A Scalable Data Mining Approach for Providing Public Health with Disease Incidence Predictions Weeks in Advance

Anna L. Buczak, Erhan Guven, Steven M. Babin, Erin N. Hahn, David W. George, Yevgeniy Elbert, Liane C. Ramac-Thomas, Benjamin D. Baugher, Jacqueline S. Coberly, and Sheri H. Lewis

The Johns Hopkins University Applied Physics Laboratory (APL) has developed a novel and scalable data mining and fuzzy association rule-making approach to deriving disease incidence predictions several weeks in advance of an outbreak. This capability provides a new set of information that may be used by decision makers in conjunction with other complementary information about the country (e.g., infrastructure, disease history, agriculture, and U.S. and local military and civilian populations) from a variety of other sources (e.g., intelligence and disease experts). The prediction of the future infectious disease incidence provides the decision maker with enhanced ability to determine whether to enable deployment of measures to increase and focus biosurveillance and/or to plan and enable mitigation efforts to reduce morbidity and mortality well in advance of the start of the outbreak.

INTRODUCTION

Infectious disease outbreaks result from interactions among the host, the pathogen, and the environment, which are components of the epidemiological triad.[1] For many years, study of these outbreaks has focused on using models for the dynamics of disease spread[2] or for outbreak surveillance and detection.[3] The sooner an outbreak is detected, the more timely and effective are the measures that can be deployed by public health agencies to mitigate the morbidity and mortality due to the disease.[4] Recent studies have moved from outbreak detection to the prediction of outbreaks before they occur. Most of these studies rely on varying types of regression analysis,[5] while others use such techniques as neural networks.[6] Because the incidence of many diseases is recognized as being influenced by the environment (e.g., vector-borne diseases), investigators have turned to using environmental data such as temperature and rainfall[7] to make predictions. One advantage of using such environmental variables is that many of them can be obtained by satellite remote sensing, thereby providing the ability to study remote areas and avoid expensive field measurements. Other types of environmental variables include climate indices,[8] such as the Southern Oscillation Index, the West Pacific Index, and the NINO3, which is an eastern Pacific Ocean sea surface temperature anomaly index. Such cli-

mate indices are used to enhance further the predictive capability because they are known to be leading indicators of future changes in seasonal and nonseasonal weather patterns.[9] Vegetation indices are derived from satellite remote sensing data and provide indications of vegetation types and conditions, soil moisture, and the effects of fires and human land use, all of which may have impacts on disease vector habitat.[10–11]

It is important to emphasize that the software system described herein substantially differs from systems designed for the early detection of disease outbreaks, such as the Johns Hopkins University Applied Physics Laboratory (APL) Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE).[12] Early-detection systems use a variety of data to detect a disease outbreak that has already begun but is not yet obvious because, for example, there are few cases or the disease is still in its prodromal stage. This article will instead describe a system that makes a prediction of disease incidence several weeks into the future, even before a disease outbreak has begun. When testing their predictive capabilities, most authors of published modeling studies have a tendency to (*i*) use input data that were already used in model development, (*ii*) assume all input data are available at the exact time the prediction is made (time T), or (*iii*) both. Both of these tendencies will lead to exaggerated measures of model performance compared to how the model would be expected to perform in an operational environment. Models developed and tested on the assumption that all the most up-to-date data are available for model input at time T are, in effect, making retroactive predictions compared to how the model would be used in a realistic operational environment. Assume, for example, that the prediction model requires weekly disease incidence input data and the users want a prediction made on Monday, but the data for the previous week will not be available until Friday. Because data are not actually available at the time a prediction is made, the model will either fail to perform at its tested level of accuracy or will not run at all. Because the APL team is focused on the needs of the user (e.g., force health protection and local public health professionals), we take great care in avoiding these two tendencies so that our tested model prediction accuracy will more reliably indicate how the user may expect the model to perform. The APL disease prediction system is designed and tested by using data that are actually available to the user at the time the prediction is made, so the resulting prediction accuracy is more realistic for the user. The APL team sought and received input from users ranging from civilian public health departments in other countries to U.S. military public health professionals. There was consensus among these users that a system that can predict disease incidence a month or more in advance would be especially valuable to them both for planning purposes and for implementation of mitigation measures (e.g., ranging from public education efforts such as reducing standing water to spraying insecticide).

This article will further show that using techniques involving fusion of data from disparate sources along with fuzzy association rule mining (FARM)[13] can result in the development of prediction models with promising results for the public health professional responsible for mitigating the effects of disease outbreaks. The project described in this article is called the PRedicting Infectious Disease Scalable Method (PRISM) and was developed for the Joint Program Manager–Information Systems of the DoD. The PRISM software system was built to automate the FARM process prototyped previously by APL[9] for predicting infectious disease and to provide an easily interpreted visualization of the results.

PREDICTION METHOD

The PRISM algorithms and software suite have three tasks to accomplish (*i*) building a prediction model; (*ii*) establishing, defining, and maintaining a database, as well as automating input data download, data preprocessing, and prediction generation; and (*iii*) visualizing prediction output. The first task is the most computationally intensive of the three tasks, but it does not have to be repeated once the prediction model is finalized for a particular disease and geographic region. Running the model to generate a prediction and visualizing the output are not computationally intensive tasks and can be performed on a typical laptop computer.

Building a Prediction Model

Figure 1 provides an overview of the method for building a prediction model and generating predictions. In step 1, subject matter experts and analysts determine what variables might be relevant for a certain disease and location (e.g., dengue fever in a district in Peru) and the data sources for these variables. Most of the time, relevant variables can be determined by reviewing the scientific literature. Once the variables are determined, the sources of data for those variables must be found.

Step 2 (Fig. 1) is building the prediction model and testing it. This is first done one time for a particular disease and location to see whether a model can be built that performs with reasonable accuracy. If this is the case, a new model should be retrained about once a year to make certain that the accuracy of prediction remains reasonably high. In building a prediction model, an extended historical database is especially valuable in performing the data mining process. RapidMiner software (http://rapidminer.com/products/), with a large number of extensions developed by APL, is used to implement the FARM methodology and build a classifier based on these historical data. The classifier uses machine-learning techniques to predict class member-

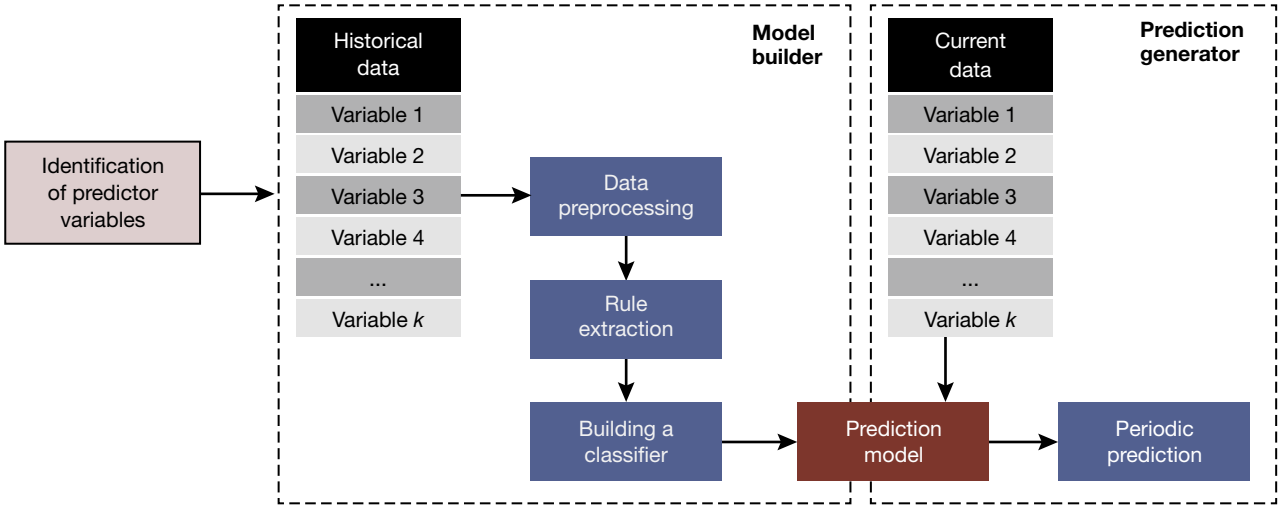

Figure 1. An overview of the PRISM method for creating and using a prediction model.

ship for data instances. For example, this classification can be for HIGH or LOW disease incidence, where a threshold between these two classes is based on the scientific literature, opinions of subject matter experts, and the desire of the user to have as low false-positive and false-negative rates as feasible for operational use. While the system so far has used classification to define two classes, the system is not limited to this number.

The classifier uses a set of rules to define the classes. Fuzzy association rules are of the form

$$\text{IF } (X \text{ is A}) \text{ THEN } (Y \text{ is B}),$$

where X and Y are variables, and A and B are membership functions that characterize X and Y, respectively. X is called an antecedent and Y is called a consequent of the fuzzy association rule. The rules are fuzzy because there can be overlapping memberships where the amount of overlap is quantified. As an example, four fuzzy membership functions (SMALL, MED, LARGE, and VERY LARGE) for the variable rainfall are shown in Fig. 2. Fuzzification is defined as the process in which a number (e.g., rainfall value in millimeters) is transformed into a membership value lying between 0 and 1, thereby allowing for a smooth transition between full membership (1) and nonmembership (0). The degree of membership in a set is generally considered to be the extent to which a corresponding fuzzy set applies. In Fig. 2, a rainfall of 50 mm will be transformed into two membership functions: SMALL, with a degree of membership 0.5; and MED, with a degree of membership 0.5.

Because the model builder typically discovers thousands of such rules based on the historical data, a set of criteria must be used to select the best rules to be used in the final prediction model. These criteria include metrics called confidence, lift, and support.[9,13] Confidence is the conditional probability that if the antecedents of a rule are true, then the consequent of the rule is also true. A confidence equal to unity means that if the antecedents are true, the consequent is always true. Support is a metric for how general the rule applies. For example, a support equal to 0.01 means that the rule applies to 1% of the data. Lift is another metric, and the higher it is, the more dependent the variables are on one another. More details on these metrics may be found in Agrawal et al.[13] and Buczak et al.[9] The selection of the rules and

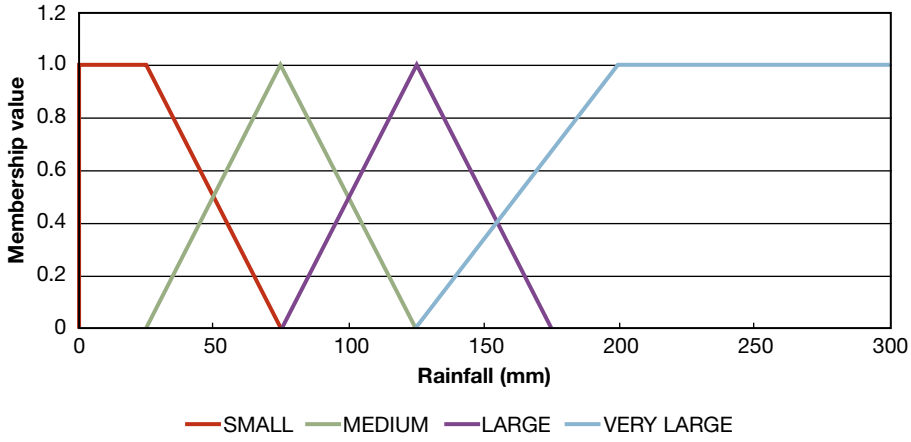

Figure 2. Example of fuzzy membership functions (SMALL, MED, LARGE, and VERY LARGE) for the variable rainfall.

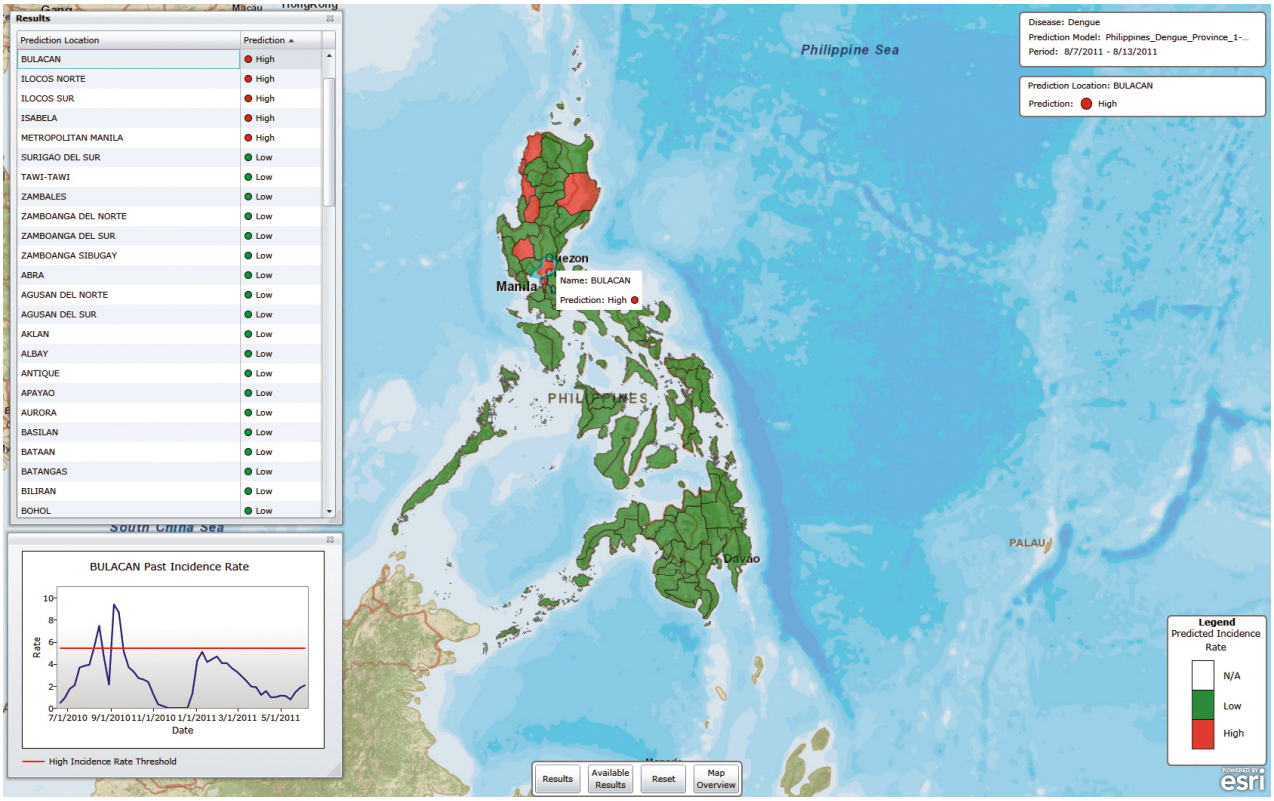

Figure 3. Results viewer for predicted dengue incidence in the Philippines, using the province-level resolution.

the order in which they are applied in the prediction model are determined by an automatic method that minimizes the misclassification error.[9]

Automation of Data Download and Prediction Generation

To make a prediction by using the model described above, the most recently available data are downloaded from the source (step 3 in Fig. 1). The data sets are described in more detail in the *Data Sets* section. Note that the system does not assume that all data are from time T, the time the prediction is made. At time T, if only data from week T-2 (i.e., 2 weeks prior to time T) are available, then those are the data used by the model. Because these data are from disparate sources and typically have different temporal and spatial resolutions, the data are preprocessed into a uniform space resolution (e.g., a defined geographic region such as a province) and time resolution (e.g., 1 week). Once all the predictor variable data have been downloaded, the classifier selected in the previous task is automatically used to generate disease incidence predictions at the same spatial and temporal resolution (e.g., a province and weekly).

Visualization of the Prediction Results

A visualization software tool was developed to illustrate the prediction results on a map. This tool is a viewer based on the Esri ArcGIS application programming interface (API) for the Microsoft Silverlight application (http://www.microsoft.com/silverlight/). The tool uses a catalog service to create a list of currently available results and a results service to display graphically the locations of the predictions color-coded by the classification used for the prediction results (e.g., red for HIGH and green for LOW). In addition to these two Extensible Markup Language (XML) services, Keyhole Markup Language (KML) files saved locally may be browsed and loaded to visualize the results. Further details will be described in the *Software Architecture* section. An example of the results map generated by this process is shown in Fig. 3.

SOFTWARE ARCHITECTURE

An overview of the PRISM software architecture is shown in Fig. 4. PRISM has two major software components that are used to create disease incidence prediction models and prediction result displays for the user: the model builder and the PRISM web application. The model builder is responsible for creating a classifier that, along with a current data feed, produces the prediction results for a geographic region. The PRISM web application handles the extraction, transformation, and loading (ETL) of external data sources for use in both model building and prediction. In addition, the PRISM web

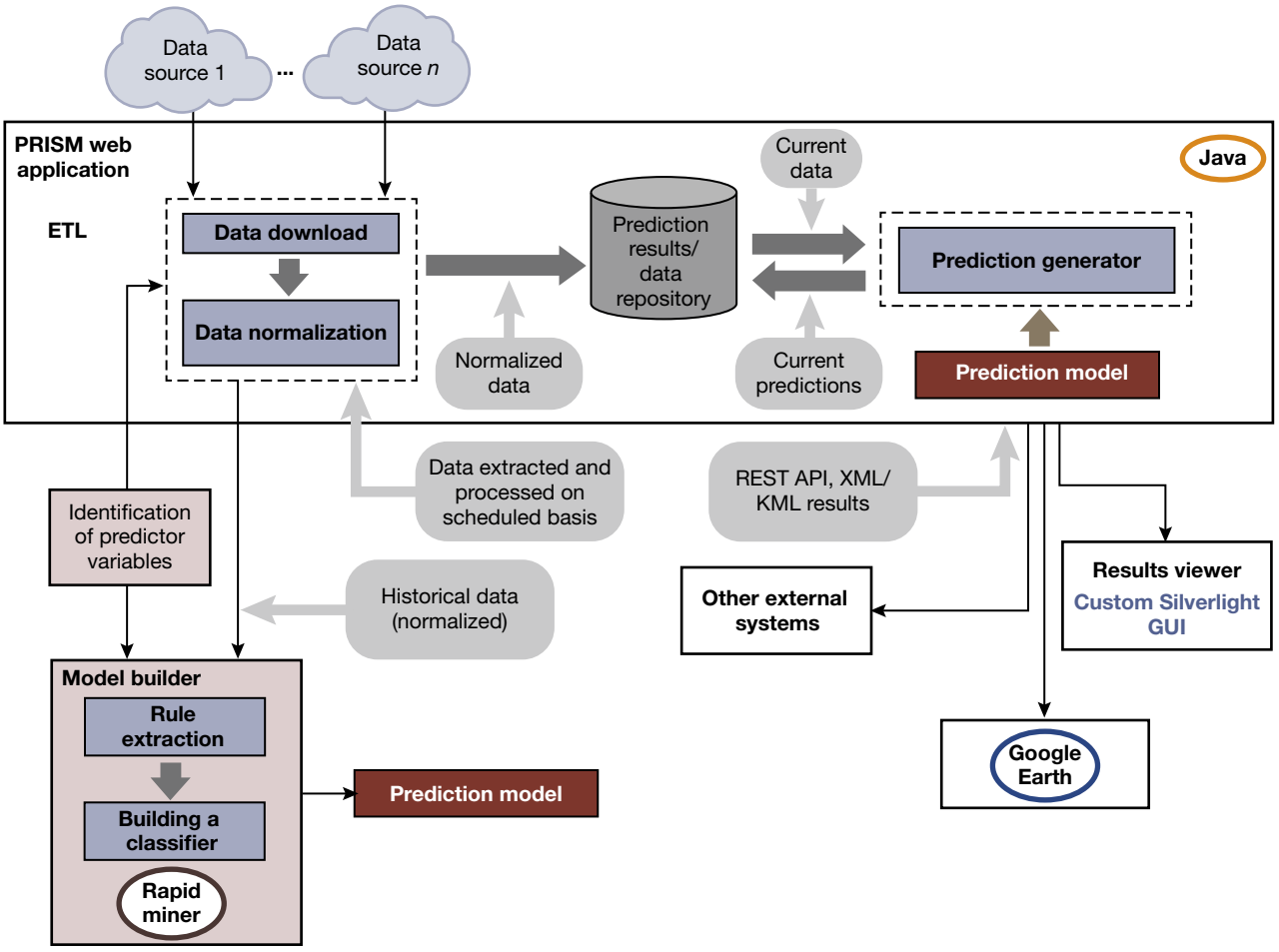

Figure 4. PRISM software architecture.

application executes the available classifiers to produce prediction results and provide these prediction results through an external Representational State Transfer (REST) interface.

The model builder is based on the RapidMiner data mining software framework. RapidMiner uses the Java programming language, which provides an easy way to modify its functionality for use with the PRISM system. APL has extended this framework with specific operators that implement the FARM and classifier building algorithms. This model building begins with the selection of the appropriate predictor variables (e.g., rainfall, incidence rates, etc.), which depends heavily on a literature review and the subject matter expertise of the analysts. Historical data that comprise all of the selected predictor variables for the geographic region of interest are collected. As with any data mining technique, the process is more effective with more data. These data are then divided into three disjoint subsets: training data, fine-tuning data, and testing data. The training data are then used in performing the FARM data mining methodology to extract the rules and build the classifiers. The best performing classifier on the fine-tuning data subset becomes the prediction model. The testing data subset, which has not been previously used, is used to measure the accuracy of this model. The model builder process needs to be repeated for every new disease and geographic region. Also, when a new complete year of data becomes available, the model builder process should be repeated so that the training on these new data will help the model maintain reasonable accuracy of its predictions.

The PRISM web application is responsible for scheduling and executing collection and normalization of data from the identified sources, scheduling and running the classifier used to compute a prediction, implementing the web services that display the prediction results, and hosting the standalone viewer (see example of viewer results in Fig. 3). The web application is logically separated into three primary functional areas: the ETL functionality, the prediction generation functionality, and the web services functionality.

1. ETL refers to a process in database usage and especially in data warehousing that involves extracting (downloading) data from outside sources, transforming (or normalizing) the data to fit operational

needs, and loading the data into the end database or operational store. In the context of PRISM, the raw predictor-variable data are downloaded, transformed to work with the prediction model, and loaded into a geospatially enabled database for easy retrieval by the model. The database can be loaded with static or downloaded files, including large amounts of dynamic data. The ETL also handles missing data errors, including error notifications and repeated attempts at data downloading. The raw data are referenced by jurisdictional division and mapped to geographical resolution. The data are also selected and arranged with respect to both geographic and temporal resolution for exporting to the model.

2. Prediction generation is the process of generating prediction results by using the normalized data and a classifier (prediction model) developed during the model-building phase. The necessary data include data from previous weeks (e.g., T-1, T-2, T-3, T-12) but never from the present week (T) because it always takes a few days for those data to become available. For certain variables (such as Normalized Difference Vegetation Index and Enhanced Vegetation Index) that come in 16-day averages, the most recent week that can be used is T-4 to ensure that the data are really available for downloading. In addition, predictions can be made when some data

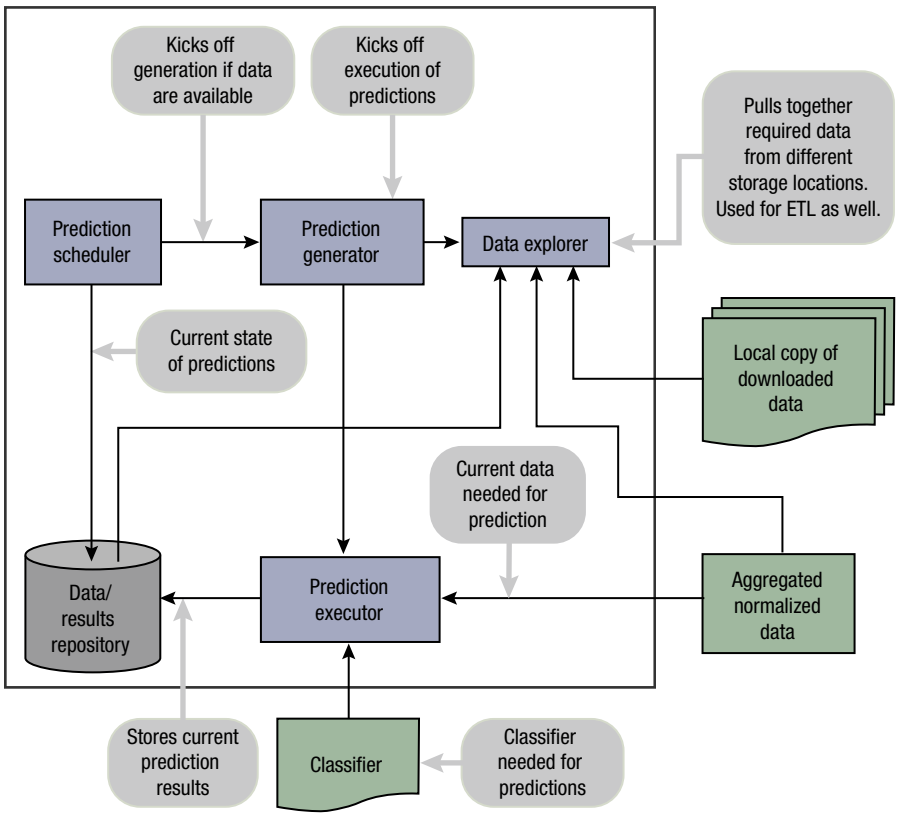

Figure 5. Prediction generation architecture.

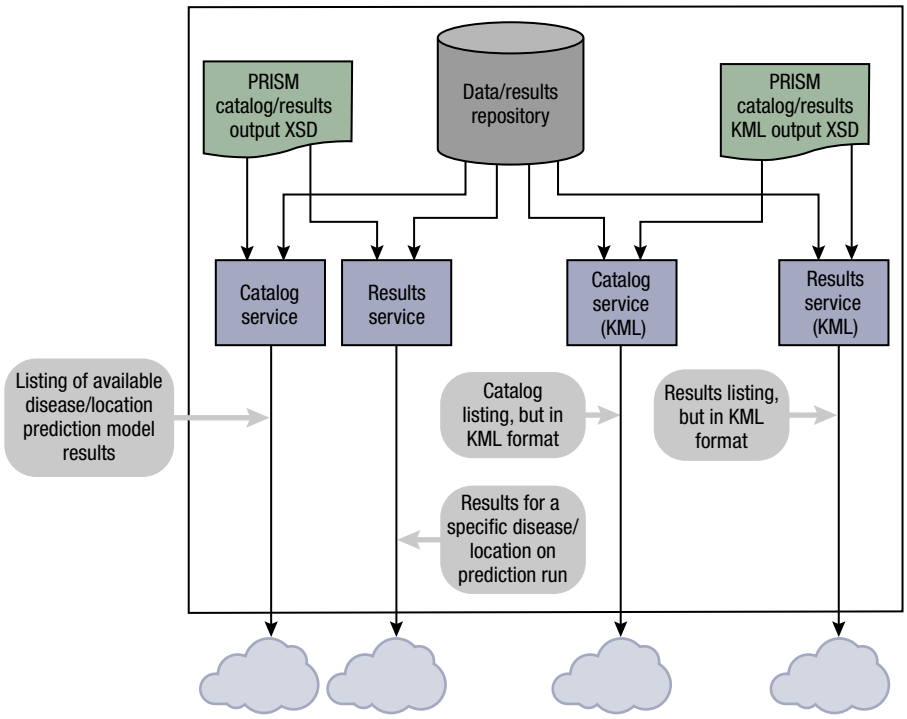

Figure 6. Web service architecture. The catalog and results services on the left side are in xsd format.

are missing for other reasons. The prediction results are stored in the database. The prediction generation architecture is shown in Fig. 5. This software is responsible for scheduling and running the preconfigured prediction models by using the data exported by ETL. Not all data sources are updated at the same interval, so predictions are scheduled for the time that the input data are typically available. The prediction model and the defined membership functions used to map predictor variables to their attributes are configured using the Spring Framework (http://projects.spring.io/spring-framework/).

3. The PRISM predictions are made accessible through several REST web services. REST web services facilitate transactions between web servers by allowing loose coupling between different services. REST is used because it allows greater flexibility in output format than the Simple Object Access Protocol (SOAP) and, unlike SOAP, it requires less configuration and does not need a message header. The web service architecture is shown in Fig. 6. The PRISM web services list the available prediction model results in both XML Schema Definition format (xsd) and KML. The resulting catalog of available disease/location/model results uses the xsd format to define the required and optional fields and provide the interface to the web service. The KML format is used to allow the catalog results to be displayed in any KML-compliant application, such as Google Earth.

DATA SETS

The APL data mining system for developing and using predictive models involves a large amount of disparate types of data (see Fig. 7). Some of these data, such as land elevation, are static. Some data sources are updated only annually and manually, such as population and demographics. The Southern Oscillation Index and sea surface temperature anomaly indices are automatically updated monthly and weekly, respectively. Satellite-derived vegetation indices, land surface temperature, and rainfall are automatically updated every 16 days, every 8 days, and every 3 hours, respectively. Disease incidence data may be automatically updated daily or weekly. Therefore, assembling these disparate data into a database compatible for use by the disease prediction model involves preprocessing into a uniform spatiotemporal resolution.

As mentioned earlier, the data were divided into three disjoint subsets: training data, fine-tuning data, and testing data. Only the first two subsets were used to develop a prediction model, while the third set was used to measure the accuracy of the prediction.

RESULTS

The model is built using only the training data, and it is fine-tuned using the fine-tuning data set mentioned above. Current data are input to the prediction model, and the predictions (4 weeks ahead) can be compared against the testing data set that the model has never seen. For such a comparison to be made, at least two

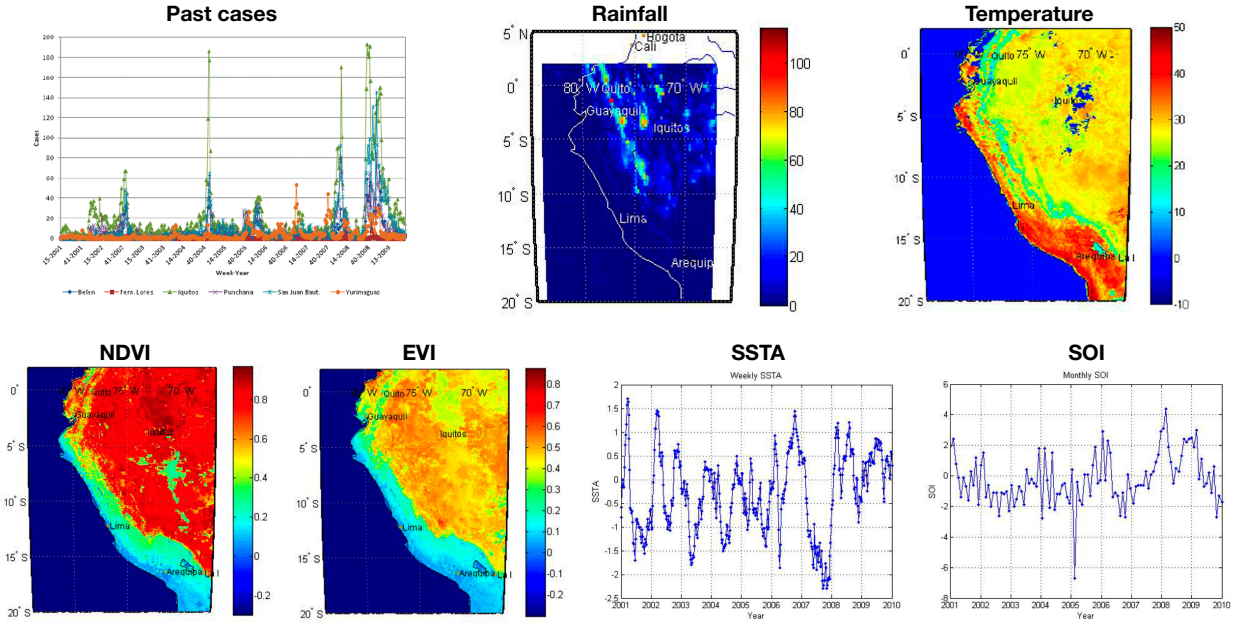

Figure 7. Examples of disparate types of data used in the disease prediction model.

disease states have to be defined. In this case, we defined HIGH incidence as a weekly incidence rate greater than or equal to a fixed number (e.g., 1.5) of standard deviations above the mean historical incidence based on the training and fine-tuning data sets. LOW incidence is anything below that rate. Note that there is no uniform definition of disease outbreak that involves standard deviations above mean historical incidence, and furthermore, the statistical distribution of disease incidence is often not Gaussian.[14] Different diseases exhibit different temporal fluctuations in incidence, and this may also vary by region and population. For the user, the threshold should be operationally meaningful, meaning that it is set at some level above which the user would typically take some kind of action. Depending on where this threshold is set, the accuracy of the prediction can then be evaluated to see when it predicted a HIGH or LOW incidence and how this compared to the actual data.

True positives (TP) are defined as instances where a prediction of HIGH disease incidence corresponded to actual data confirming a HIGH disease incidence. True negatives (TN) are similar except that the prediction of LOW disease incidence corresponds to actual data confirming this result. A false positive (FP) occurs when the prediction says incidence is going to be HIGH but the actual data show it to be LOW. A false negative (FN) occurs when the predicted incidence is LOW, but the actual data show it be HIGH. To determine the accuracy of these predictions, four commonly used metrics were used:

1. Positive Predictive Value (PPV): $PPV = \frac{TP}{TP + FP}$ or the proportion of positive predictions that are outbreaks;
2. Negative Predictive Value (PPV): $NPV = \frac{TN}{TN + FN}$ or the proportion of negative predictions that are non-outbreaks;
3. Sensitivity: $Sensitivity = \frac{TP}{TP + FN}$ or the proportion of correctly predicted outbreaks (also called Probability of Detection);
4. Specificity: $Specificity = \frac{TN}{TN + FP}$ or the proportion of correctly predicted non-outbreaks; note that 1 – Specificity is the False Alarm Rate.

The public health users in Peru wanted a prediction of 4-week dengue incidence for the district of interest made 4 weeks ahead (i.e., 4 weeks from the date the prediction was made), and they wanted a threshold between HIGH and LOW of 2 standard deviations above the multiyear mean[9] because this was the level above which they would consider a response. Accordingly, for the Loreto district of Peru, our predictions of 4-week incidence made 4–7 weeks into future resulted in a PPV, NPV, Sensitivity, and Specificity of 0.81, 0.98, 0.64, and 0.99, respectively. For the Philippines, a threshold between HIGH and LOW of 1.5 standard deviations above the multiyear mean was used on the basis of feedback from the users and the quality of the data (e.g., noise, missing data, etc.). The users in the Philippines wanted predictions of weekly incidence made 4 weeks in advance. In this case, for year-2011 predictions 4 weeks in advance of weekly dengue incidence in the Abra province in the Philippines, the PPV, NPV, Sensitivity, and Specificity were 0.75, 0.82, 0.64, and 0.88, respectively.

LESSONS LEARNED FROM THE TABLETOP EXERCISE

Over the course of PRISM's development, end users and stakeholders were familiarized with PRISM through a series of discovery workshops. The purpose of the discovery workshops was to give individuals and organizations tasked with addressing infectious disease outbreaks an opportunity to learn about PRISM, provide feedback, and influence the development of a culminating tabletop exercise (TTX). The purpose of the TTX was to obtain and document more substantial feedback from a broader community of potential end users and to support further development of the model and the creation of a formal concept of employment.

The TTX was held March 5–7, 2013, at Lawrence Livermore National Laboratory in Livermore, California. The event brought together 35 key attendees, including representatives from seven Combatant Commands, the Armed Forces Health Surveillance Center, the National Center for Medical Intelligence, and the Centers for Disease Control and Prevention. The TTX used different scenarios to examine the value to the end users of a disease prediction capability such as PRISM and to better understand customized features the end users recommended to make the tool a useful part of their workflow.

The TTX participants were very enthusiastic about PRISM, and the TTX resulted in several key findings. First, it became clear that an infectious disease predictive capability could provide significant improvement to accomplishing the Combatant Commands' force health protection mission by conserving time, funding, and resources. Second, when this capability is used in conjunction with other information, it enables Combatant Command planners to choose from a wide range of proactive options to address a potential infectious disease outbreak. Third, resources for infectious disease mitigation and consequence management can be applied more discriminately in a proactive rather than reactive manner. Overall, a predictive capability like PRISM can help improve operational readiness.

The TTX concluded that PRISM should be further developed and should focus, as much as possible, on user-identified features. The TTX identified the desire among users to develop predictive disease incidence models for

all designated infectious diseases of interest in areas of concern, including risks from both natural and man-made outbreaks. While such an objective cannot be achieved in a short period of time, this is a clear demonstration of the end users' enthusiasm for the capability. Some of the other user-desired features included the ability to provide three levels of risk prediction (e.g., red, yellow, and green, instead of only red and green); a broader selection of spatial and temporal granularity; the ability to provide predictions that can be integrated into existing health risk assessment procedures in order to achieve an integrated composite view of the health risk; greater flexibility for models to be easily updated with real-time information (e.g., to account for a socioeconomic crisis); and the ability to do "what if" analysis within the model to anticipate the influence of certain variables on the likelihood of a disease outbreak.

The TTX was a capstone event for PRISM, the results of which will be used to guide future development. End user and stakeholder feedback are critical for the model to be used as broadly and effectively as possible. The TTX format was particularly effective because it allowed participants to brainstorm and build on each other's ideas. The collaboration that took place led to clear recommendations and highly useful feedback for future development of PRISM.

CONCLUSIONS

Among the many challenges faced by public health professionals are the large efforts in planning and allocating resources when disease outbreaks occur. Having a reliable predictive capability can help them to anticipate and prepare more effectively and efficiently for an outbreak. A predictive capability does not replace an early detection capability, which allows for prediction validation and more refined efforts to mitigate the disease outbreak that is already happening. Therefore, prediction and detection systems complement one another. This article describes a prediction capability that is unique as well as novel. It combines advanced techniques in FARM and disparate data fusion to provide a prediction of disease incidence several weeks in advance, even when there is not yet any evidence of an actual outbreak.

The method developed by APL requires one-time model development (CPU intensive) for a specific disease in a specific country or region. This resulting predictive model (not CPU intensive) is then run when requested and when new data are available. New epidemiological data are typically collected and compiled by the public health user weekly so current versions use weekly data. Output of a model run is a country map showing provinces with disease incidence higher than normal or not; this output is provided at least a month in advance and possibly before an outbreak has even begun. PRISM is a tool for prediction and not detection. Detection of an outbreak can occur only when the earliest stages of the outbreak have begun.

The users of a predictive capability for disease incidence have indicated that they prefer results with as low a proportion of false positives and false negatives as possible. Too many false positives can lead to wasted resources and alarm fatigue.[15] Too many false negatives may lead users to wonder whether the predictive capability is adding any value to what they have traditionally done. The method developed by APL reduces both false positives and false negatives to levels deemed acceptable by the public health user, thereby providing added value. According to these users, having at least 4 weeks lead time before a prediction of high disease incidence allows for a wide range of mitigation options. For example, the public health user may choose to begin enhancing ongoing biosurveillance by instigating more frequent data collection and the addition of more data sources, thereby improving the odds for early detection. Alternatively, instead of waiting for early detection to confirm the predicted outbreak, the user may decide that more proactive measures are needed. Depending on the disease, the potentially impacted population, and other factors, these measures may range from less to more aggressive, including such measures as intensified public health educational outreach, using resources to reduce vector habitat, or even employing disease quarantine or personnel evacuation at the earliest signs of positive detection. APL has begun efforts to extend the methodology described in this article to new diseases, including malaria in the Republic of Korea and influenza in the United States.

ACKNOWLEDGMENTS: Funding for this work is provided by the U.S. Department of Defense Joint Program Manager, Medical Countermeasures Systems, JPEO-CBD.

REFERENCES

[1]Wallace, R. B., Doebbeling, B. N., and Last, J. M. (eds.), *Public Health and Preventive Medicine*, 14th Ed., Appleton and Lange, Stamford, CT (1998).

[2]Chowell, G., Diaz-Duenas, P., Miller, J. C., Alcazar-Velasco, A., Hyman, et al., "Estimation of the Reproduction Number of Dengue Fever from Spatial Epidemic Data," *Math. Biosci.* **208**(2), 571–589 (2007).

[3]Xing, J. Burkom, H., and Tokars, J., "Method Selection and Adaptation for Distributed Monitoring of Infectious Diseases for Syndromic Surveillance," *J. Biomed. Informat.* **44**(6), 1093–1101 (2011).

[4]Bootsma, M. C., and Ferguson, N. M., "The Effect of Public Health Measures on the 1918 Influenza Pandemic in US Cities," *Proc. Natl. Acad. Sci. USA* **104**(18), 7588–7593 (2007).

[5]Halide, H., and Ridd, P., "A Predictive Model for Dengue Hemorrhagic Fever Epidemics," *Int. J. Environ. Health Res.* **18**(4), 253–265 (2008).

[6]Husin, N. A., Salim, N., and Ahmad, A. R., "Modeling of Dengue Outbreak Prediction in Malaysia: A Comparison of Neural Network and Nonlinear Regression Model," in *Proc. International Symp. on Information Technology* (ITSim 2008), Vol. 3, Kuala Lumpur, 1–4 (2008).

[7]Hii, Y. L., Zhu, H., Ng, N., Ng., L. C., and Rocklov, J., "Forecast of Dengue Incidence Using Temperature and Rainfall," *PLoS Negl. Trop. Dis.* **6**(11), e1908 (2012).

[8]NOAA Earth System Research Laboratory Physical Sciences Division, "Climate Indices: Monthly Atmospheric and Ocean Time Series," http://www.esrl.noaa.gov/psd/data/climateindices/list/ (accessed 26 Sep 2013).
[9]Buczak, A. L., Koshute, P. T., Babin, S. M., Feighner, B. F., and Lewis, S. H., "A Data-Driven Epidemiological Prediction Method for Dengue Outbreaks Using Local and Remote Sensing Data," *BMC Med. Informat. Decis. Making* **12**, 124 (2012).
[10]Kalluri, S., Gilruth, P., Rogers, D., and Szczur, M., "Surveillance of Arthropod Vector-Borne Infectious Diseases Using Remote Sensing Techniques: A Review," *PLoS Pathog.* **3**(10), e116 (2007).
[11]Ferreira, N., Ferreira, L., and Huete, A., "Assessing the Response of the MODIS Vegetation Indices to Landscape Disturbances in the Forested Areas of the Legal Brazilian Amazon," *Int. J. Remote Sens.* **31**(3), 745–759 (2010).
[12]Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S., et al., "A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE)," *J. Urban Health* **80**(2, Suppl 1), i31–i42 (2003).
[13]Agrawal, R., Imielinski, T., and Swami, A., "Mining Association Rules Between Sets of Items in Large Databases," in *Proc. International Conf. on Management of Data*, Washington, DC, 207–216 (1993).
[14]Texier, G., and Buisson, Y., "From Outbreak Detection to Anticipation," *Revue d'Epid. Et de S. Publique* **58**, 425–533 (2010).
[15]Raso, R., and Gulinello, C., "Creating Cultures of Safety: Risk Management," *Nurs. Manage.* **41**(12), 26–33 (2010).

The Authors

Anna L. Buczak is a project manager, section supervisor, and Principal Professional Staff data mining specialist. She leads the PRISM team. **Erhan Guven** is a Senior Professional Staff computer scientist. He developed the majority of PRISM software. **Steven M. Babin** is a Senior Professional Staff medical doctor and remote sensing specialist. He contributed medical expertise as well as satellite remote sensing and atmospheric science expertise to the analysis and interpretation of the data. **Erin N. Hahn** is a Senior Professional Staff member. She serves as the Assistant Project Manager for PRISM. **David W. George** is an Acting Assistant Group Supervisor and a Senior Professional Staff software engineer. He serves as software lead for PRISM. **Yevgeniy Elbert** is a Senior Professional Staff statistician. He performed preprocessing of the epidemiological data. **Liane C. Ramac-Thomas** is a Senior Professional Staff Data fusion specialist. She developed the dengue and malaria prediction models. **Benjamin D. Baugher** is a Senior Professional Staff mathematician. He developed the dengue and malaria prediction models and algorithms. **Jacqueline S. Coberly** is a project manager, section supervisor, and Senior Professional Staff epidemiologist with a focus in infectious disease epidemiology. **Sheri H. Lewis** is a Principal Professional Staff public health specialist. She serves as the Program Manager for Global Health Surveillance. For further information on the work reported here, contact Anna Buczak. Her e-mail address is anna.buczak@jhuapl.edu.

The *Johns Hopkins APL Technical Digest* can be accessed electronically at **www.jhuapl.edu/techdigest**.