

# Computational Prediction of Cholera Outbreaks

MARTIN MUBANGIZI

ERNEST MWEBAZE

JOHN A. QUINN

Faculty of Computing and I.T., Makerere University, Kampala, Uganda

[mmubangizi, emwebaze, jquinn]@cit.mak.ac.ug

---

Timely detection of cases of cholera is essential in order to limit the severity and duration of an outbreak. By the time that reports of cholera are received, however, an outbreak may already be well underway. This paper describes methodologies to take data from different sources and apply machine learning techniques to predict the risk of cholera outbreak over time in different areas. Using data collected across 80 regions in Uganda, we first analyse regions which have similar dynamics in cholera prevalence over time. We then formulate a probabilistic model to predict future levels of cholera cases and demonstrate its operation.

Categories and Subject Descriptors: I.2.1 [Computing Methodologies]: Artificial Intelligence – Medicine and Science.

General Terms: Disease outbreak prediction, clustering, dynamical Bayesian networks.

---

## 1. INTRODUCTION

Timely detection of cases of cholera is essential in order to limit the severity and duration of an outbreak. By the time that reports of cholera are received, however, an outbreak may already be well underway.

Background knowledge is useful to preempt reports of infection in such cases. By observing, e.g. that there has been heavy rainfall in an area which has previously been affected by cholera, a high risk of repeated outbreak can be predicted. This forewarning can be used to plan the best use of available resources to mitigate the risk (e.g. sanitation engineering), or to provide response teams with a “red alert”. However, this insight is currently applied ad-hoc, without being explicitly quantified.

This paper describes methodologies to take data from different sources and uses machine learning techniques to predict the risk of cholera outbreak over time in different areas. Other work has looked at prediction of cholera in a non-probabilistic setting, see e.g. [Pascual et al. 2000; Codeco 2001; Cazelles and Hales 2006; de Magny et al. 2008]. Bayesian techniques have been recently applied in the field of biosurveillance [Wong et al. 2004], though mainly for problems where we have data about the effects rather than the causes of an epidemic.

The data used in the work is summarised in section 2. In section 3 we cluster different areas together, grouping together areas with similar historical patterns of disease occurrence. In section 4 we describe the dynamical probabilistic model used for representing the evolution of an outbreak, and give details of parameter estimation and prediction in section 5.

## 2. EXPERIMENTAL DATA

Data on disease prevalence and related factors were used to analyse patterns of cholera occurrence and the relationship between cholera and environmental factors.

### 2.1 Epidemiological Data

The Ministry of Health of Uganda set up the Health Management Information System (HMIS) as the primary tool for collecting epidemiological data from all areas of the country. A special section of the HMIS is devoted to the Integrated Disease Surveillance and Reporting unit that is tasked with monitoring specific diseases that are epidemic prone. Currently this unit is monitoring 12 diseases including Acute Flaccid Paralysis, Rabies, Cholera, Dysentery, Guinea Worm, Malaria, Measles, Meningitis, Neonatal tetanus, Plague, Yellow Fever and other Viral Hemorrhagic Fevers.

Epidemiological data for this study was obtained from the Epidemiology and Surveillance Unit at the Ugandan Ministry of Health. The data included weekly measurements from 80 districts in Uganda for every week from January 2003 to September 2007. This data was presented as cases of disease events and the corresponding deaths for every week.

### 2.2 Climatic Data

There is known to be a strong connection between between climatic conditions and outbreaks of cholera. Climatic data was collected from seven meteorological centres, covering most districts in Uganda, for selected years from 1960-2008, although there are missing data occurrences. In this study, daily rainfall, humidity (at both 0600Hrs and 1200Hrs), and temperature measurements were used; weekly averages for each were derived except in the case of rainfall where maximum measurements were used.

## 3. ANALYSIS OF RELATED LOCATIONS

To aid the process of building a probabilistic model to make predictions based on the observational data, a process of clustering was undertaken to group the distinct districts into groups with similar characteristics, based on a weekly history of cholera outbreaks over a period of 5 years.

### 3.1 Clustering Algorithm

We used a generic clustering algorithm that employs an Euclidean distance measure to quantify the similarity of a given vector point representing cholera incidences in a district over time, with a particular class or cluster. The Euclidean distance for two vectors  $x, y \in R^N$  can be defined as

$$d(x, y) = \sqrt{\sum_{j=1}^N (x_j - y_j)^2}. \quad (1)$$

In our case,  $x_j$  and  $y_j$  are the number of cholera cases in week  $j$  for areas  $x$  and  $y$ . Since all these vector points inherently have the same meaning, a distance measure such as Euclidean distance, we found, suffices to classify the data. The algorithm

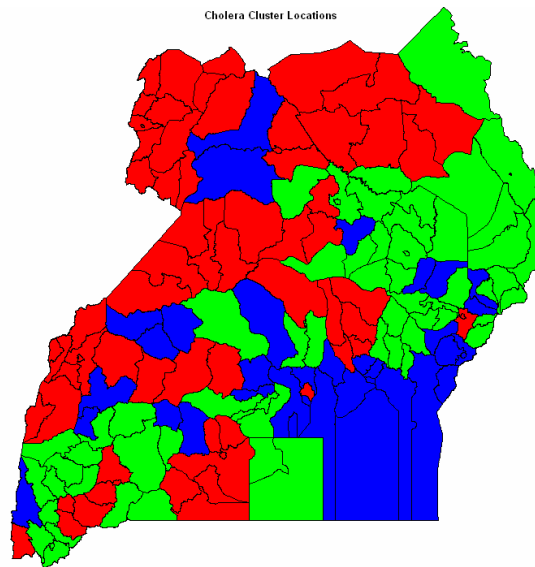


Fig. 1. Map showing the 80 districts of Uganda clustered according to similarity in history of cholera prevalence for the years 2003 – 2007. Green indicates cluster 1; blue indicates cluster 2; red indicates cluster 3.

initially obtains an Euclidean distance matrix based on the data. A hierarchical cluster tree is then generated from the distances in the distance matrix and then thresholded to give the different clusters.

### 3.2 Location Clusters

Figure 2 shows the clustering of the 80 districts of Uganda based on the weekly incidences of cholera over the 5 years 2003-2007. Three clusters shaded in different colors were generated by the algorithm representing similarities within the districts in each cluster. The clusters as shown in Figure 2 appear to a great extent to be geographically related, cluster 1 (red) forming most of the North/Eastern part of the country, cluster 2 (green) forming most of the central part and cluster 3 (blue) forming the rest of the Northern/Western part of the country.

An analysis of total number of cholera cases for the different districts helps to explain the basis for this categorisation of the districts. Figure 2 shows the total number of cases per district in each cluster. The areas in cluster 1 (31 districts) have consistently had no reports of cholera for the time of the recordings; cluster 2 has had a low number of cases (18 districts, 1–10 reports of cholera); and cluster 3 contains areas with a severe number of cases (31 districts, 11–5175 reports of cholera).

Note that in Figure 1, clusters 2 and 3 appear to be made up of a number of smaller sub-groups. In future we aim to apply hierarchical clustering, or clustering into a greater number of classes  $k$ , to see if there are significant patterns in the data at a finer level.

The analysis of related areas is useful for splitting up the area into separate

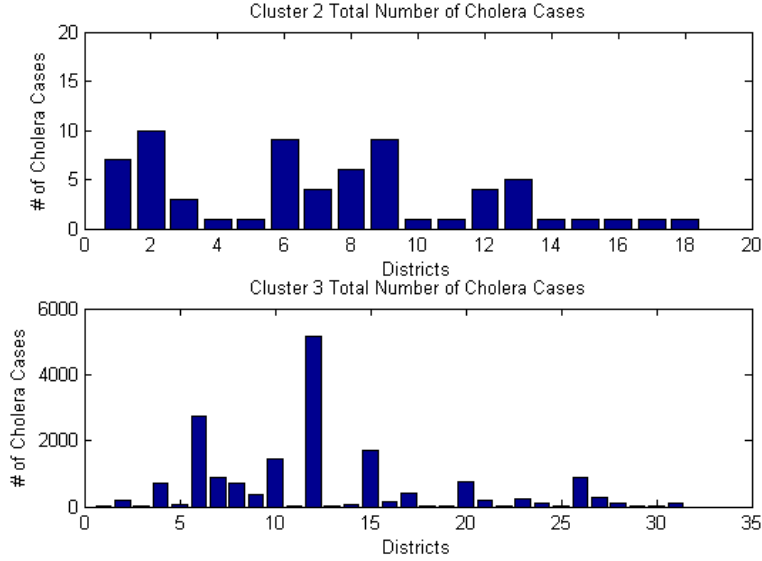


Fig. 2. Numbers of recorded cholera cases in the 18 regions of cluster 2 (top) and the 31 regions of cluster 3 (bottom). There were no recorded cases of cholera in the regions of cluster 1.

classes, in order to process areas with similar dynamics together in the same sequential model. We now describe the dynamical modelling component of this work.

#### 4. MODEL FOR OUTBREAK PREDICTION

Figure 3 shows a dynamical Bayesian network (DBN) [West and Harrison 1999] modelling the evolution in time of cholera infection rates in a particular area, affected by rainfall, temperature, and humidity (both at 0600Hrs and at 1200Hrs). We model the number of cases of infection as being directly affected by the number of cases the week before, and the average humidities, average temperature and maximum rainfall for that week.

We take each variable in the model to be continuous and Gaussian distributed. The mean of the disease rate is taken to be a weighted sum of the dependent variables plus some constant. Each variable is therefore distributed as:

$$r_t \sim \mathcal{N}(\mu_r, \sigma_r^2)$$

$$h_t \sim \mathcal{N}(\mu_h, \sigma_h^2)$$

$$p_t \sim \mathcal{N}(\mu_p, \sigma_p^2)$$

$$q_t \sim \mathcal{N}(\mu_q, \sigma_q^2)$$

$$\log(d_t) | \log(d_{t-1}), r_t, h_t, p_t, q_t \sim \mathcal{N}(w_1 r_t + w_2 h_t + w_3 p_t + w_4 q_t + w_5 \log(d_{t-1}) + \mu_d, \sigma_d^2).$$

We deal with all disease rates in log space. This ensures that we never predict a negative count, and is equivalent to a Poisson regression model. The joint probability

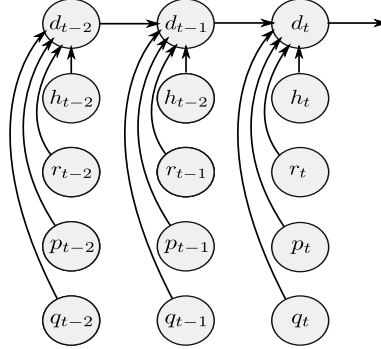


Fig. 3. The variable  $d_t$  denotes the disease rate at time  $t$ ,  $h_t$  denotes the temperature,  $r_t$  denotes the amount of rainfall,  $p_t$  denotes the humidity at 0600Hrs and  $q_t$  denotes the humidity at 1200Hrs.

distribution for a sequence of  $N$  observations under this model is given by

$$P(r_{1:N}, h_{1:N}, p_{1:N}, q_{1:N}, d_{1:N}) = P(d_1|r_1, h_1, p_1, q_1) \left[ \prod_{n=1}^N P(r_n)P(h_n)P(p_n)P(q_n) \right] \prod_{n=2}^N P(d_n|d_{n-1}, r_n, h_n, p_n, q_n). \quad (2)$$

Given appropriate parameter settings, we are then in a position to make predictions for future values of the sequence.

## 5. LEARNING AND INFERENCE

The parameters of the model (the means  $\mu_r, \mu_h, \mu_p, \mu_q, \mu_d$ , the weights  $w_1, w_2, w_3, w_4, w_5$  and variances  $\sigma_r^2, \sigma_h^2, \sigma_p^2, \sigma_q^2, \sigma_d^2$ ) were learnt separately for each region using maximum likelihood. That is, given observed sequences of data  $r_{1:N}, h_{1:N}, p_{1:N}, q_{1:N}$  and  $d_{1:N}$ , we find the parameter settings which maximise Eq. (2). The weights, for example, are found using the maximum likelihood estimator

$$\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top d_{2:N} \quad (3)$$

where the columns of matrix  $X$  are  $r_{2:N}, h_{2:N}, p_{2:N}, q_{2:N}, d_{1:N-1}$ .

Data imputation for missing values was done by first initializing missing data to averages and later refining estimates by performing iterations of Expectation (finding expected values of the missing data points based on the current parameter estimates) and Maximization (finding the maximum likelihood parameters based on the current expected complete data). The data was then split into training sets and test sets, with two thirds of the data used for training. We trained on a number of different areas, allowing us to make predictions for those areas and compare the predicted number of cases each week with the actual number. We show the week-by-week predictions in Figure 4. We used the Bayes Net Toolbox for Matlab [Murphy 2007] to carry out experiments.

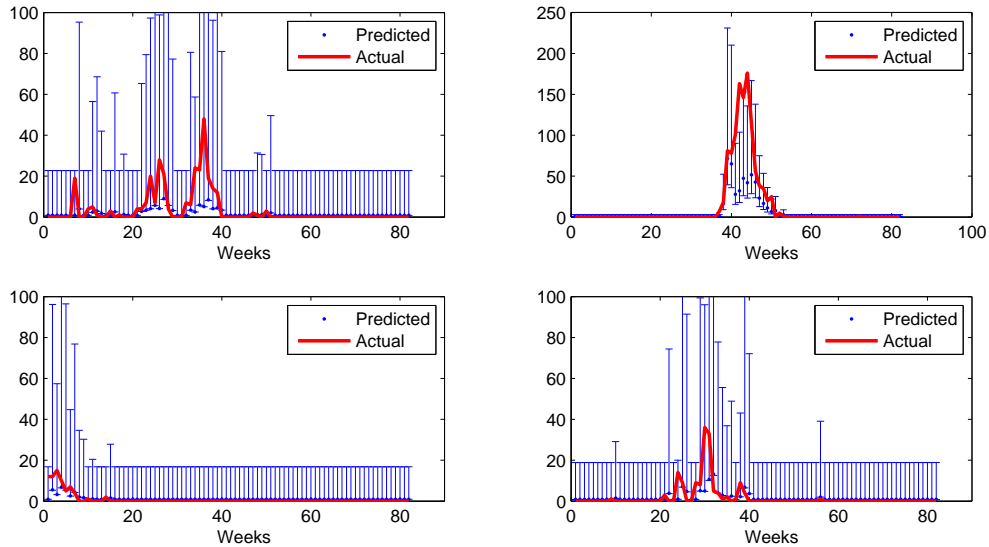


Fig. 4. Week ahead predictions for the number of cholera cases for the districts of Kasese(top left), Masindi(top right), Gulu (bottom left) and Arua (bottom right), plotted against the actual number of cases. Each time frame is one week. Error bars represent two standard deviations of the predicted distributions of the number of cases.

### 5.1 Learning within clusters of areas

In addition to training and testing with data from a single region, we experimented with aggregating training data from multiple regions in the same cluster. We used data from 5 districts of Kasese, Kampala, Masindi, Gulu and Arua. All these districts are from cluster 3. The choice of these districts was based on the magnitude of number of cases of cholera and the availability of reasonably complete climatic data. Data from the 5 districts were used to set parameters for the model of each district.

We evaluate the predictions by first calculating whether there was an outbreak or not in each time frame (class 1 if the number of cases of cholera was above some threshold, 0 otherwise). These gold standard classes were compared with the model's predictions by means of a Receiver Operating Characteristic (ROC) curve. From Fig. 5, the accuracy of the predictions are found to be the same at threshold=0, but at threshold=5 training with data from a single region overall gave better predictions than training with data from multiple regions. We therefore cannot conclude that training using data from different regions improves on predictions, possibly because our testing data was limited. Future experiments with more data will be used to clarify this. The prediction accuracy in both cases was good as is evident from the values of the Area Under the ROC Curve (AUC) in Fig. 5. Looking at the ROC curves it can be seen that with the models constrained to allow 5 percent false positive rates, would give true positive rates of up to 70 percent. That is, if we determined that a 5 percent false alarm rate was acceptable,

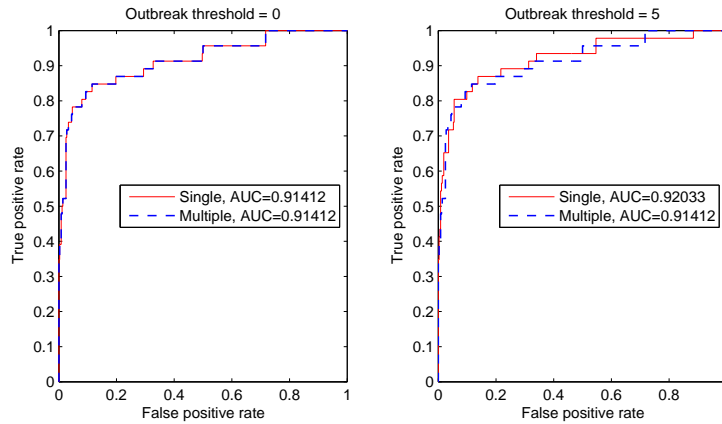


Fig. 5. Receiver Operating Characteristic (ROC) curve with outbreak thresholds set at 0 and 5 cases per week. “Single” indicates that each area was trained independently, whereas “Multiple” indicates that training data from districts in the same cluster was aggregated.

we would expect to correctly predict 70 percent of outbreaks a week in advance.

## 6. CONCLUSION

In the first part of this work, we have shown that areas of disease prevalence can be clustered to give regions with similar historical dynamics. We then formulated a dynamical probabilistic model which can be used to make predictions of cholera infection rates given the history of infection in that area and climatic data.

There are several directions in which both of these components can be extended. Hierarchical clustering may reveal groups of districts with more subtle similarities. By showing which districts are mutually informative about each other, we also have a basis for knowing whether a cholera outbreak in one district implies a risk in a neighbouring district.

In terms of dynamical modelling, we aim to investigate whether accuracy of predictions increases consistently when we try to learn the dynamics of outbreak when looking at clusters of areas. We also aim to add other variables to the model which influence cholera prevalence, for example environmental and demographic factors such as latrine coverage, drainage and average income.

## 7. ACKNOWLEDGMENTS

We thank Luswa Lukwago and Joseph Wamala for expert advice about cholera prediction. Disease prevalence data was provided by the Ugandan Ministry of Health, and climatic data was provided by the Ministry for Water, Lands and the Environment. The project was supported in part by an IBM Faculty Award.

## REFERENCES

- CAZELLES, B. AND HALES, S. 2006. Infectious diseases, climate influences, and nonstationarity. *PLoS Med* 3, 8, epub.
- CODECO, C. 2001. Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. *BMC Infect Dis* 1, 1, epub.

- DE MAGNY, G. C., MURTUGUDDE, R., SAPIANO, M., NIZAM, A., BROWN, C., BUSALACCHI, A., YUNUS, M., NAIR, G., GIL, A., LANATA, C., CALKINS, J., MANNA, B., RAJENDRAN, K., BHATTACHARYA, M., HUQ, A., SACK, R., AND COLWELL, R. 2008. Environmental signatures associated with cholera epidemics. *Proc Natl Acad Sci USA* 105, 46, 17676–81.
- MURPHY, K. 2007. The Bayes Net Toolbox for Matlab. Tech. rep., University of British Columbia. <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>; accessed February 07 2008.
- PASCUAL, M., RODO, X., ELLNER, S., COLWELL, R., AND BOUMA, M. 2000. Cholera Dynamics and El Nino-Southern Oscillation. *Science* 289, 5485, 1766–1769.
- WEST, M. AND HARRISON, J. 1999. *Bayesian Forecasting and Dynamic Models*. Springer.
- WONG, W., MOORE, A., COOPER, G., AND WAGNER, M. 2004. Bayesian Biosurveillance of Disease Outbreaks. In *Proc. Uncertainty in Artificial Intelligence*.