Data Integration and Predictive Analysis System for Disease Prophylaxis: Incorporating Dengue Fever Forecasts

John Freeze
Knowledge Based Systems, Inc.
jfreeze@kbsi.com

Madhav Erraguntla
Knowledge Based Systems, Inc.
merraguntla@kbsi.com

Akshans Verma
Knowledge Based Systems, Inc.
akverma@kbsi.com

Abstract

The goal of the Data Integration and Predictive Analysis System (IPAS) is to enable prediction, analysis, and response management for incidents of infectious diseases. IPAS collects and integrates comprehensive datasets of previous disease incidents and potential influencing factors to facilitate multivariate, predictive analysis of disease patterns, intensity, and timing. We have used the IPAS technology to generate successful forecasts for Influenza Like Illness (ILI). In this study, IPAS was expanded to forecast Dengue fever in the cities of San Juan, Puerto Rico and Iquitos, Peru. Data provided by the National Oceanic and Atmospheric Administration (NOAA) was processed and used to generate prediction models. Predictions were developed with modern machine learning algorithms, identifying the one-week and four-week forecast of Dengue incidences in each city. Prediction model results are presented along with the features of the IPAS system.

1. Introduction

A plethora of diseases create a continuously evolving threat to public health. Some diseases experience seasonal cycles, resurfacing every year or every few years around the same time. For example, in some regions of the world the prevalence of Dengue fever jumps every couple of years. The prevalence of influenza fluctuates consistently enough to earn the phrase 'flu season.' Other diseases such as the Zika and Ebola may be rare until they experience unexpected outbreaks.

Spread of these diseases could be prevented with appropriate disease management measures if they could be predicted before they occurred. Foresight into the location, timing, peak intensity, and potential number of infected individuals could enable public health officials to take proactive disease containment and management efforts [1, 2, 3]. Health organizations such as the CDC have recognized this fact, and have sponsored several competitions and workshops to encourage development of viable prediction models [7, 8, 13].

Accurate prediction of disease outbreak enables effective disease-management activities. These activities, such as coordinating medical personnel and pharmaceutical resources in advance of outbreaks, can reduce the severity of the outbreaks and prevent hospital overcrowding. Forewarning of outbreaks can also enable development of medical interventions, prophylaxis to disease hazards, and containment of disease vectors.

Traditional epidemiology has focused on system dynamics, network, and compartmental models-based approaches (susceptibility, exposed, infected, recovered (SEIR)) for forecasting disease progression [4, 5, 6]), which require intensive data analysis and modeling by field experts. IPAS takes advantage of innovative machine learning and predictive analytics to facilitate the creation of generic disease prediction models that require greatly reduced user interaction.

Other researchers have analyzed prior data to predict outbreaks of disease in the future. Investigators have studied the link between disease outbreaks and weather [14], consumption of bushmeat [9], socioeconomic status [10, 11], and an array of other factors. Collection and curation of data is critical to these disease-outbreak predictions that are based on observed trends in historical data. Required data include quantitative and qualitative data related to disease occurrences, weather and environmental data, and vector data, among others (see Figure 1).

Recognizing the importance of data collection, research groups like HealthMap [12] and VectorMap [13] have curated valuable datasets; however, these datasets are isolated. Researchers attempting to leverage these diverse, isolated datasets must invest time merging these data sets before using them. These researchers must wrestle with issues with data quality, inconsistent and incompatible data, and mismatched terminology. To merge the data, researchers perform ad-hoc, proprietary, behind the scenes data manipulations that are neither transparent nor scalable.

IPAS provides fundamental approaches to data cleaning and spatial and temporal harmonization which

URI: http://hdl.handle.net/10125/50001
ISBN: 978-0-9981331-1-9

allow researchers to integrate and analyze the collected data. Spatial harmonization includes creating regional and national environmental data from individual sensor data. Temporal harmonization includes conversion of data from one time scale to another, for example, transforming daily data into weekly data. Supported core processing steps include spatio-temporal clustering, correlation analysis, and determination of the factors influencing the spread of a disease.

There have also been a slew of attempts to use machine learning techniques to predict the next disease outbreak [25, 27]. For example, research at the University of Punjab developed techniques to predict Dengue fever using Naïve Bayesian, Reduced Error Pruning (REP), REP Tree, Random tree, J48, and Sequential Minimal Optimization (SMO) models. However, most of these research efforts do not have support for gathering and preprocessing data, performing predictive analytics, and generating visualizations oriented to public health stakeholders. IPAS provides the requisite analytics within a robust framework for end-to-end epidemiological analysis.

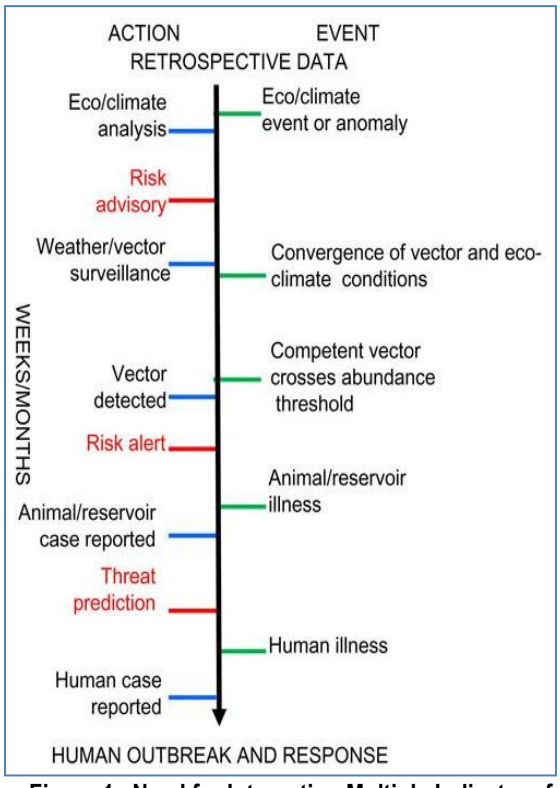


Figure 1. Need for Integrating Multiple Indicators for Effective Predictive Models[2]

IPAS leverages machine learning and predictive analytics-based techniques such as kernel-based support vector machines, nearest-neighborhood calculation, decision trees, random forests, and boosted trees for the purposes of disease incidence prediction [9, 10]. We have previously presented the results of application of these analytical models to predict the ILI across the USA and within Health and Human Services (HHS) regions [24]. In this paper, we present the application of IPAS technology to Dengue fever forecasting.

Accurate forecasting of disease outbreaks means little if decision makers and public health officials are not appropriately informed. Currently, analysts at the CDC, DoD, and WHO create and share weekly disease biosurveillance reports to summarize and communicate the state of the focus disease [19, 26]. These reports summarize the disease, reported incidents, and detected patterns and observations (see Figure 2). While this activity involves analysis and insights from the epidemiological experts, most of the analyst time is spent on collecting and aggregating the necessary information.

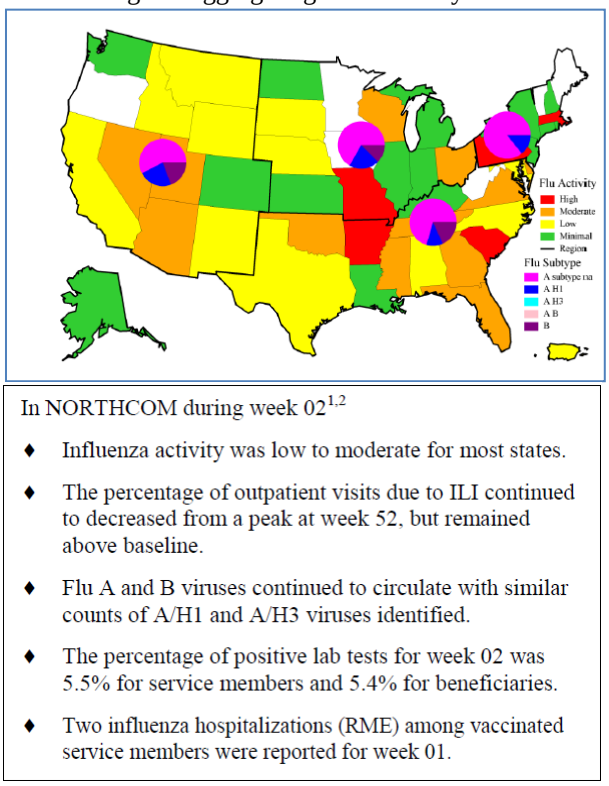


Figure 2. Sample Influenza Biosurveillance Report

IPAS could reduce the burden on biosurveillance and epidemiology analysts who currently spend thousands of hours each month on collecting and aggregating the necessary information. IPAS automates the collection and aggregation of relevant disease data to facilitate comprehensive epidemiological analysis. Natural language processing is used to extract specific disease, syndromic, and zoonotic details from sources such as

news feeds, public health reports, medical publications, and social media. The mined data is used in disease analysis and the creation of biosurveillance reports such as the one shown in Figure 2.

2. Dengue-Related Data Set

A set of Dengue-related data was provided by the NOAA. This data contains Dengue incidents from San Juan, Puerto Rico and Iquitos, Peru, as well as climatological variables for both of these cities [23].

2.1. Base Data

The Dengue data set describes the weekly incidences of Dengue at two cities: Iquitos, Peru, and San Juan, Puerto Rico. The incidence data set for Iquitos, Peru ranges from years 2000-2013, and the set for San Juan, Puerto Rico ranges from 1990-2013. The features of this data set describe when and how often a specific strain of Dengue occured for both locations (see Figure 3, Figure 4).

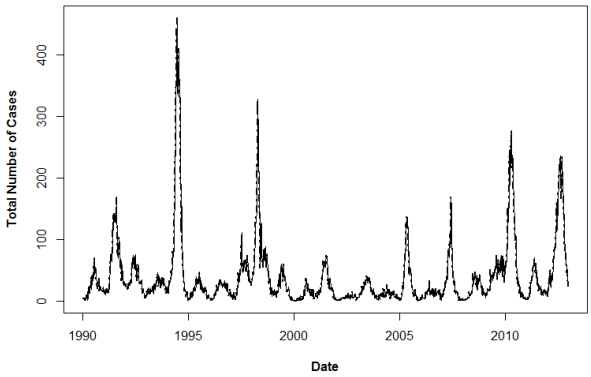


Figure 3. Dengue Cases in San Juan, Puerto Rico

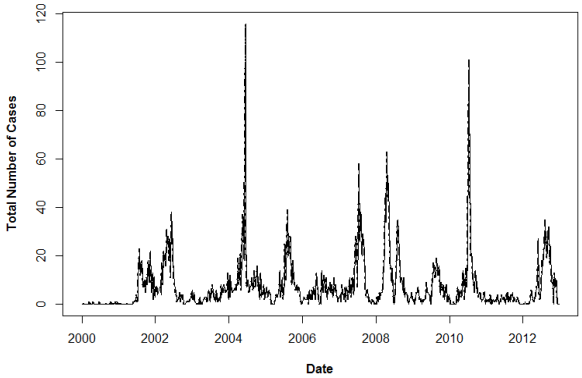


Figure 4. Dengue Cases in Iquitos, Peru

Population data for both cities is also available. For Iquitos, the data consists of the population of the four districts in the metropolitan area of Iquitos from 2000-2014. For San Juan, it consists of the population of the San Juan-Carolina Metropolitan Statistical Arena for 1990 and 1999-2014.

Environmental data is available in the form of weather station data. Iquitos has records of the daily temperature and precipitation from 1979-2015. San Juan has records of that data but for 1956-2015. Satellite data provides the daily precipitation data from both cities from 1983-2014. Temperature and precipitation data were reanalyzed to focus on the daily relative humidity, specific humidity, dew point, and temperature. Data from both cities is available from 1979-2015.

Finally, satellite data on vegetation was collected from the Weekly Normalized Difference Vegetation Index (NDVI) derived from the NOAA Climate Data Record (CDR) of Advanced Very High Resolution Radiometer (AVHRR) Surface Reflectance. Data from both cities is available from 1981-2013.

2.2. Feature Extraction and Data Processing

Significant data processing and feature extraction was performed on the data. The first step was to aggregate all daily data into "weekly" data, aggregated across seven consecutive days. The first week of each year started on January 1 and ended on January 7, and all further weeks followed the same pattern. Using this definition, weeks from the same year all start on the same day.

Some of the daily data was not available for all days, especially in the weather data set; missing data were left blank. Non-missing weather-related data was averaged over each week during aggregation. It was assumed that data for missing days was comparable to reported days. If any week contained no weather data for any of the seven days, that entry was left empty. Any week containing a blank entry, and associated data for that week, was removed from the final dataset for model generation. All of the weather-based data was converted to a weekly time frame.

Incidence data was normalized to incidence per hundred thousand people, based on annual population data. For the years 1991-1998 that were missing in San Juan, for which we used population data from 1990. Next, the percent change in incidence (slope) for one, two, three, and four week horizons was calculated. For weeks with no incidence reported, the slope was set to 0.

Next, cumulative values were calculated on both a yearly basis and on a fixed horizon. Cumulative yearly incidences were aggregated, starting from the first week of the year and summing through the current week at each data point. Fixed horizon cumulative incidences were calculated on four- and eight-week horizons,

summing the Dengue incidences from the current week with data from the prior three or seven weeks, respectively.

Nonlinear parameters were introduced by taking the square and cube of a subset of the parameters. These nonlinear terms include the total incidence, slopes, and cumulative cases observed. Inclusion of these terms introduces nonlinear effects into the models.

It was observed that Dengue outbreaks have a yearly cycle, and exhibit severe outbreaks every three to five years. To capture this phenomenon, we included "Time since last severe outbreak" as a predictor variable in the model. To begin, the highest incidence of each year was calculated for each location. Then, an outbreak threshold was defined as the seventieth percentile of a yearly peak incidences for each location. All weeks with incidence above the location-specific outbreak threshold were identified. Then, the number of weeks since the most recent outbreak was calculated for each data point.

3. Prediction Modeling

Once the aforementioned features were generated, the data from both locations were combined into a single dataset. The combined data was then separated into a training set and a test data set based on years. The test data set spans the Dengue seasons starting in 2010-2012 for both locations. The data for the remaining years (2000-2009 for Iquitos, and 1990-2010 for San Juan) were used to train the models.

Before the final model was generated, a linear regression model was used to determine the significance of each of the features in the data set. Non-significant features were removed from the final prediction models. Excluded features included most of the weather data, although the minimum temperature and mean precipitation were still included.

Several prediction models were generated to provide forecasts of two modeling goals: one-week and four-week look ahead predictions. These models were generated in R using linear regression (with non-linear terms), Support Vector Machine (SVM), Random Forests (RF), and Boosting (B) modeling techniques. The *caret* package[1] provided convenient access to model generation for each of these models. Additionally, the *caret* package includes support for n-fold cross validation.

For the four models generated here, 10-fold cross validation was implemented parameter tuning in each model. Each model was trained and tested using the same training and testing data sets. To simulate real-time predictions, all parameters selected for use in prediction were constrained to the list of data available on or before the current week; prior weeks' data could be used for prediction, but no forward-looking features were used.

All models performed similarly, although the Random Forest model showed the best performance based on Mean Square Error (MSE) over the test data set for both prediction goals (see Table 1). The residuals were compared visually to fitted values for the training data to verify that no visible residual correlations are present (see Figure 5). Then, the fitted values were compared to truth for one-week and four-week forecast models (see Figure 6). While all models were generated, only Random Forest plots are shown here for the sake of brevity.

Table 1. Model Performance on Test Data for 1- and 4-week Look Ahead

Model Type	Mean Square Error	
	1-week forecast	4-week forecast
Regression	1.77	5.00
Support Vector Machine	1.84	5.01
Random Forest	1.47	4.71
Boosting	1.70	5.28

Figure 5 shows a tight cluster of points near the origin because most of the weekly incidence counts are low. As incidences increase, residual variance increases, pointing to heteroscedasticity in the model. The effect is small, especially when the outlier (Fitted Value near 20) is ignored. This effect points to difficulties in the models' ability to forecast extreme values. In our future effort, we will explore log data transformation to address the observed heteroscedasticity in the residuals.

As can be expected, the four-week forecast model performs worse than the one-week forecast model for all model types, showing a higher MSE value. The increased lead time of four-week forecasts results in increased variability that is not fully captured by the existing models. Boosting models showed relatively poor performance compared to the other models shown, likely due to the required exclusion of categorical variables. It is expected that including these variables would improve boosting model performance.

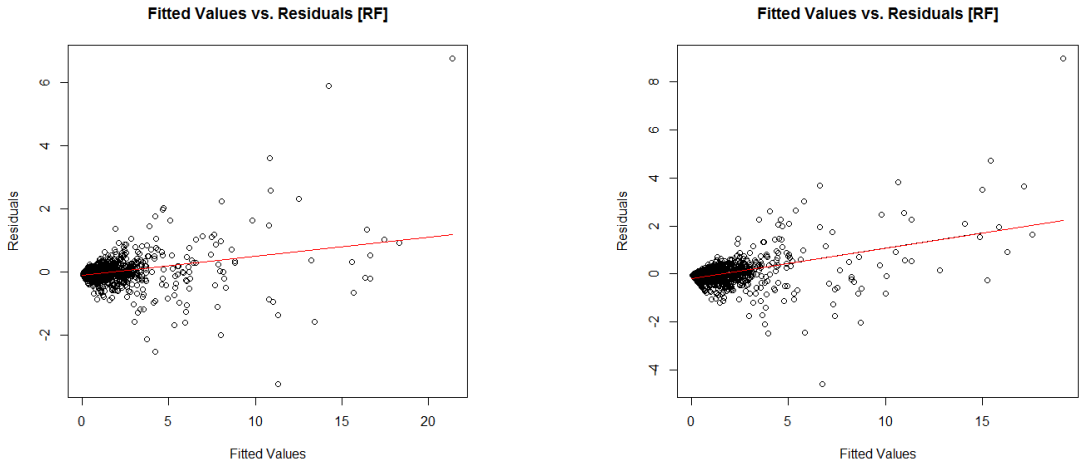
[1]https://cran.r-project.org/web/packages/caret/caret.pdf

Significant factors were observed for the models, showing a few trends. The week number was significant, implying a seasonal effect in Dengue outbreaks. The total number of cases and the square and cube of the total cases were all significant, indicating the nonlinear nature of the models. A previous weeks' incidences and the percent change in incidences from previous weeks were both significant, showing the importance of trend. Temperature and daily temperature range features were significant, highlighting the importance of weather in Dengue outbreaks. The region feature was still significant, suggesting the existence of other factors that were not included here.

All models had trouble forecasting the severe outbreak for Iquitos in 2010 (see Figure 6). The one-week Random Forest model was able to predict a severe outbreak in 2010 for Iquitos, but did not fully capture the magnitude of the outbreak. The four-week forecast using Random Forest was able to predict an increase, but did not fully capture the timing or the peak of the outbreak. The model performed well at non-severe outbreak times for both forecast horizons. Future efforts will incorporate features to enable better predictions for severe outbreaks by detecting patterns that can provide early warning indicators for severe, outlier outbreaks.

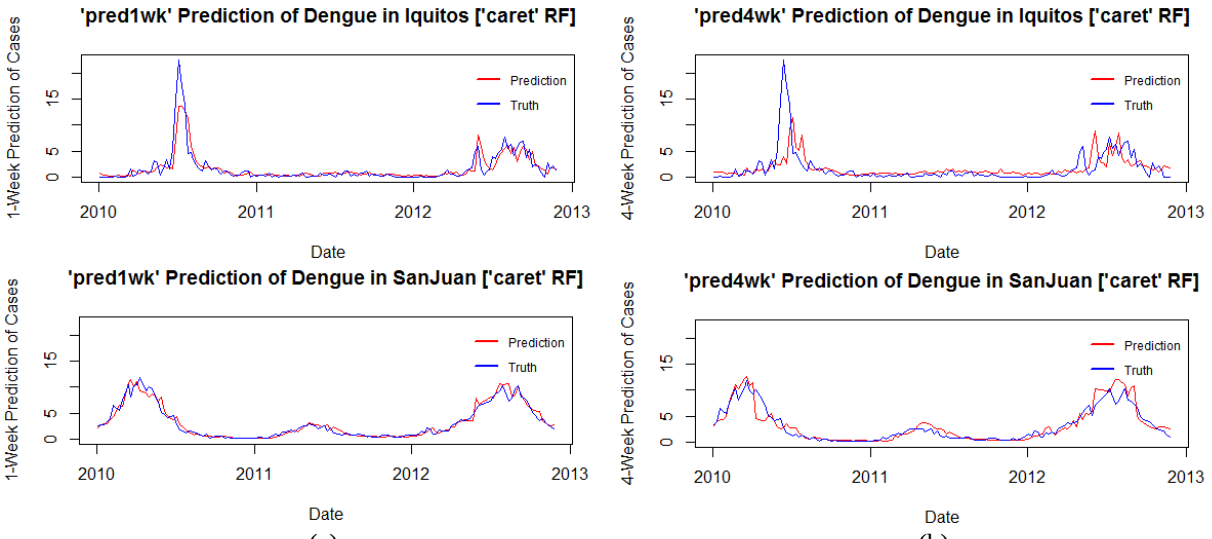
4. IPAS Disease Management Application

The IPAS Application aggregates data gathered through a natural language processing (NLP) information extraction engine. This component is used



(a) (b)

Figure 5. Residual Comparison in Random Forest Model for (a) 1-Week and (b) 4-Week Forecast



(a) (b)

Figure 6. (a) 1-Week and (b) 4-Week Forecasts of Dengue Incidents for Random Forest Model

to extract relevant meaningful information, such as disease outbreaks, from unstructured data and render it in a format useful for analysis [14, 15]. The flexibility of this tool enables the incorporation of disease, syndromic, and zoonotic details from news feeds, public health reports, and medical publications. Prediction models and surveillance reports are generated using this data.

Once prediction models have been developed, the historical data and applicable predictions are portrayed to stakeholders in a meaningful way. Although the models shown here forecast Dengue incidences, the IPAS application enables the exploration of other diseases in general. The IPAS application is intended to provide this functionality through three primary views:

1) Situational Awareness View
2) Explore and Analyze View
3) Predictions View

4.1. Situational Awareness View

The situational awareness view of the IPAS application provides the high level overview of the global disease infections that have occurred in the past (Figure 7). This view is designed to support epidemiologists, and public health analysts to:

- View the history of disease at a particular region (temporal perspective of a disease in a region).
- Determine which diseases occur in a given region (spatial perspective of a region).
- Observe the spread or migration of disease (temporal perspective of a disease).

The situational awareness view consists of three different visual interfaces, which correspond to the three main dimensions of the epidemic data: disease type (top left), location (top right), and time (bottom). All the different components of the situational awareness view are integrated to provide the user with "brushing and linking" features. The idea of brushing and linking is to facilitate visual analytics and let the end user choose the focus disease, location, or time and adjust the display accordingly. Interactive changes made in one interface are automatically reflected in other interfaces.\

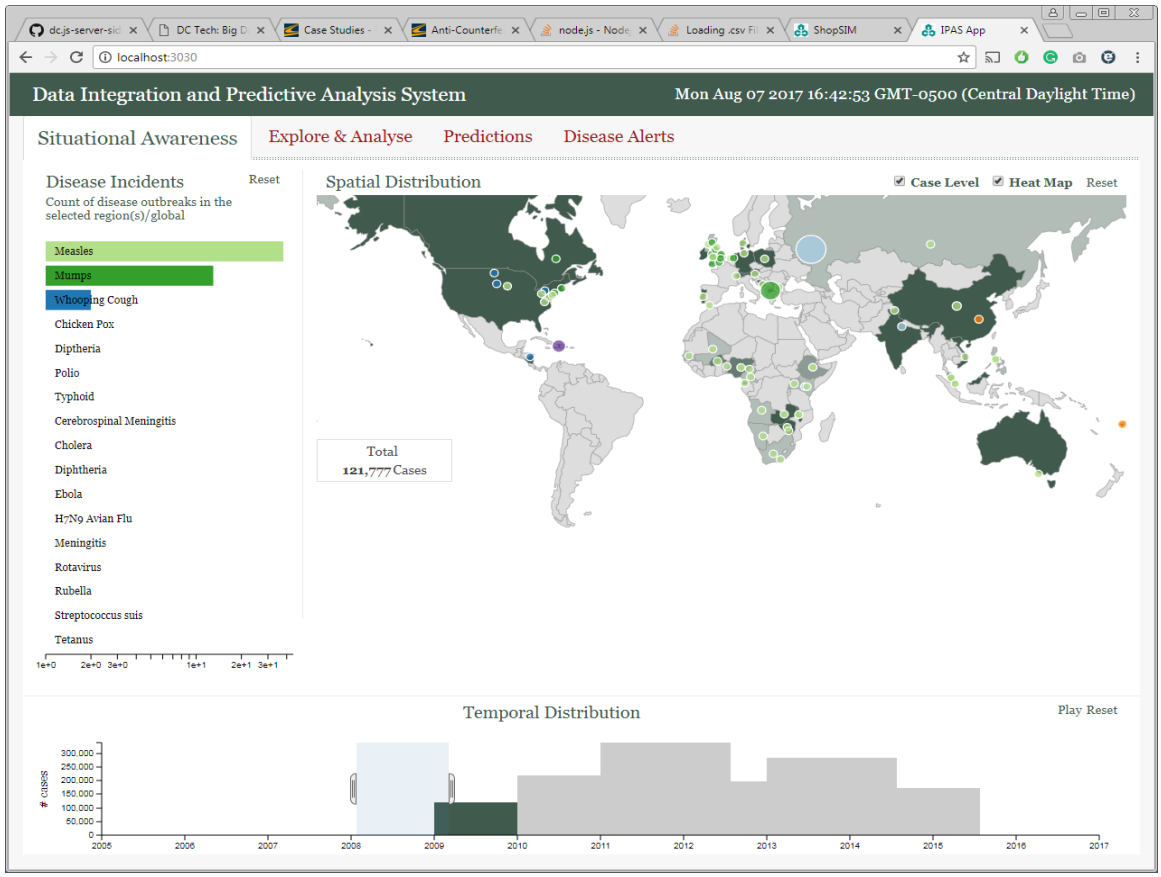


Figure 7. Situational Awareness View of IPAS Application

Connecting multiple interfaces through interactive brushing and linking provides the ability to the end users to creatively drill-down, explore, and analyze the data. The disease incidents map on the top left side provides the histogram of all diseases, with the length of the chart representing the number of reported cases of that disease (Figure 7).

The spatial distribution map on the top helps the analyst to understand the location of the selected disease outbreak. The location of the bubble on the map shows where the cases are recorded and the color corresponds to the disease on the disease incidents map on the top left. The bigger the bubble, the higher the number of cases recorded. Along with the bubble, countries are color coded based on the relative distribution of the selected disease(s) among countries. The user can get additional details about each bubble by hovering the mouse over the bubble. The time line chart on the bottom shows the temporal distribution of the selected disease(s). The X-axis shows the year in which the cases are recorded and the Y-axis represents the number of cases recorded. The screenshot given above shows how filters can be applied to different interfaces of the situational awareness view. It shows the cases of whooping cough, mumps, and measles all over the world between 2009 and 2010. An additional animation in this view displays the origin and spread of a disease over time.

4.2. Explore and Analyze View

The Explore & Analyze view of the IPAS application (Figure 8) helps to drilldown into the data to perform exploratory analysis, causal analysis, and hypothesis testing. This view will assist analysts and epidemiologists to identify factors related to disease intensity such as:

- Temporal Correlation Analysis and Visualization (Is the temporal pattern of a disease correlated with weather patterns, school schedule, vector prevalence?).
- Spatial Correlation Analysis and Visualization (Is the spatial pattern of the disease correlated with spatial coordinates of vectors, contaminated water, contaminated air?). This functionality will be completed in the next phase of IPAS implementation.
- Hypothesis Testing (Is the pattern of the disease different before and after a milestone event?). This functionality will be completed in the next phase of IPAS implementation.

One of the key challenges in building the predictive model is to find the factors that influence the disease pattern. This view helps in gaining the insights required to develop predictive models. The screenshot (Figure 8) shows US flu data by regions, as defined by the US Department of Health and Human Sciences (HSS).

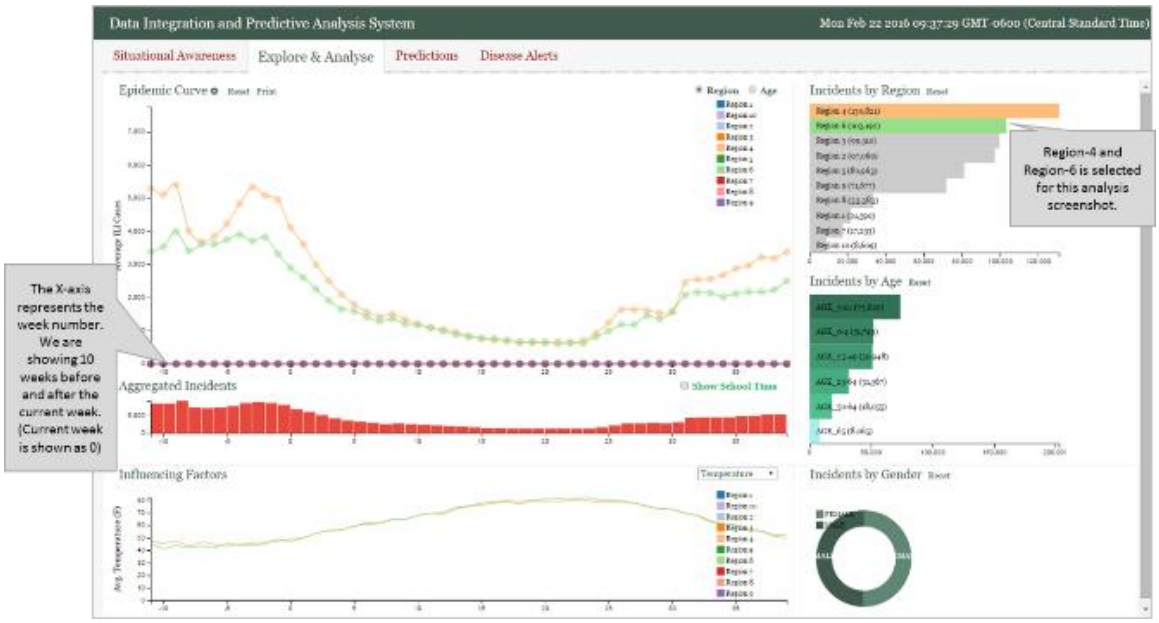


Figure 8. Explore & Analyze View of IPAS

The "Epidemic curve" shown in Figure 8 gives the graphical display of the number of incidence cases in an outbreak, plotted over time. Each line in the chart corresponds to a HHS region which are differentiated using different colors. The bar chart "Aggregated Incidents" on the bottom (red) of the Epidemic curve shows the aggregated number of cases in the selected regions. When no region is selected, the display shows the aggregated number of cases for the US. The user can filter the time focus by lasso select of a time region.

Also shown in this display are school open and break times. In the case of most endemic infectious diseases like flu, schools play a major role in transmitting the disease; overlaying the school information on top of the epidemic curve will help an epidemiologist understand that phenomenon. In order to see the school timing, the checkbox "Show School Time" on the top right corner of "Aggregated Incidents" is enabled (Figure 9).

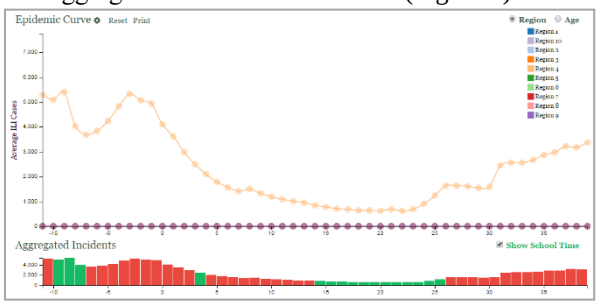


Figure 9. Comparing Epidemic Curve with School Calander

Similar to comparing the epidemic curve between different regions, this interface supports the comparison of the epidemic curve between age groups, gender, and other demographic factors by selecting associated filters and selection (Figure 8). The first-row chart of the right side shows the incidents of disease by region (Figure 9). It helps to identify the most affected and least affected regions. It also helps to compare disease patterns in different regions. The second row chart shows the incidents by age groups. The pie chart on the bottom right sorts the incidents by gender (Figure 8).

The "Influencing Factors" chart helps the analysts to understand the various factors that influence, affect, and correlate with the disease intensity. Figure 10 shows various factors that can be displayed on the chart. For example, comparing the weather patterns and the epidemic curve helps to understand whether a weather pattern has any correlation with the epidemic curve. All the charts in the Explore & Analyze view are integrated to provide the brushing and linking feature.

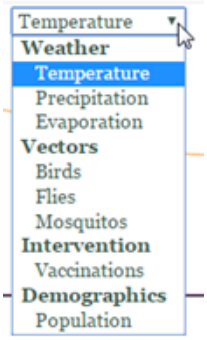


Figure 10. Disease Influencing Factors

4.3. Prediction View

The Prediction View of the IPAS application (Figure 11) helps to explore the results of the analytical and machine learning models which forecast the spread of diseases (Section 3). This view shows the model results in an intuitive fashion, supporting epidemiologists and healthcare officials to analyze:

- Expected disease pattern in the next few weeks and months.
- Peak timing and intensity.

This view also assists the users in taking prophylaxis decisions based on advanced knowledge. The filters shown on the left side enable the selection of disease(s) and regions of interest (Figure 11).

Based on the selection, the geospatial map on the top right highlights the regions and the prediction timeline chart on the bottom shows the actual progression of the selected diseases in the past 26 weeks and the prediction for next 26 weeks. The actual progression is shown to facilitate the comparison of the accuracy of the prediction model. The predicted peak week is also highlighted in the chart. Information about the current values and predicted peak values are also shown in the information box.

5. Summary

Advance knowledge about the location, timing, and intensity of infectious diseases will help public health stakeholders in taking proactive disease containment and management efforts. IPAS is motivated to provide predictive modeling infrastructure to support this important public health functionality. IPAS prediction modelling was expanded to Dengue fever incidences from Iquitos, Peru and San Juan, Puerto Rico.

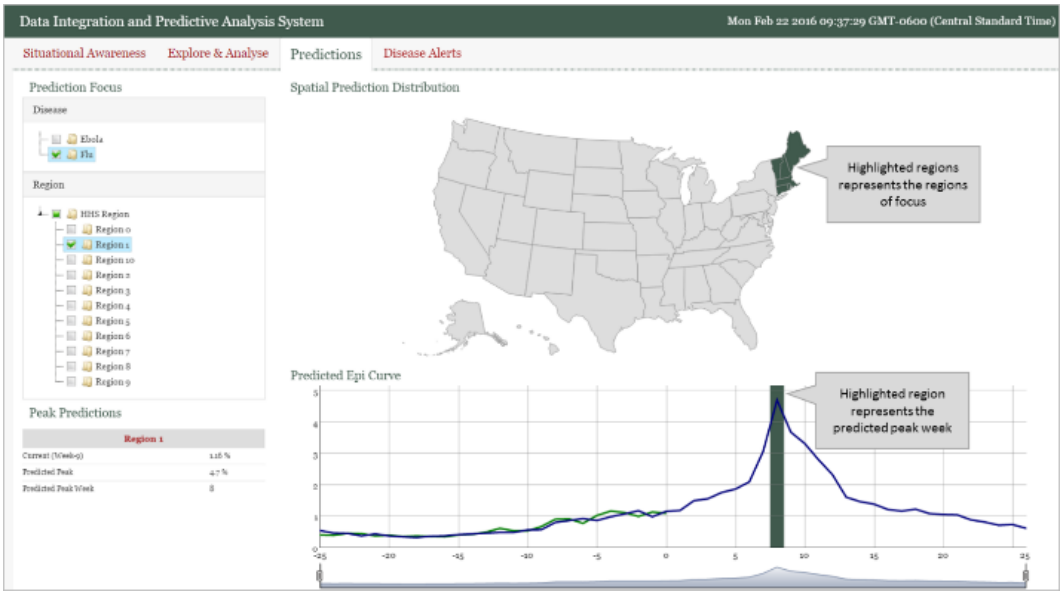


Figure 11. Prediction View of IPAS

IPAS supports comprehensive, end-to-end exploratory analysis, temporal correlation analysis, and prediction. The features of the IPAS system supporting the different stages in epidemiological data collection, integration, and analysis have been presented here.

In future work, we will attempt to improve the ability of models to predict outbreaks. Potential paths could include incorporating additional prediction techniques, and adding new features, especially for early detection of severe and outlier outbreaks.

Additional prediction techniques will be investigated. In particular, SARIMA and K-nearest neighbor models are currently being investigated, but other machine learning techniques may be incorporated as well.

Additional data will be incorporated into the prediction models to improve model performance. Country health status is a particularly interesting set of data being investigated. Country health status captures information such as number of hospital beds per capita, number of doctors per capita, and other information that may indicate how prepared a country is to contain the spread of a disease. We have collected this data from World Health Organization (WHO) and our future work will explore inclusion of this data in making disease predictions.

Additional work will also include the creation of new prediction models. For instance, instead of forecasting an incidence value, a prediction could be made on whether or not the incidence will be above or below a threshold. Very often, public health responders do not need a specific disease intensity value, but a range of values that correspond to different impacts. Alternatively, models could be adjusted to produce both a forecast and an expected error, allowing forecasts made with less certainty to be viewed with more caution. We will also attempt to predict other metrics, including the week at which the peak will occur and the maximum weekly incidence in a year.

6. Acknowledgements

This work was supported by the Defense Health Program (DHP) under Small Business Innovative Research (SBIR) Contract No. W81XWH-16-C-0130.

7. References

[1] Stephen S Morse, Jonna A K Mazet, Mark Woolhouse, Colin R Parrish, Dennis Carroll, William B Karesh, Carlos Zambrana-Torrelio, W Ian Lipkin, and Peter Daszak, “Prediction and prevention of the next pandemic zoonosis,” Lancet. 2012 December 1; 380(9857): 1956–1965. doi:10.1016/S0140-6736(12)61684-5.

[2] Sarah H. Olson, Corey M. Benedum, Sumiko R. Mekaru, Nicholas D. Preston, Jonna A.K. Mazet, Damien O. Joly, John S. Brownstein, “Drivers of Emerging Infectious Disease Events as a Framework for Digital Detection,” Emerging Infectious Diseases, Vol. 21, No. 8, August 2015.

[3] Corley CD, Pullum LL, Hartley DM, Benedum C, Noonan C, Rabinowitz PM, et al. (2014) Disease Prediction Models and Operational Readiness. PLoS

ONE 9(3): e91989. doi:10.1371/journal.pone.0091989.

[4] H. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review*, vol. 42, no. 4, p. 599–653, 2000.

[5] R. J. D. Tebbens, "A Dynamic Model of Poliomyelitis Outbreaks: Learning from the Past to Help Inform the Future," *American Journal of Epidemiology*, vol. 162, no. 4, pp. 358-372, 2005.

[6] B. T. Mayer, J. N. S. Eisenberg, C. J. Henry, G. M. Gomes, E. L. Ionides and J. S. Koopman, "Successes and Shortcomings of Polio Eradication: A Transmission Modeling Analysis," *American Journal of Epidemiology*, vol. 177, no. 11, pp. 1236-45, 2013.

[7] "Ebola modeling workshop at Georgia Tech", http://dx.doi.org/10.6084/m9.figshare.1301267

[8] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2015) Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol vol. 11, no. 8: e1004382. doi:10.1371/ journal.pcbi.1004382

[9] Wolfe, N., Daszak, P., Kilpatrick, A. M., and Burke, D. "Bushmeat hunting, deforestation, and prediction of zoonotic disease emergence." *Emerging Infectious Diseases*, Vol. 11 No. 12, 2005

[10] Jones, K., Patel, N., Levy, M., Storeygard, A., Balk, D., Gittleman, J., and Daszak, P. "Global trends in emerging infectious diseases." *Nature*, Vol. 451, 2008.

[11] Paull, S., Song, S., McClure, K., Sackett, L., Kilpatrick, A. M., and Johnson, P. "From superspreaders to disease hotspots: linking transmission across hosts and space." *Frontiers in Ecology and the Environment*, Vol. 10, no. 2. March 2012.

[12] HealthMap, "HealthMap," [Online]. Available: http://www.healthmap.org/en/. [Accessed 08 March 2016].

[13] VectorMap, "Know the vector, Know the threat," Walter Reed Biosystematics Unit, [Online]. Available: www.vectormap.org/. [Accessed 08 March 2016].

[14] M. Erraguntla, S. Ramachandran, C.-N. Wu and R. J. Mayer, "Avian Influenza Data mining Using Environment, Epidemiology, and Etiology Surveillance and Analysis Toolkit (E3SAT)," in Hawaii International Conference on System Sciences, Hawaii.

[15] S. Ramachandran, M. Erraguntla, R. Mayer and P. Benjamin, "Data Mining in Military Health Systems – Clinical and Administrative Applications," in IEEE Conference on Automation Science and Engineering, 2007.

[16] CDC, "FluView," CDC, [Online]. Available: http://www.cdc.gov/flu/weekly/fluviewinteractive.htm. [Accessed 08 March 2016].

[17] Wan Yang, Marc Lipsitchb, and Jeffrey Shamana, "Inference of seasonal and pandemic influenza transmission dynamics," PNAS, March 2015, vol. 112, no. 9, pp. 2723–2728.

[18] E. J. Pedhazur, 1997, Multiple Regression in Behavioral Research, 3rd ed, Harcourt, Inc: Troy, MO.

[19] M. Erraguntla, L. May, B. Gopal and R. J. Mayer, "Open Data Sources Based Biovigilance," in International Conference on Artificial Intelligence, Las Vegas, 2012.

[20] Benjamin, P., Madanagopal, K., Erraguntla, M., & Corlette, D. (2016, January). Distributed Information Gathering, Exploration and Sensemaking Toolkit (DIGEST). In Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 449). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[21] Google, "The Next Chapter for Flu Trends," [Online]. Available: http://googleresearch.blogspot.com/2015/08/the-next-chapter-for-flu-trends.html. [Accessed 08 March 2016].

[22] DTRA, "Biosurveillance Ecosystem (BSVE)," DTRA, 08 March 2016. [Online]. Available: http://www.dtra.mil/Portals/61/Documents/bsve-fact-sheet_draft_05-01-2014_pa-cleared-distro-statement.pdf. [Accessed 08 March 2016].

[23] NOAA. "Dengue Forecasting," [Online]. Available: http://dengueforecasting.noaa.gov/about.php [Accessed 12 Jan 2017].

[24] Erraguntla, Madhav, et al. "Data Integration and Predictive Analysis System for Disease Prophylaxis." Proceedings of the 50th Hawaii International Conference on System Sciences. 2017.

[25] Erraguntla, M., Tomasulo, P., land, K., Kamel, H., Bravo, M., Whitaker, B., Mayer, R. J., Khaire, S. 2014, "Data Mining to Improve Safety of Blood Donation Process," Hawaii International Conference on System Sciences - 47, 2014.

[26] Erraguntla, M., Gopal, B., Ramachandran, S., Mayer, R. J. 2012, "Inference of Missing ICD9 Codes Using Text Mining and Nearest Neighbor Techniques," Hawaii International Conference on System Sciences, 2012.

[27] Erraguntla, M., Ramachandran, S., Wu, C. N., & Mayer, R. J. (2010, January). Avian Influenza Datamining Using Environment, Epidemiology, and Etiology Surveillance and Analysis Toolkit (E3SAT). In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1-7). IEEE.