a given country by:

$$\mu_I = \frac{1}{|\{t \in I\}|} \sum_{t \in I} P_t$$

Considering $C$ different sources and hence $C$ different models, let us denote the prediction for the $t^{th}$ time point from the $c^{th}$ model by ${}_c\widehat{P}_t$.

Using these definitions we can now proceed to describe the fusion model. Essentially, the model is similar to the one described in Section 3.1.3, with the difference being the way we construct the feature vectors. Similar to Eqn 3.3, we construct a prediction $m' \times n'$ matrix for fusion given by ${}_C\mathcal{M}$ where the $t^{th}$ row is represented by equation 4.9.

$$ {}_C\mathcal{M}_t = \left[ \begin{array}{cccc} {}_1\widehat{P}_t & \ldots & {}_C\widehat{P}_t & P_t \end{array} \right] \tag{4.9}$$

Then similar to Eqn 3.7, we factor this matrix into latent factors, ${}_CU$, ${}_CF$, ${}_Cb_*$ as given by Eqn 4.10:

$$\begin{aligned} {}_C\widehat{\mathcal{M}}_{i,j} &= \mu_i + {}_Cb_j + {}_CU_i^T \times {}_CF_j \\ &+ {}_CF_j \times |{}_C\mathcal{N}(i)|^{-\frac{1}{2}} \textstyle\sum_{k \in {}_CN(i)} ({}_C\mathcal{M}_{i,k} - \mu_i + {}_Cb_k) {}_Cx_k \end{aligned} \tag{4.10}$$

so that the final prediction for the $T$th data point is given by

$$\widehat{P}_T = {}_C\widehat{\mathcal{M}}_T, n'.$$

The fitting function is given by equation 4.11:

$$\begin{aligned} {}_Cb_*, {}_CF, {}_CU, {}_Cx_* = \text{argmin}( & \sum_{i=1}^{m'-1} \left( {}_C\mathcal{M}_{i,n'} - {}_C\widehat{\mathcal{M}}_{i,n'} \right)^2 \\ & + \lambda_3 ( \sum_{j=1}^{n'} {}_Cb_j^2 + \sum_{i=1}^{m'-1} ||{}_CU_i||^2 \\ & + \sum_{j=1}^{n'} ||{}_CF_j||^2 + \textstyle\sum_k ||{}_Cx_k||^2 )) \end{aligned} \tag{4.11}$$

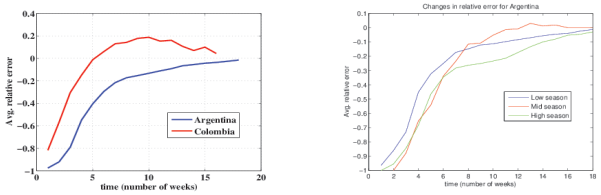
As before the free parameters are estimated through cross-validation.

## 5 Forecasting a Moving Target

One of the key challenges in creating a prospective ILI case count predictor is the fact that the official estimates are often delayed and, furthermore, even when published the estimates are revised over a number of weeks before these become finally stable. For this paper, we concentrate on 15 Latin American countries as described in Section 6 and consider the official ILI estimates from the Pan American Health Organiation (PAHO).Thus we can categorize PAHO count values downloaded on any week into three different types: (a) the uknnown PAHO counts reprenseted by $\ddot{P}_t$, (b) the known and stable PAHO counts denoted by $\dot{P}_t$, and (c) the known and unstable PAHO counts denoted by $\tilde{P}_t$. While we desire to predict $\ddot{P}_t$, the uncertainity associated with $\tilde{P}_t$ introduces errors in the predictions. In this section, we study the effects of such unstable data and propose three different models to adjust these unstable values to more accurate ones.

Figure 2a plots the relative error of an unstable PAHO data series w.r.t. its final estimate, as a function of time. It can be seen that different countries have different stability characteristics: for some countries, PAHO count values are stabilized very slowly whereas for others they stabilize faster (esp as the number of updates for a week increases). Stability behavior of PAHO count values were also found to be dependent on the time of the year as shown in Fig. 2b. To plot this curve for Argentina, we categorized any week with less than 100 cases to belong to a low season, greater than 300 to be a high season, and the remaining values to be mid season (the thresholds were different for different countries).

At the same time, the PAHO official updates provide an indication of the number of samples used to generate the case count estimate. Preliminary experiments show that this size is correlated with the accuracy of ILI case counts. In other words, in general, larger values of statistical population size results in smaller relative errors for ILI case count. Thus using both the number of samples and the lag in uploading the week data, we




(a) (b)

Figure 2: Average relative error of PAHO count values with respect to stable values. (a) Comparison between Argentina and Colombia (b) Comprison between different seasons for Argenitna.

can use machine learning techniques to revise the officially published PAHO estimates. Preliminary results show that for different seasons and different countries, we encounter different stability patterns. Therefore, any PAHO count adjustment method should be customized for seasons and countries separately.

Let us assume that $\dot{\mathcal{P}}$ is the set of stable PAHO counts for a specific country. Also, assume that the sequence of updates for each stable PAHO count value is available. In other words, for $\dot{P}_i$ we have the following set:

$$\dot{\mathcal{P}}_i = \left\{P_i^{(1)}, P_i^{(2)}, ..., P_i^{(m)}, ...\right\} \tag{5.12}$$

where $P_i^{(m)}$ is the value of $P_i$ after $m$ weeks of update.

After recognizing high, low, and mid-season months for the country, we can categorize each $\dot{P}_i$ to belong to one of these categories. Then, for category S, an adjustment dataset is constructed named as ${\mathcal{P}_A}^S$ which is defined as follows:

$${\mathcal{P}_A}^S = \left\{(1, P_i^{(1)}, \dot{P}_i, N_i^{(1)}), ..., (m, P_i^{(m)}, \dot{P}_i, N_i^{(m)}), ...\right\} \tag{5.13}$$

Each member of ${\mathcal{P}_A}^S$ is a tuple with four entries: the first entry denotes the time slot that the sample belongs to; the second entry is the actual unstable value of $P_i$; the third entry is the related stable value; and finally, $N_i^{(m)}$ is the size of the statistical population for that week.

In the next step, a linear regression algorithm is used to adjust unstable PAHO values. In order to adjust value of the PAHO values in the $m$th time slot of season S, we use ${\mathcal{P}_A}^S$ set to learn $a_0$, $a_1$, $a_2$, and $a_3$ coefficients in the following equation:

$$\hat{\dot{P}}_i^{(m)} = a_0 + a_1 \times m + a_2 \times P_i^{(m)} + a_3 \times N_i^{(m)} \tag{5.14}$$

where $\hat{\dot{P}}_i^{(m)}$ is the adjusted PAHO count value for the $m$th time slot.

Experimental results show that this adjustment method results in more accurate known PAHO values. Average relative errors of the published unstable PAHO values before and after correction for each country are shown in Figure 3. While in a few cases, we do not experience any improvement, in countries such as Argentina and Paraguay, we experience significant improvements.

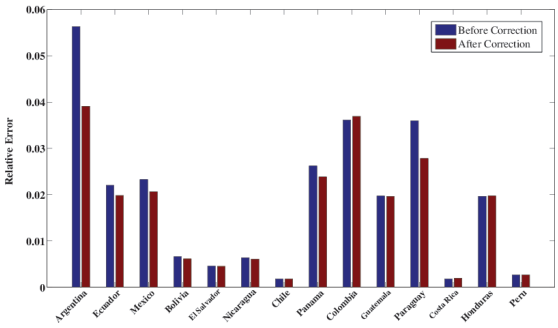



Figure 3: Average relative error of PAHO count values before and after correction for different countries.

Finally, similar to Eq. 5.14, in addition to $P_i^{(m)}$, one can use only time difference $(m)$ or size of population $(N_i^{(m)})$ to correct unstable PAHO values. Effect of these corrections on overall accuracy of predictions are explored in Section 7.

## 6 Experimental Setup

**6.1 Reference Data.** In this paper, we focus on 15 Latin American countries viz. Argentina, Bolivia, Costa Rica, Colombia, Chile, Ecuador, El Salvador, Guatemala, French Guiana, Honduras, Mexico, Nicaragua, Paraguay, Panama and Peru. We collected weekly ILI counts from the official Pan American Health Organization (PAHO) website(`http://ais.paho.org/php/viz/ed_flu.asp`), **every day** from January 2013 to August 2013. The estimates downloaded every day for each country contain data from January 2010 to the latest available week on the day of collection. This dataset is stored in a database we refer to as the *Temporal Data Repository* (TDR). The TDR is also timestamped so that for any given day, we can readily retrieve the ILI case counts that were download on that day. This is important as historic data may be updated by PAHO even a number of weeks after the first update. For the purpose of experimental validation we used the data for the period Jan 2010 to December 2012 as the static training set. We considered Wednesdays of the weeks as a reference day within a week. For each Wednesday from Jan 2013 to July 2013, we used the latest available PAHO data in TDR for that day and predicted 3 weeks from the last available week for which the PAHO data was available. These predictions are next evaluated against the final ILI case count as downloaded on September 1, 2013 and we report the performance of our algorithms in Section 7.

**6.2 Evaluation criteria.** We evaluate the prediction accuracy of the different algorithms using a modified version of percentage relative error:

$$\mathcal{A} = \frac{4}{N_p} \sum_{t=t_s}^{t_e} \frac{|P_t - \hat{P}_t|}{max(P_t, \hat{P}_t, 10)} \tag{6.15}$$

where $t_s$ and $t_e$ indicate the starting and the ending time point for which predictions were generated. $N_p$ indicates the number of time points over the same time period (i.e. $N_p = t_e - t_s + 1$). Note that the measure is scaled to have values in $[0, 4]$ and the denominator is designed to not over-penalize small deviations from the true ILI case count (e.g., when the true case count is 0 and the predicted count is 1). It is to be noted that the accuracy metric so defined is non-convex and is in general multi-modal.

**6.3 Surrogate data sources.** Before describing our data sources in detail, we describe our overall methodology for organizing a flu-related dictionary (for tracking in multiple media such as news, tweets, and search queries).

**6.3.1 Dictionary creation.** The keywords relating to ILI were organized from a seed set of words and expanded using a combination of time-series correlation analysis and pseudo-query expansion. The seed set of keywords (e.g., *gripe*) was constructed in Spanish, Portugese, and English using feedback from our in-house subject mattter expers.

**Pseudo-query expansion.** Using the seed set, we crawled the top 20 web sites (according to Google Search) associated with each word in this set. We also crawled some expert sites such as the official CDC website and equivalent websites of the countries under consideration, detailing the causes, symptoms and treatment for influenza. Additionally we crawled a few hand-picked websites such as `http://www.flufacts.com` and `http://health.yahoo.net/channel/flu_treatments`. We filtered the words from these sites using standard language processing filtering techniques such as stopword removal and Porter stemming. The filtered set of keywords were then ranked according to the absolute frequency of occurrence. The top 500 words for Spanish and English were then selected. For example, words such as *enfermedad* and *pandemia* were obtained from this step.

**Time-series correlation analysis.** Next we used Google Correlate (now a part of Google Trends) to identify keywords most correlated with the ILI case count time-series for each country. Once again these words were found to be a mix of both English and Spanish. As an added step in this process, we also compared time-shifted ILI counts: left-shifted to capture the words searched leading up to the actual flu infection and right-shifted to capture the words commonly searched during the tail of the infection. This entire exercise provided us some interesting terms like *ginger* which has been used as a natural herbal remedy in the eastern world. We also found popular flu medications such as *Acemuk* and *Oseltamivir*, which are also sold under the trade name of *Tamiflu* as highly correlated search queries, especially particularly for Argentina.

**Final filtering.** The set of terms obtained from query expansion and correlation analysis were then pruned by hand to obtain a vocabulary of 151 words. We then performed a final correlation check and retained a final set of 114 words.

**6.3.2 Google Flu Trends ($\mathcal{F}$):** Google Flu Trends (GFT `http://www.google.org/flutrends`) is a tool based on [15] and provided by Google.org which gives weekly and up-to-date ILI case count estimates using search query volumes. Of the countries under consideration, GFT provides weekly estimates for only 6 of them viz. Argentina, Bolivia, Chile, Mexico, Peru and Paraguay. These estimates are typically at a different scale than the ILI case counts provided by PAHO and therefore need to be scaled accordingly. We collected this data weekly on Monday from Jan 2013 to Aug 2013. (The data downloaded on a particular day contains the entire time-series from 2004 to the corresponding week.)

**6.3.3 Google Search Trends ($\mathcal{S}$):** Google Search Trends `http://www.google.com/trends` is another tool provided by Google. Using this tool we can download an estimate of search query volume as a percentage over its own temporal history, filtered geographically. We download the search query volume time series for the 114 keywords described earlier and convert the percentage measures to absolute values using a static dataset we downloaded on Oct 2012 when Google Search Trends used to provide absolute query volumes.

**6.3.4 Twitter ($\mathcal{T}$):** Twitter data was collected from Datasift.com and geotagged using an in-house geocoder. We lemmatized the tweet contents and used language detection and POS tagging to help differentiate relevant from irrelevant uses of our keywords (e.g., the Spanish word *gripe*, meaning flu, is part of our flu keyword list as opposed to the undesired and unrelated English word 'gripe'). The resulting analysis yields a weekly occurrence count of our dictionary in tweets.

**6.3.5 HealthMap ($\mathcal{H}$):** Similar to Twitter data, we also collect flu-related news stories using HealthMap `http://healthmap.org`, an online global

disease alert system capturing outbreak data from over 50,000 electronic sources. Using this service we receive flu-related news as a daily feed which is similarly enriched and filtered to obtain a multivariate time series over lemmatized version of the keywords. While Twitter is more suitable to ascertain general public response, the HealthMap data provides more detailed information but may capture the trends at a slower rate. Thus each of these sources offers utility in capturing different surrogate signals: Twitter offers leading but noisy indicators whereas HealthMap provides a slightly delayed but more reliable indicator.

**6.3.6 OpenTable ($\mathcal{O}$):** We also use data on trends of restaurant table reservations, initially studied in [17] to be a potential early indicator for outbreak surveillance, as another surrogate for ILI detection. This novel data stream is based on the postulate that a higher than average number of restaurants with table availabilities in a region can serve as an indicator of an event of interest, such as increase in flu cases. Table availability was monitored using OpenTable `http://www.opentable.com`, an online restaurant reservation site with 28,000 restaurants at the time of this writing. Daily searches were performed starting from September 2012 for a table for two persons at lunch and dinner; between 12:30-3pm, and between 6-10:30pm. Data was collected for Mexico by city (Cancun, Mexico City, Puebla, Monterrey, and Guadalajara) and for the entire country. The daily proportion (proportion used due to changes in the number of restaurants in the system) of restaurants with available tables was aggregated as a weekly time-series.

**6.3.7 Weather ($\mathcal{W}$):** All of the previously described data sources can be termed as non-physical indicators which can work suitably as indirect indicators about the state of the population with respect to flu by exposing different population characteristics. On the other hand, meteorological data can be considered a more direct and physical driver of influenza transmission [18]. It has been shown in [9, 10, 8] that absolute humidity can be directly used to predict the onset of influenza epidemics. Here, we collect several other meteorological indicators such as temperature and rainfall in addition to humidity from the Global Data Assimilation System (GDAS). We accessed this data in GRIB format from `http://ladsweb.nascom.nasa.gov/` at a resolution of 1 degrees lat/long interval. However, looking at all the lat/long for a country can often lead to noisy data. As such we filtered the downloaded data and used the indicators only around the surveillance centers. We also aggregate this data using weekly averages and thus obtain a resultant time series for each country. We collected this data weekly from Jan 2013 to August 2013.

Table 5: ILI case count prediction accruacy for Mexico using OpenTable data as a single source, and by combining it with all other sources using model level fusion on uncorrected ILI case count data.

| Method | Lunch | Dinner | Lunch & Dinner |
|---|---|---|---|
| MF | 1.92 | 2.23 | 2.31 |
| NN | 1.99 | 1.83 | 2.11 |
| MFN | 2.11 | 2.31 | 2.44 |
| Model Fusion | 2.96 | 2.87 | 2.99 |

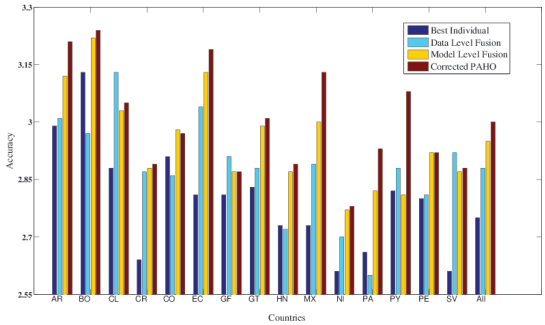



Figure 4: Accuracy of different methods for each country.

## 7 Results

We present an exhaustive set of experiments evaluating our algorithms over 6 months of predictions from Jan 2013 to August 2013. The final estimates of ILI case counts are stable according to the estimates downloaded from PAHO on Oct 1, 2013. All models considered here were used to forecast 2 weeks beyond the latest available PAHO ILI estimates. Key findings are presented in Table. 1. We analyze some important observations from this table next.

**Can we 'beat' Google Flu Trends with our custom dictionary?** The key difference between Google Flu Trends and Google Search Trends is that the former uses a closed dictionary whereas we constructed the dictionary to use with GST. As can be seen Table 1, for majority of the common countries (countries for which data from both GST and GFT is present), regressors running on GST consistently outperform those running on GFT (with Mexico and Peru being the exception). Thus we posit that the GST model devised here is a sufficiently close approximation to GFT, with the added advantages of having access to raw level data and being available for more countries than GFT (among the 15 countries we consider, only 6 of them are present in the GFT database).

**Which is the optimal regression model?** From Table 1, we can also analyze the three different regressors proposed in Section 3.1 with respect to overall accuracy. With respect to each individual source, we can see that matrix factorization with nearest neighbor embdedding (MFN) performs the best in average over the

Table 1: Comparing forecastng accuracy of models using individual sources.

| Model | Sources | AR | BO | CL | CR | CO | EC | GF | GT | HN | MX | NI | PA | PY | PE | SV | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MF | $\mathcal{W}$ | 2.78 | 2.46 | 2.39 | 2.14 | 2.70 | 2.22 | 2.12 | 2.63 | 2.52 | 2.73 | 2.31 | 2.21 | 2.49 | 2.77 | 2.61 | 2.47 |
| | $\mathcal{H}$ | 2.81 | 2.31 | 2.22 | 1.92 | 2.43 | 2.04 | 2.11 | 2.57 | 2.33 | 2.48 | 2.39 | 2.15 | 2.18 | 2.47 | 2.33 | 2.32 |
| | $\mathcal{T}$ | 2.37 | 2.35 | 2.18 | 2.03 | 2.21 | 2.12 | 1.83 | 2.12 | 2.29 | 2.03 | 1.89 | 2.06 | 1.96 | 2.20 | 2.21 | 2.12 |
| | $\mathcal{F}$ | 2.34 | 2.11 | 2.29 | N/A | N/A | N/A | N/A | N/A | N/A | 2.71 | N/A | N/A | 2.31 | 2.24 | N/A | 2.33 |
| | $\mathcal{S}$ | 2.48 | 2.21 | 2.33 | 2.04 | 2.31 | 2.21 | 1.93 | 2.03 | 2.15 | 2.51 | 2.42 | 2.52 | 2.33 | 1.93 | 2.30 | 2.24 |
| NN | $\mathcal{W}$ | 2.92 | 2.93 | 2.63 | 2.52 | 2.66 | 2.51 | 2.71 | 2.82 | 2.59 | 2.62 | 2.55 | 2.59 | 2.61 | 2.80 | 2.52 | 2.66 |
| | $\mathcal{H}$ | 2.73 | 3.10 | 2.42 | 2.27 | 2.83 | 2.64 | 2.43 | 2.25 | 2.71 | 2.31 | 2.61 | 2.35 | 2.43 | 2.39 | 2.52 | 2.53 |
| | $\mathcal{T}$ | 2.72 | 2.86 | 2.31 | 2.62 | 2.77 | 2.52 | 2.71 | 2.66 | 2.51 | 2.44 | 2.13 | 2.01 | 1.77 | 2.51 | 2.20 | 2.45 |
| | $\mathcal{F}$ | 2.11 | 2.21 | 2.33 | N/A | N/A | N/A | N/A | N/A | N/A | 2.19 | N/A | N/A | 2.41 | 2.32 | N/A | 2.26 |
| | $\mathcal{S}$ | 2.51 | 2.31 | 2.41 | 1.81 | 2.52 | 2.41 | 2.12 | 2.29 | 2.51 | 2.13 | 2.61 | 2.14 | 2.51 | 1.87 | 2.12 | 2.28 |
| MFN | $\mathcal{W}$ | 2.99 | 3.01 | 2.88 | 2.53 | 2.78 | 2.81 | 2.77 | 2.83 | 2.61 | 2.70 | 2.56 | 2.66 | 2.82 | 2.79 | 2.51 | 2.75 |
| | $\mathcal{H}$ | 2.81 | 3.13 | 2.63 | 2.58 | 2.91 | 2.77 | 2.57 | 2.63 | 2.73 | 2.50 | 2.61 | 2.54 | 2.51 | 2.69 | 2.61 | 2.68 |
| | $\mathcal{T}$ | 2.74 | 3.03 | 2.51 | 2.64 | 2.83 | 2.51 | 2.81 | 2.71 | 2.60 | 2.48 | 2.13 | 2.55 | 2.19 | 2.57 | 2.31 | 2.57 |
| | $\mathcal{F}$ | 2.33 | 2.41 | 2.34 | N/A | N/A | N/A | N/A | N/A | N/A | 2.69 | N/A | N/A | 2.54 | 2.48 | N/A | 2.46 |
| | $\mathcal{S}$ | 2.61 | 2.44 | 2.55 | 2.22 | 2.61 | 2.52 | 2.71 | 2.31 | 2.62 | 2.48 | 2.61 | 2.31 | 2.53 | 2.23 | 2.13 | 2.46 |

Table 2: Comparison of prediction accuracy while combining all data sources and using MFN regression.

| Fusion Level | AR | BO | CL | CR | CO | EC | GF | GT | HN | MX | NI | PA | PY | PE | SV | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | 3.12 | 3.22 | 3.03 | 2.88 | 2.98 | 3.13 | 2.87 | 2.99 | 2.87 | 3.00 | 2.77 | 2.82 | 2.81 | 2.92 | 2.87 | 2.95 |
| Data | 3.01 | 2.97 | 3.13 | 2.87 | 2.86 | 3.04 | 2.91 | 2.88 | 2.72 | 2.89 | 2.70 | 2.60 | 2.88 | 2.81 | 2.92 | 2.88 |

Table 3: Comparison of prediction accuracy while using model level fusion on MFN regressors and employing PAHO stabilization.

| Correction Method | AR | BO | CL | CR | CO | EC | GF | GT | HN | MX | NI | PA | PY | PE | SV | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| None | 3.12 | 3.22 | 3.03 | 2.88 | 2.98 | 3.13 | 2.87 | 2.99 | 2.87 | 3.00 | 2.77 | 2.82 | 2.81 | 2.92 | 2.87 | 2.95 |
| Weeks Ahead | 3.15 | 3.24 | 3.04 | 2.87 | 2.97 | 3.17 | 2.87 | 2.99 | 2.88 | 3.05 | 2.77 | 2.91 | 3.02 | 2.91 | 2.88 | 2.98 |
| Num samples | 3.20 | 3.24 | 3.03 | 2.88 | 2.96 | 3.12 | 2.87 | 3.01 | 2.89 | 3.12 | 2.78 | 2.92 | 3.04 | 2.91 | 2.87 | 2.99 |
| Combined | 3.21 | 3.24 | 3.05 | 2.89 | 2.96 | 3.19 | 2.87 | 3.00 | 2.89 | 3.13 | 2.77 | 2.93 | 3.08 | 2.92 | 2.88 | 3.00 |

Table 4: Discovering importance of sources in Model level fusion on MFN regressors by ablating one source at a time.

| Sources | AR | BO | CL | CR | CO | EC | GF | GT | HN | MX | NI | PA | PY | PE | SV | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All | 3.21 | 3.24 | 3.05 | 2.89 | 2.96 | 3.19 | 2.87 | 3.00 | 2.89 | 3.13 | 2.77 | 2.93 | 3.08 | 2.92 | 2.88 | 3.00 |
| w/o $\mathcal{W}$ | 2.91 | 2.99 | 2.77 | 2.71 | 2.61 | 2.59 | 2.66 | 2.69 | 2.49 | 2.78 | 2.62 | 2.87 | 2.60 | 2.43 | 2.67 | 2.69 |
| w/o $\mathcal{H}$ | 3.04 | 2.85 | 2.89 | 2.56 | 2.81 | 2.77 | 2.61 | 2.75 | 2.75 | 2.82 | 2.57 | 2.75 | 2.51 | 2.87 | 2.71 | 2.75 |
| w/o $\mathcal{T}$ | 2.92 | 3.14 | 2.95 | 2.61 | 2.72 | 2.81 | 2.88 | 2.79 | 2.61 | 2.93 | 2.74 | 2.63 | 2.79 | 2.74 | 2.81 | 2.80 |
| w/o $\mathcal{S}$ | 3.19 | 3.11 | 2.92 | 2.64 | 2.69 | 2.70 | 2.89 | 2.88 | 2.78 | 3.07 | 2.75 | 2.91 | 2.80 | 2.71 | 2.86 | 2.86 |
| w/o $\mathcal{F}$ | 3.20 | 3.12 | 2.88 | 2.89 | 2.96 | 3.19 | 2.87 | 3.00 | 2.83 | 3.02 | 2.77 | 2.93 | 2.98 | 2.88 | 2.88 | 2.96 |

countries. For some countries such as Panama, when using only Google Search Trends, MFN performs poorer than vanilla MF; nevertheless the average accuracy over all countries for any given data source is best when using MFN.

**Which is the best strategy to combine multiple data sources?** As shown in Table 2, in overall, model level fusion works better than data level fusion. For 8 of the 15 countries, model level fusion works appreciably better than data level fusion, while the reverse trend is seen for 4 other countries. This showcases the importance of considering both kinds of fusion depending on the country of interest.

**How effective are we at forecasting a moving PAHO target?** As shown in Table 3, our corrected estimates using both the number of samples and the *weeks ahead* from the upload date are generally better. It is instructive to note that our correction strategy is able to increase the overall accuracy only by a score of approximately 0.05 over all the countries, for some countries such as Mexico and Argentina (for which the data update is typically noisy) we obtain a substantial imporovement of scores. This suggests that the correction strategy may be selectively applied when forecasting for certain countries.

**How do physical vs social indicators fare against each other?** From Table 1, we see that the data source with the best single accuracy happens to be the physical indicator source, i.e., weather data. However, Table 4 conveys a mixed story. Here we conduct an *ablation test*, wherein we remove one data source at a time from our model level MFN fusion framework and contrast accuracies. While removing the weather data degrades the accuracy score the most, removing the social indicators also degrades the score to varying degrees. Thus we posit that it is important to consider both the physical and social indicators to get a refined signal about the prevalent ILI incidence in the population.

**How relevant is restaurant reservation data to forecasting ILI?** All the results thus far do not consider the OpenTable reservation data, since this source is available only for Mexico (among the countries studied here). We considered table availability for different time ranges and compared performance using our MFN model. As Table 5 demonstrates, we obtain the best performance when considering both lunch and dinner

reservation data. Nevertheless, we have observed that including this source as part of the ensemble decreases the overall accuracy by 0.01 over the uncorrected ILI case count data. Thus it is our opinion that although the reservation data could exhibit some signals about prevalent ILI conditions, it likely is also a surrogate for non-health conditions (e.g., social unrest) which must be factored out to make the data source more useful.

Finally we present Figure 4 where we compare for each country the accuracies of prediction form the best individual source, with the theose from booth data level and model level fusion of the different sources and the the model level fusion of MF regressors applied on the corrected PAHO estimates rather than the raw ones. As can be seen, we progresively increase our accuracies with the corrected PAHO estimates providing the final increase in predictive power to our model level fusion framework.

## 8 Conclusions and Further Work

To forecast ILI over a range of Latin American countries, we have explored a gamut of options pertaining to data sources, fusion possibilities, and corrections to track a moving target. Our results demonstrate that there are significant opportunities to improve forecasting performance and selective superiorities among data sources that can be leveraged. Our future work focuses on reconciling the phenomenological models developed here with true epidemiological models to that we can develop not just near-term forecasts as done here but also identify long-range characteristics of the epidemic as it unfolds.


## Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.


## References


[1] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in china with search query from baidu," *PlosOne*, vol. 8, no. 5, p. e64323, 2013.

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant1, "Influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[3] K.Denecke, P.Dolog, and P.Smrz, "Making use of social media data in public health," in *Proceedings of WWW '12*, 2012, pp. 243–246.

[4] K. Lee, A.Agrawal, and A.Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the KDD '13*, 2013, pp. 1474–1477.

[5] N.Kanhabua and W.Nejdl, "Understanding the diversity of tweets in the time of outbreaks," in *Proceedings of WWW '13*, 2013, pp. 1335–1342.

[6] C. Chew and G. Eysenbach, "Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak," *PlosOne*, vol. 5, no. 11, p. e14118, 2013.

[7] R. Sugumaran and J.Voss, "Real-time spatio-temporal analysis of west nile virus using twitter data," in *Proceedings of COM.Geo '12*, 2012, pp. 1335–1342.

[8] J. D. Tamerius, J. Shaman, W. J. Alonso, K. Bloom-Feshbach, C. K. Uejio, A. Comrie, and C. Viboud, "Environmental predictors of seasonal influenza epidemics across temperate and tropical climates," *PLoSPathog*, vol. 9, no. 3, pp. 68–72, 2013.

[9] J. Shaman, E. Goldstein, and M. Lipsitch, "Absolute Humidity and Pandemic Versus Epidemic Influenza," *American journal of epidemiology*, vol. 173, no. 2, pp. 127–135, 2010.

[10] J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, and M. Lipsitch, "Absolute humidity and the seasonal onset of influenza in the continental United States." *PLoS biology*, vol. 8, no. 2, p. e1000316, 2010.

[11] P. Kostkova, "A roadmap to integrated digital public health surveillance: the vision and the challenges," in *Proceedings of WWW '13*, 2013, pp. 687–694.

[12] A. Apolloni, V. Kumar, M. Marathe, and S. Swarup, "Computational epidemiology in a connected world," *Computer*, vol. 42, no. 12, pp. 83–86, 2009.

[13] J. Canny, "Collaborative filtering with privacy via factor analysis," in *Proceedings of SIGIR '02*, 2002, pp. 238–245.

[14] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of KDD '08*, 2008, pp. 426–434.

[15] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.

[16] V. Chase, "Promed: a global early warning system for disease." *Environmental Health Perspectives*, vol. 104, no. 7, p. 699, 1996.

[17] E. O. Nsoesie, D. L. Buckeridge, and J. S. Brownstein, "Who's not coming to dinner? evaluating trends in online restaurant reservations for outbreak surveillance," *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.

[18] W. Yang, S. Elankumaran, and L. C. Marr, "Relationship between humidity and influenza A viability in droplets and implications for influenza's seasonality." *PloS one*, vol. 7, no. 10, p. e46789, 2012.