A Disease Outbreak Detection System using Autoregressive Moving Average in Time Series Analysis

Richard John M. Buendia [1,2]
[1]University of the Philippines, Manila
Manila, Philippines
[2]Ideyatech Inc, Ortigas,
Pasig City, Philippines
E-mail: richard@ideyatech.com

Geoffrey A. Solano [1,3]
[3]Algorithms and Complexity Laboratory,
Department of Computer Science
University of the Philippines, Diliman
Quezon City, Philippines
E-mail address: geofsolano@gmail.com,
gasolano@upm.edu.ph

Abstract—**Disease outbreak detection is an immensely beneficial yet a very delicate procedure. Its output can potentially provide immeasurable help and save countless lives in preventing further problems. Thus, the method must be as accurate and precise as possible. This paper discusses the Disease Outbreak Detection System, which is an online system developed to aid public health workers as they resolve this problem. Specifically, it helps epidemiologists analyze the behavior of a certain disease outbreak by providing them prediction values for a specific time interval. The system is able to perform such feature with the aid of R software which performs computations of Time Series analysis using Autoregressive Moving Averages (ARMA) Model to generate values based on the present condition of the outbreak. These generated values will serve as basis to know how the outbreak will turn out, thus giving the epidemiologist sufficient time to respond to major public health threats and formulate preventive measures to control and solve the outbreak. The tool was developed for the Philippine setting, specifically for the use of outbreak monitoring agencies such as the National Epidemiological Center (NEC), and thus uses Philippine health data which basically comes from two major sources: surveys and censuses, as well as from administrative records of health and health related agencies.**

Keywords— Disease Outbreak Detection, Time Series Analysis, ARMA

I. INTRODUCTION

Infectious diseases are clinically evident diseases resulting from the presence of pathogenic microbial agents, including pathogenic viruses, pathogenic bacteria, fungi protozoa, multicellular parasites and aberrant proteins. These pathogens are able to cause disease in humans, animals or plants, and are called communicable diseases due to their potential of transmission from one person or species to another. Transmission of an infectious disease may occur through one or more of diverse pathways including physical contact with infected individuals. These infecting agents may also be transmitted through liquids, food, body fluids, contaminated objects, airborne inhalation, or through vector-borne spread. [9]

Third-world countries often have a large number of casualties resulting from such diseases due to lack of facilities to detect, track and treat them. In the Philippines, the public health community has historically relied on the watchful eyes of doctors, who have reported individual cases or clusters of cases of particular diseases to the authorities for disease outbreak detection. But these days, the availability of electronic health-care data should facilitate more automated and earlier outbreak detection and intervention if there are systems that can make use of such data. Besides diagnoses of known diseases, other indicators--such as primary complaints of patients coming to the emergency room or calling a nurse hotline--are being collected in electronic formats and could be analyzed if suitable methods existed [13]. These solve one of the causes of inconsistency among data gathered. It minimizes the load on reporting cases therefore it will be easy to submit data which result to a lesser deviation.

A disease outbreak detection system will enable epidemiologists to generate an approximation on how many will be infected in a given time utilizing the set of data gathered from different time intervals. These will be beneficial to the health sector for the reason that they will have an idea on what will be the status of the outbreak for a given time interval therefore providing them ideas on what degree the outbreak will turn out. As a result, it will provide our health sectors enough time to plan an action on preventing or controlling the situation.

Accurate and timely detection of infectious disease outbreaks provides valuable information which can enable public health officials to respond to major public health threats in a timely fashion. However, disease outbreaks are often not directly observable. For surveillance systems used to detect outbreaks, noise caused by routine behavioral patterns and special events, can further complicate the detection task [6]. For this system, only the set of data gathered in a given time interval are analyzed, excluding all other factors contributing to disease outbreak detection for the purpose of avoiding the occurrences of such noise.

A number of studies have been done in detecting health risks using such models in time series analysis. SPSS13.0

software was used to construct the Autoregressive Integrated Moving Average (ARIMA) model based on the monthly malaria incidence of Huaiyuan and Tongbai countries in Huaihe River Valley, from Jan. 1998 to Dec. 2005. The constructed model was then applied to predict the monthly malaria incidence in 2006 and the incidence from ARIMA model was compared with the actual incidence, so as to evaluate the model's validity. Malaria incidence of 2007 was predicted by ARIMA model based on malaria incidence from 1998 to 2006 [11]. A univariate time-series analysis method has been used to model and forecast the monthly number of dengue haemorrhagic fever (DHF) cases in southern Thailand. They developed ARIMA models on the data collected between 1994–2005 and then validated the models using the data collected between January–August 2006. The results showed that the regressive forecast curves were consistent with the pattern of actual values. The ARIMA (1,0,1) model fitting was adequate for the data with the Q-statistic (Q=4.446). [8] Another study was done to identify the stochastic ARIMA model for short term forecasting of hepatitis C virus (HCV) seropositivity among volunteer blood donors in Karachi, Pakistan, ninety-six months (1998-2005) data on HCV seropositive cases among male volunteer blood donors tested at four major blood banks in Karachi, Pakistan were subjected to ARIMA modelling. Subsequently, a fitted ARIMA model was used to forecast HCV seropositive donors for 91-96 months to contrast with observed series of the same months. To assess the forecast accuracy, the mean absolute error rate (%) between the observed and predicted HCV seroprevalence was calculated. Finally, a fitted ARIMA model was used for short-term forecasts beyond the observed series. This short-term forecast beyond the observed series adequately captured the pattern in the data and showed increasing tendency of HCV seropositivity over the forecast interval. [1]

II. Health Data : The Philippines Setting

The Philippines obtains its health data from two major sources: surveys and censuses; and from administrative records of health and health related agencies.

A. *Surveys and Censuses*

The National Demographic and Health Survey (NDHS) is undertaken by the National Statistics Office (NSO) once every 5 years to provide national and regional estimates of levels and trends of fertility as well as examines the differentials and determinants of fertility. It also yields information on family planning, childhood and adult mortality, maternal and child health, and knowledge and attitudes related to HIV/AIDS and other sexually transmitted infections.

B. *Administrative Records of Health and Health Related Agencies*

The Department of Health (DOH) and other government agencies separately collect and maintain data that can be used for monitoring and assessing the health status of the Philippines. The administrative data serves as supplementary data to augment the extensive data required for planning, policy formulation and program interventions on health. With the large contribution of the private sector, monitoring and evaluation of their services are best obtained through administrative data.

The DOH continues to provide / generate the bulk of data for the health sector which are obtained mostly from administrative reports regularly furnished by public hospitals, rural health units (RHUs) and other local government health units.

A major source of data for the DOH is the Field Health Service Information System (FHSIS). It provides information on the different public health programs such as: maternal and child health; nutrition; family planning; Expanded Program on Immunization; Mental health; Communicable Disease Prevention and Control (Tuberculosis, Malaria, Schistosomiasis, Leprosy); Environmental Health; Vital Statistics (natality, mortality, population); and notifiable disease reporting system. Data comes from local field health personnel through the regional and provincial health offices and consolidated at the central office. These are presented by province, city and region in the Annual FHSIS Report. [3]

A team of trained surveillance personnel conducts daily hospital rounds and weekly analysis of data. Every day, they will proceed to the patient information desk where the 24 hours logbook is located. Names of patients admitted within the last 24 hours and whose diagnosis is included among the diseases being monitored are copied. Patients are physically examined to see if they fit the case definition. Analysis of surveillance data and writing of morbidity reports are done weekly.

Surveillance data from each Regional Epidemiology and Surveillance Unit are sent to NEC by mail (diskettes) or e-mail (for regions with access to Internet) monthly for collation and merging. NEC collates all surveillance data from the regions and produces an annual report. [12]

III. Disease Outbreak Detection System

This system is an online system that enables users to generate an approximated number of infected cases provided that it was given a set of data from reported and existing cases. The output will give the public health workers an idea on how the outbreak will behave in the given amount of time. As a result, the public health workers will have a chance to integrate procedures in controlling and preventing the outbreak. Results can be easily interpreted because of the graphs that will be generated using the information the users have inputted.

The system also records and stores information regarding reported cases indicating their personal information, location and condition. This will bring several benefits to the health workers because of faster retrieval of data that will be needed, more organize and clean reports because of easier updating and a faster response to the situation.

The following concepts and tools were essential in the development of the system:

A. *Autoregressive Moving Averages (ARMA)*

The system uses the Autoregressive Moving Averages model to generate prediction values utilizing past records as basis for the computation. ARMA models are mathematical models of the persistence, or autocorrelation, in a time series.

ARMA models are widely used in hydrology, dendrochronology, and many other fields. There are several possible reasons for fitting ARMA models to data. Modeling can contribute to understanding the physical system by revealing something about the physical process that builds persistence into the series. ARMA models can also be used to predict behavior of a time series from past values alone. Such a prediction can be used as a baseline to evaluate possible importance of other variables to the system. ARMA models are widely used for prediction of economic and industrial time series.

ARMA models can be described by a series of equations. The equations are somewhat simpler if the time series is first reduced to zero-mean by subtracting the sample mean. It works with the mean-adjusted series $y_t = Y_t – Y'$, $t = 1 ... N$ where Y_t is the original time series, Y' is its sample mean, and y_t is the mean-adjusted series. One subset of ARMA models are the so-called autoregressive, or AR models. An AR model expresses a time series as a linear function of its past values. The order of the AR model tells how many lagged past values are included. [7]

B. Time Series Analysis

A time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals in statistics, signal processing, and many other fields. Time series analysis comprises methods that attempt to understand such time series, often either to understand the underlying context of the data points (background/data source), or to make forecasts (predictions). Time series forecasting is the use of a model to forecast future events based on known past events: to forecast future data points before they are measured. [4]

C. Equations

R is a freely available language and environment for statistical computing and graphics providing a wide variety of statistical and graphical techniques. It is very similar to a commercial statistics package called S-Plus, which is widely used. [10]

D. Case Fatality Rate (CFR)

Case Fatality Rate is an epidemiological term for death rate of a disease.[5] Its value is determined as follows :

$$CFR= (Deaths / Case\ Count) * 100$$

E. Akaike's Information Criterion (AIC)

Akaike's Information Criterion is a measure of the goodness of fit of an estimated statistical model. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best.

In the general case, the AIC is

$$AIC = 2k - 2 \ln(L)$$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. [2]

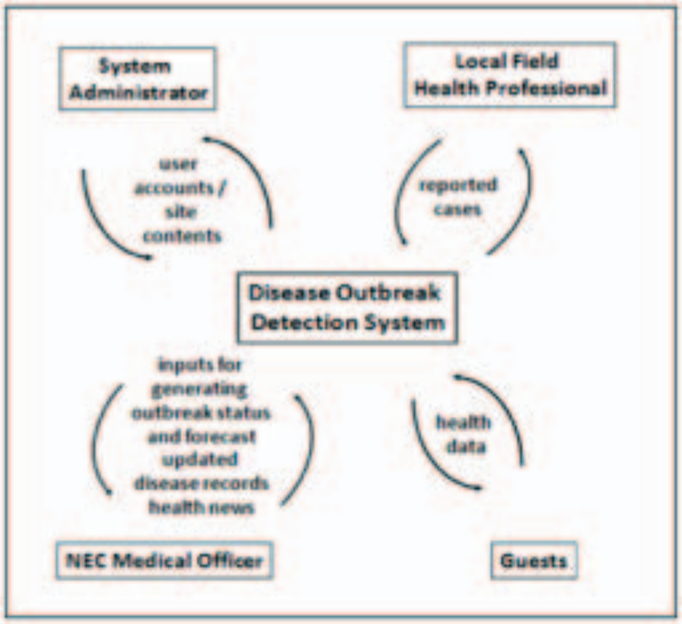
System Administrator
Local Field Health Professional
user accounts / site contents
reported cases
Disease Outbreak Detection System
inputs for generating outbreak status and forecast updated disease records health news
health data
NEC Medical Officer
Guests

Figure 1. Context Diagram of Disease Outbreak Detection System

IV. Results and Discussion

The system has four main users. First is the National Epidemiological Center (NEC) medical officer. He can perform several computation such as ARIMA forecast and Partial AutoCorrelation Function in order to produce accurate prediction values. He can also assess the present condition of a certain outbreak by using several status indicators such as case mapping and graphs that would help identify the behavior of the outbreak in a faster and easier way. Second is the Local Field Health officer. He is the one responsible in managing the case records stored in the system's database which in turn will be the basis in the assessment that will be conducted by the NEC medical officer. System Administrator is the next user. He's duty is to manage the user accounts of each registered users and also, he is responsible in managing the contents of the site. Last is the guest or the anonymous user. He is able to view the published health news about disease outbreaks.

The context diagram of the system is shown in Fig 1. Here, the overview of the interactions of all the users with the system is shown. The role of the NEC medical officer is to perform the computations to generate an approximation and provide status reports that reflect the condition of the outbreak given particular inputs and maintain the disease records. While the local field health personnel only have the ability to manage the patient's record that is stored in the system. Only they have the access to these records. Only the NEC medical officer has the capability to post health news and updates regarding disease outbreak. The role of the System Administrator is to manage the accounts of the users and the contents of the site. Viewing the summary reports about the status of the outbreak whether it is for a particular city or reports in the form of graphs is accessible only to the NEC medical officer.

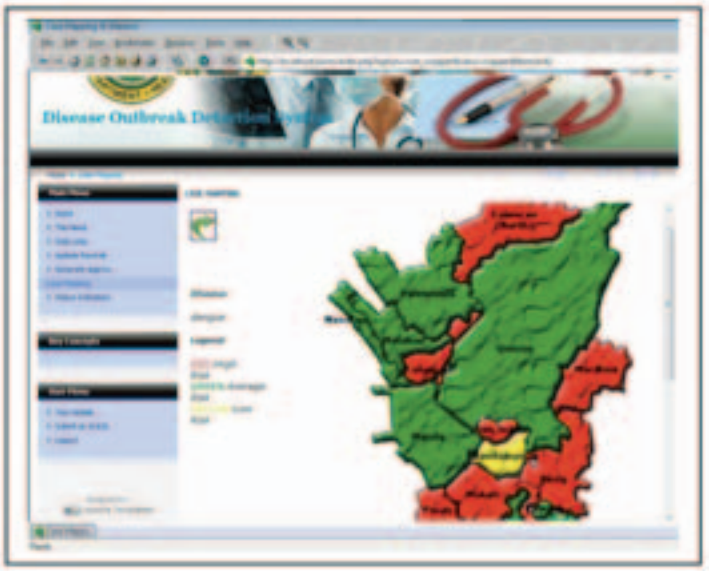

Figure 2. Case Mapping Result Page

The Case Mapping result page is shown in Fig 2. Here, cities were colored according to their risk-level for a particular disease. This page helps users to easily identify cities that require immediate attention.

The Generate Approximation result page is shown in Fig 3. Here, a short summary of the trend of the outbreak for a particular time interval is shown together with the prediction values and the standard error bounds.

The Status Graph Result page is shown in Fig 4. This page displays the appropriate graph that will reflect the outbreak status of a particular disease. In the figure, we can see that the frequency of the case reported per month is displayed for easy analysis of the behaviour of the outbreak.

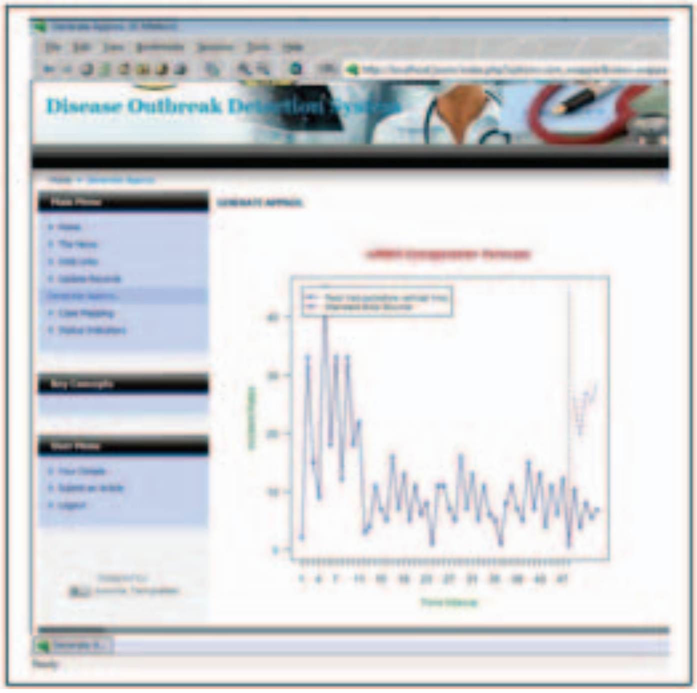

Figure 3. Generate Approximation Results Page

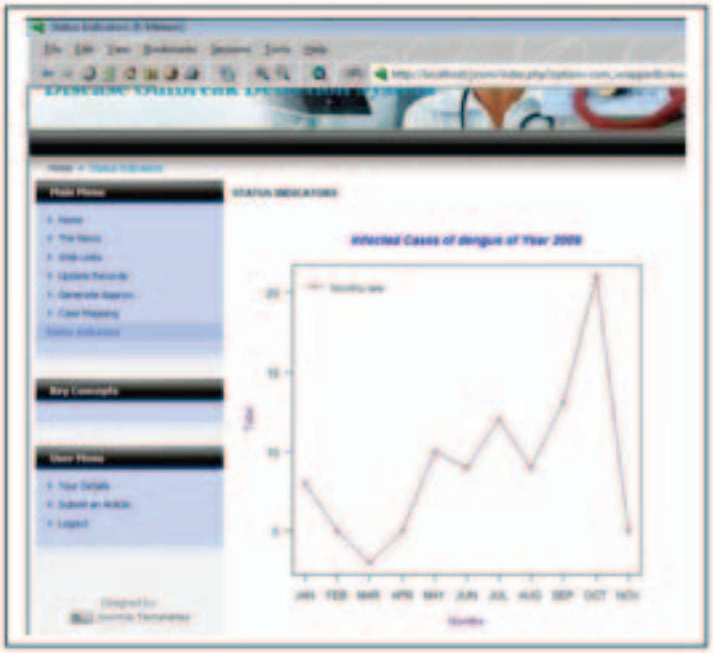

Figure 4. Page Status Graph Result Page

In general, Disease Outbreak Detection System provides useful tool for health personnel in analysing and understanding the behaviour and trend of an outbreak. The outputs generated using the system will reflect the current or even future status of the outbreak giving health personnel basis in their decision-making and in planning their actions to prevent and control an outbreak. Furthermore, Also, the system solves the problem of inconsistency and errors brought by consolidating data from local health personnel to the NEC. The system solves this because as the local field health personnel enter data in the system, it will be automatically directed to NEC and can be quickly used to generate an analysis of the outbreak.

Conclusion and Future Work

The Disease Outbreak Detection System is an online system developed to aid public health workers detect and track the spread of diseases. It specifically helps epidemiologists analyze the behavior of a certain disease outbreak by providing them prediction values for a specific time interval. The system is able to perform such feature with the aid of R software which performs computations of Time Series analysis using Autoregressive Moving Averages (ARMA) Model to generate values based on the present condition of the outbreak. These generated values will serve as basis to know how the outbreak will turn out, thus giving the epidemiologist sufficient time to respond to major public health threats and formulate preventive measures to control and solve the outbreak. The particular tool was developed for the Philippine setting, specifically for the use of outbreak monitoring agencies such as the National Epidemiological Center (NEC), and thus uses Philippine health data from surveys and censuses, as well as from administrative records of health and health related agencies.

One thing that can be done to the system modify it to accommodate other models that epidemiologists may wish to use. Another future work is integration with Geographic Information System (GIS) which can provide other facts/ demographics that can provide insights on certain locations on factors that may affect the spread of diseases.

Acknowledgment

G. A. Solano is supported by the Engineering Research and Development for Technology (ERDT) Scholarship Program of the Department of Science and Technology(DOST).

References

1. S. Akhtar, and S. Rozi ”An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan”. World Journal of Gastroenterology. 2009 April 7; 15(13): page 1607-1612
2. H. Akaike, "A new look at the statistical model identification". IEEE Transactions on Automatic Control 19 (6): 716–723. 1974. doi:10.1109/ TAC.1974.1100705. MR 0423716.
3. E.V. Domingo,"Administrative Data Key to Health Policies and Programs: Philippine Experience ". Inter-Secretariat Working Group-Health Statistics Meeting International Association for Official Statistics Conference. Shanghai China. 14-17 October 2008.
4. J. Hamilton, Time Series Analysis, Princeton: Princeton Univ. Press, 1994. ISBN 0-691-04289-6
5. F.C.K, Li, B.C.K. Choi, T. Sly, A.W.P. Pak . "Finding the real case-fatality rate of H5N1 avian influenza". Journal of Epidemiology and Community Health 62 (6): 555–559. June 2008. doi:10.1136/jech. 2007.064030. ISSN 0143-005X. PMID 18477756. Retrieved 2009-04-29.
6. H. M. Lu, D. Zeng, H, Chen, "Prospective Infectious Disease Outbreak Detection Using Markov Switching Models," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 4, pp. 565-577, April 2010, doi:10.1109/TKDE.2009.115
7. D . Meko, 2011, lecture notes on Applied Time Series Analysis, Laboratory of Tree-Ring Research, University of Arizona on Spring 2011
8. S. Promprou, M. Jaroensutasinee and K.. Jaroensutasinee”Forecasting Dengue Haemorrhagic Fever Cases in Southern Thailand using ARIMA Models”.School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhonsithammarat 80161, Thailand.
9. KJ Ryan; CG Ray (editors) (2004). Sherris Medical Microbiology (4th ed.). McGraw Hill. ISBN 0-8385-8529-9.
10. G. Shaddick,” Using R (with application in Time Series analysis)”. January 2004. p.4.
11. J.S. Zhongguo, X.Y.J.S. Chong, B.Z.Z. Chong, ”Study on the feasibility for ARIMA model application to predict malaria incidence in an unstable malaria area”.PBMed. 2007 Jun;25(3):232-6.
12. “NESSS Annual Report”. Office of the Secretary, Department of Heath. Republic of the Philippines.2001
13. Public Library Of Science (2005, February 18). A New Method For Early Detection Of Disease Outbreaks. ScienceDaily. Retrieved May 11, 2012,from http://www.sciencedaily.com- /releases/2005/02/ 050218130731.htm