

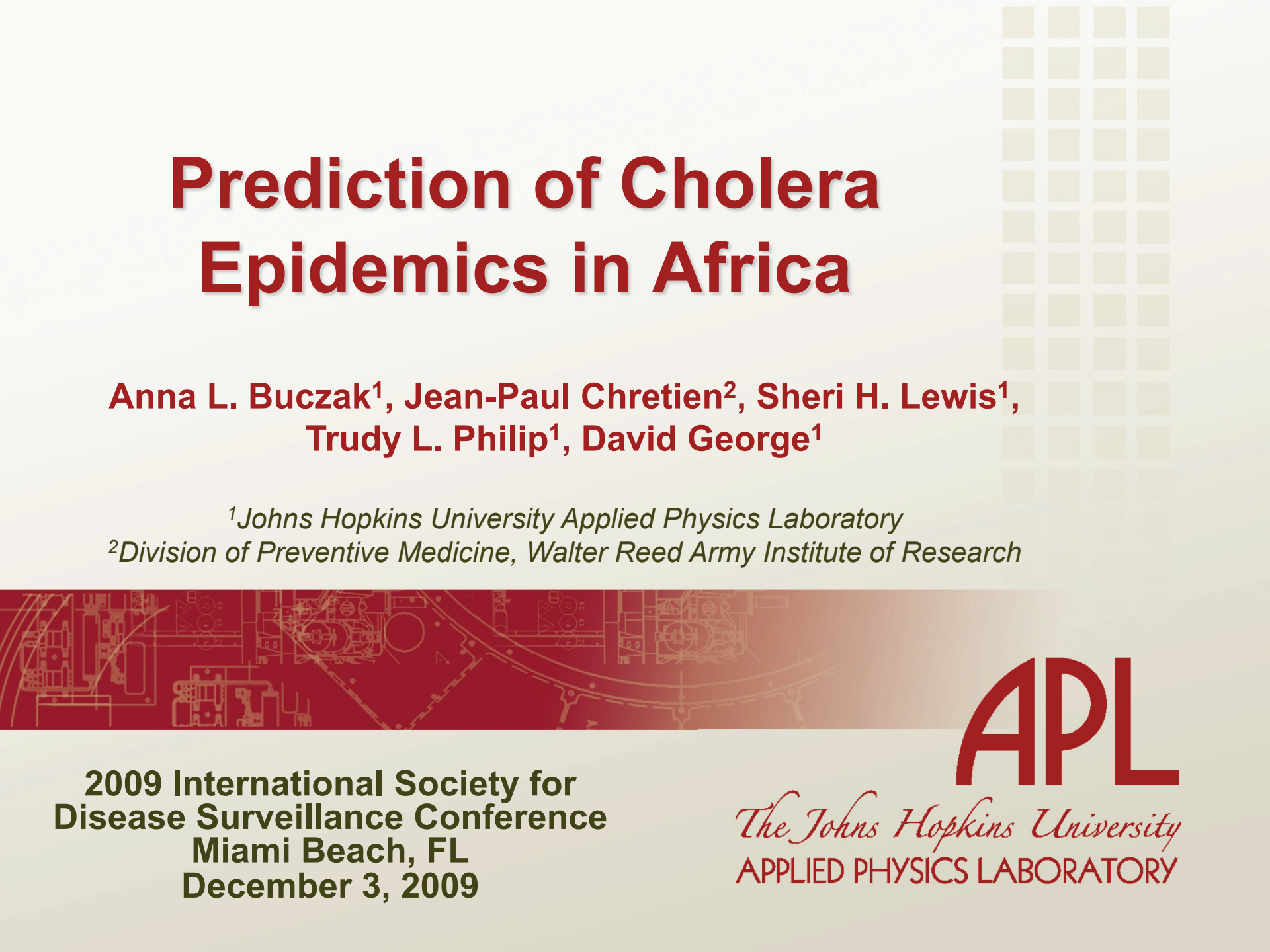
Prediction of Cholera Epidemics in Africa

Anna L. Buczak¹, Jean-Paul Chretien², Sheri H. Lewis¹,
Trudy L. Philip¹, David George¹

¹*Johns Hopkins University Applied Physics Laboratory*

²*Division of Preventive Medicine, Walter Reed Army Institute of Research*

**2009 International Society for
Disease Surveillance Conference
Miami Beach, FL
December 3, 2009**



APL

The Johns Hopkins University
APPLIED PHYSICS LABORATORY

Outline

- Background
- Related work
- Epidemiological data
- Predictor data
- Results
 - Logistic regression
 - Support Vector Machines
- Conclusions

Background

- Many climate-infectious disease links known, but need for operational forecasting systems (WHO, 2001)
- **Our focus: cholera in Africa**
 - Cholera - important global health problem:
 - ~1-3 million cases/year, ~1-2% CFR (WHO)
 - Annual economic losses likely in \$ billions
 - Complicates humanitarian emergencies
 - Most reported cases in Africa
 - No prediction systems (to our knowledge) but:
 - Accurate predictions could be actionable
 - Epidemiological data is available
 - Predictors are known in some areas

Related Work

- **Existing methods:**
 - **Limited areas (Bangladesh, Peru, Nigeria, South Africa)**
 - **Limited predictor types (often just climate)**
 - **Frequently correlations established without prediction methodology being developed:**
 - **Correlation¹ between Sea Surface Temperature (SST) in Pacific and cholera spring deaths in Bangladesh ($R^2 = 0.34$)**
 - **Correlation² between precipitation and cholera incidence in KwaZulu-Natal, South Africa ($R^2 = 0.74$)**
 - **Performance characteristics not rigorously assessed**

¹Bouma & Pascual . Seasonal and interannual cycles of endemic cholera in Bengal 1891-1940 in relation to climate and geography. Hydrobiologia 2001.

²Mendelsohn & Dawson. Climate and cholera in KwaZulu-Natal, South Africa: the role of environmental factors and implications for epidemic preparedness. Int J Hyg Environ Health 2008.

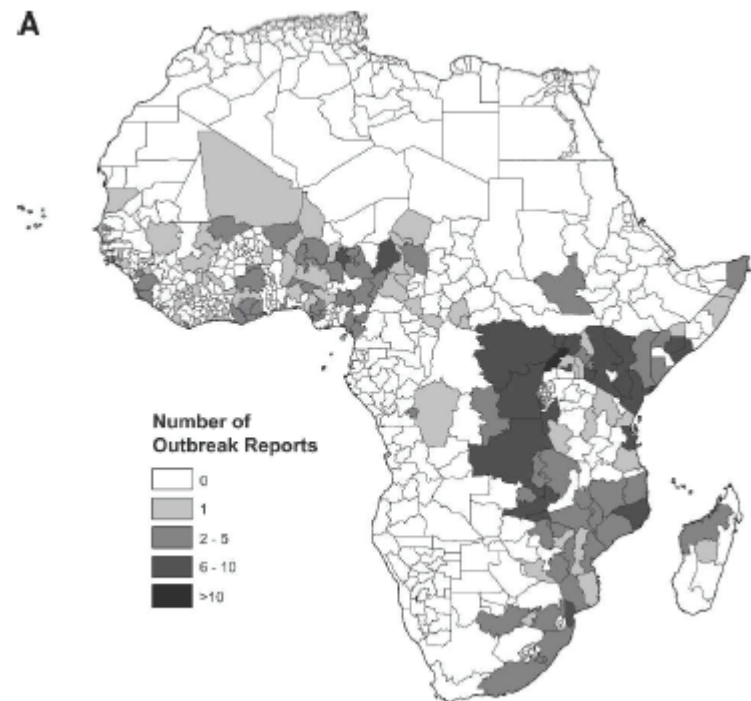
Availability of Epidemiological Data

TABLE 1

Number of outbreaks and total number of reported cholera cases by region and sub-region

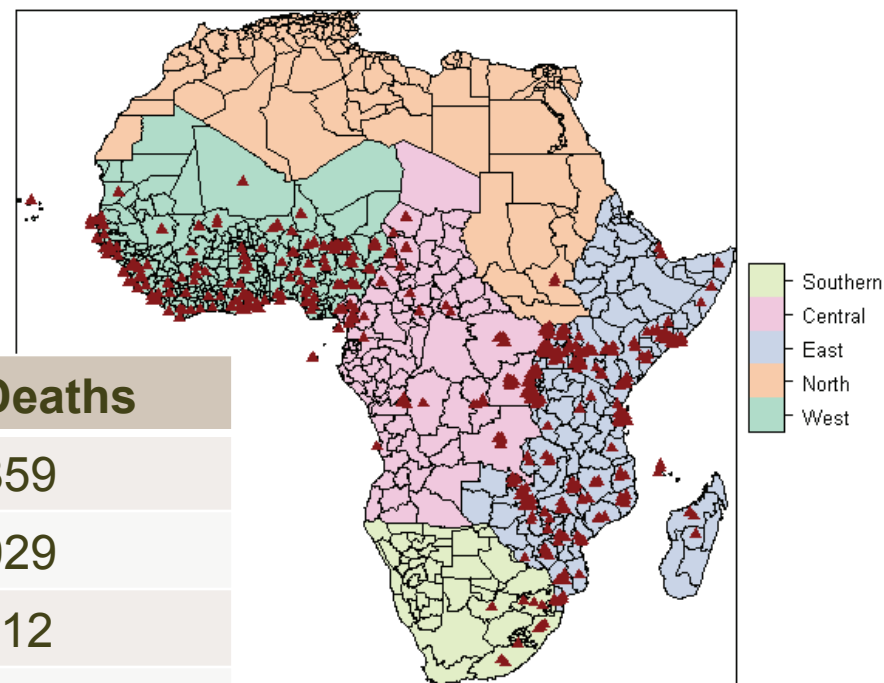
Region/sub-region	Number of outbreaks (% of total)	Total number of cases (% of total)
Africa		
West	124 (19.6)	123,012 (25.4)
Central	67 (10.6)	72,250 (14.9)
East	127 (20.1)	36,823 (7.6)
South	99 (15.7)	191,819 (39.6)
Total	417 (66.0)	423,904 (87.6)
Americas		
Central	17 (2.7)	1,794 (0.4)
South	28 (4.4)	11,305 (2.3)
Total	45 (7.1)	13,099 (2.7)
Europe		
Eastern	8 (1.3)	479 (0.1)
Total	8 (1.3)	479 (0.1)
Eastern Mediterranean		
North Africa	2 (0.3)	892 (0.2)
West Asia	34 (5.4)	28,213 (5.8)
Total	36 (5.7)	29,105 (6.0)
South-East Asia		
Southeastern Asia	44 (7.0)	8,112 (1.6)
South Asia	62 (9.8)	7,076 (1.5)
Total	106 (16.8)	15,188 (3.1)
Western Pacific		
East Asia	18 (2.9)	835 (0.2)
Oceania	2 (0.3)	1,636 (0.3)
Total	20 (3.2)	2,471 (0.5)
Global total	632 (100)	484,246 (100)

Based on ProMed (1995-2005)



Griffith et al. Review of reported cholera outbreaks worldwide, 1995-2005. Am J Trop Med Hyg 2006.

ProMed: Outbreaks by Region

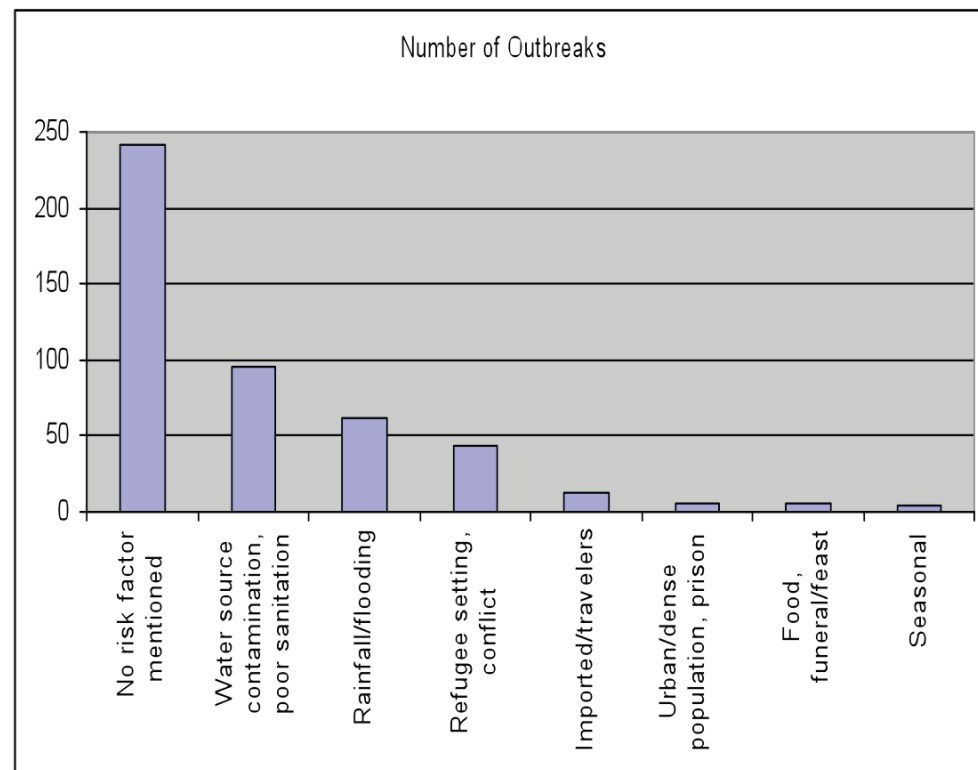


Region	# Outbreaks	# Cases	# Deaths
Eastern	257	175,221	6,859
Western	137	126,684	4,029
Middle	60	77,377	4,112
Southern	11	106,929	272
Northern	2	892	384
Total	467	487,103	15,656

Our efforts are concentrated on Eastern, Western, Middle and Southern Africa

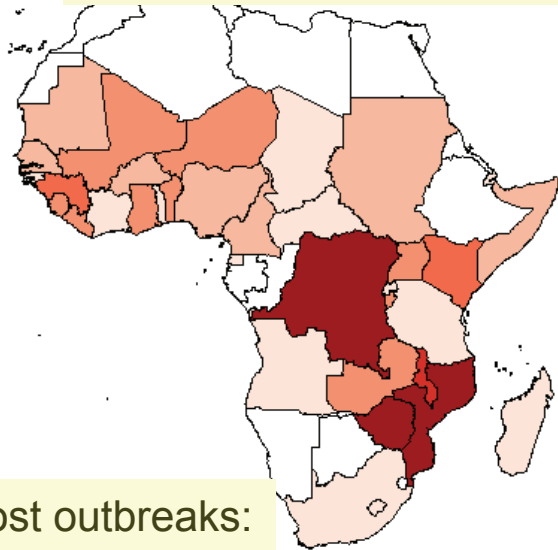
ProMed: Outbreak Causes

- No risk factor mentioned (241)
- Water source contamination, lack of potable water, poor sanitation (95)
- Rainfall/flooding (62)
- Refugee setting, conflict (43)
- Imported/travelers (12)
- Urban/dense population, prison (5)
- Food, funeral/feast (5)
- Seasonal (4)
- Total: 467



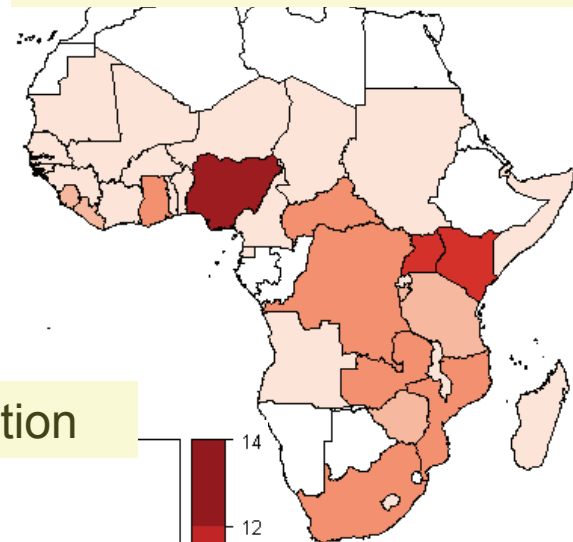
Number of Cholera Outbreaks Attributed to Different Causes

Flooding



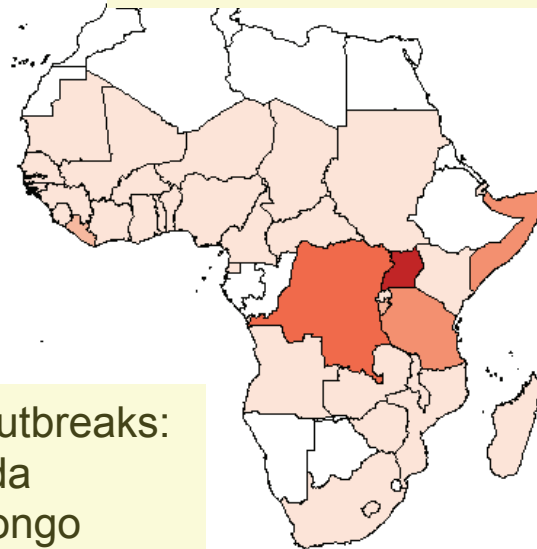
Most outbreaks:
•DR Congo
•Zimbabwe
•Mozambique

Water Contamination or Poor Sanitation



Most outbreaks:
•Nigeria
•Uganda
•Kenya

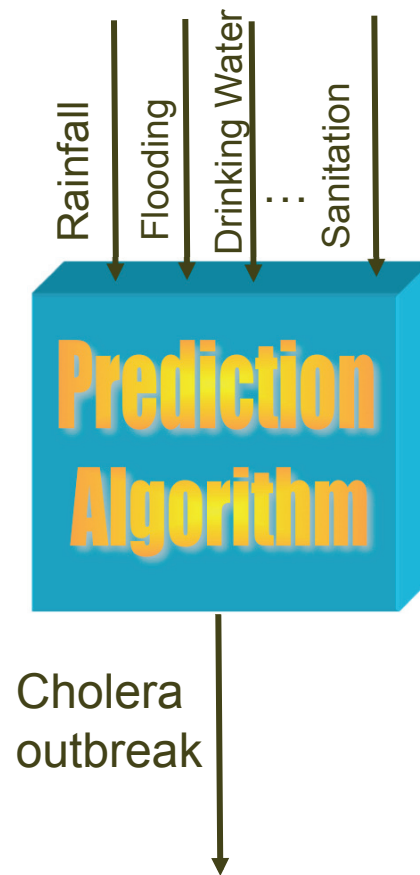
Refugee Situation



Most outbreaks:
•Uganda
•DR Congo
•Tanzania

Predictors and Their Frequency

- Predictors of interest:
 - Rainfall (monthly)
 - Flooding (monthly)
 - Percentage Improved Drinking Water (yearly at best)
 - Percentage Improved Sanitation (yearly at best)
 - Southern Oscillation Index (monthly)
 - Sea Surface Temperature (monthly)
 - Refugee situation (monthly)
 - Gross National Income Per Capita (yearly)
 - Total Health Spending Per Capita (yearly)
 - Mean Population Density (yearly at best)
 - Urban mean (yearly at best)
 - Life Expectancy at Birth (yearly at best)
 - Flood susceptibility (one value)

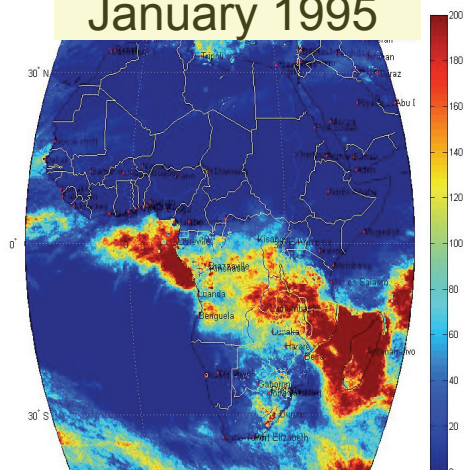


Prediction performed at province level. Cholera outbreaks to be predicted one month in advance.

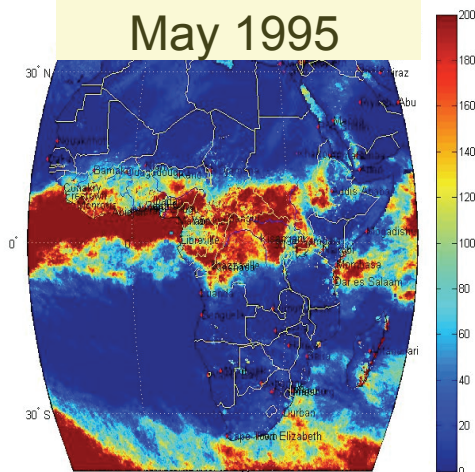
Rainfall Data

- Rainfall data for 1995-2008 obtained from NASA
- Data processed to obtain the average rainfall for each province in each country (for each month)

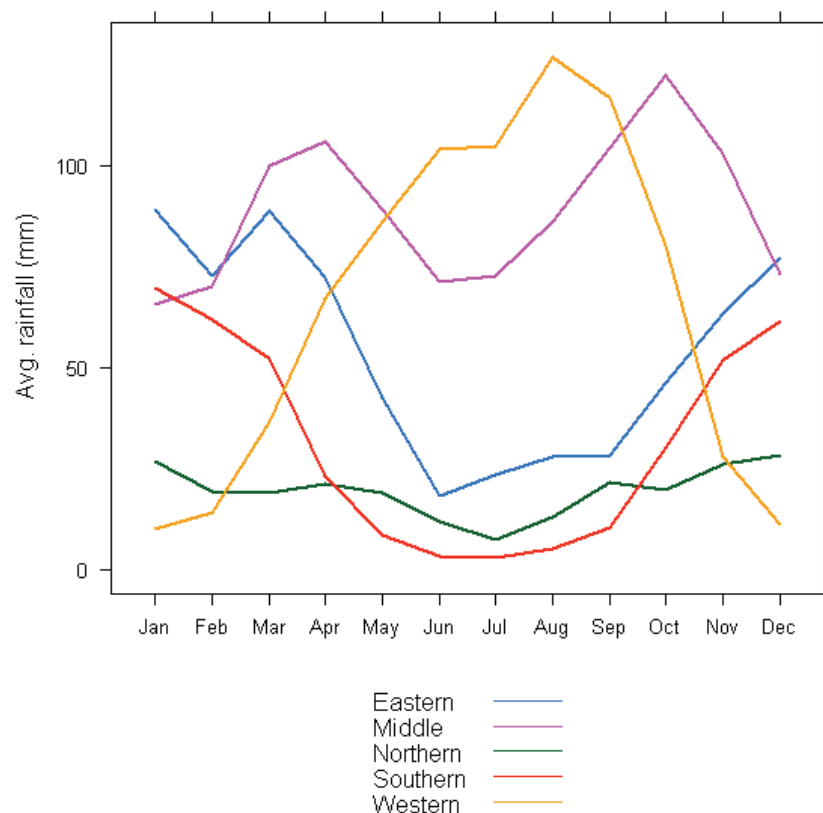
January 1995



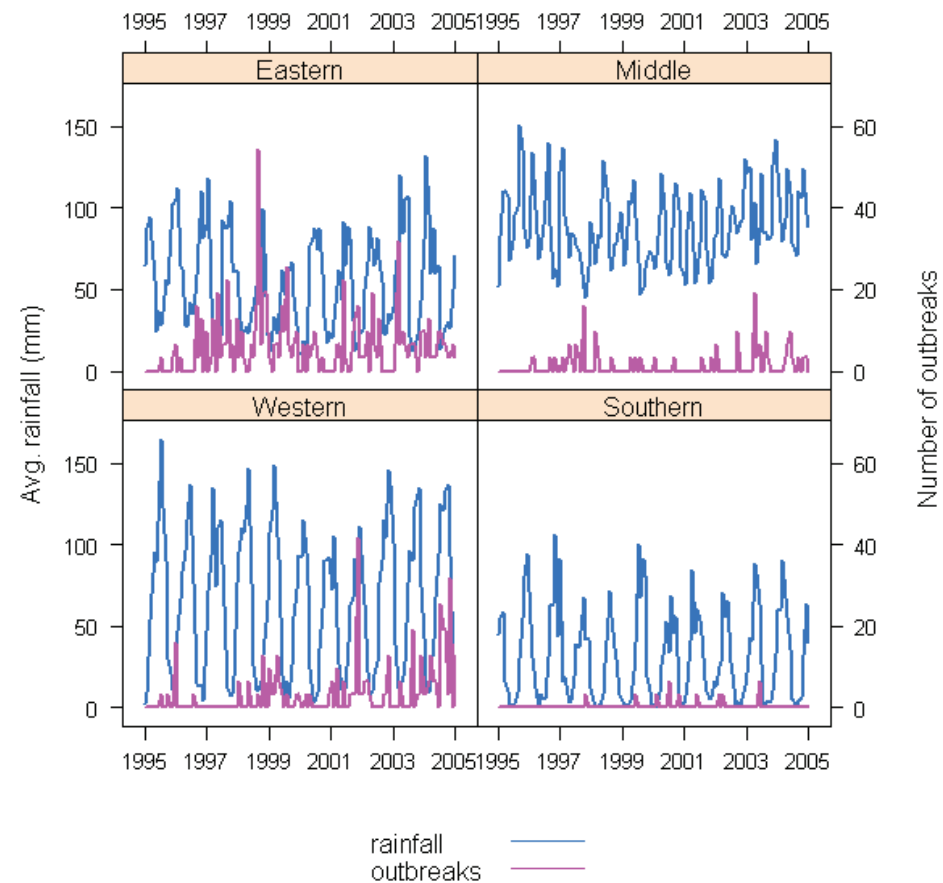
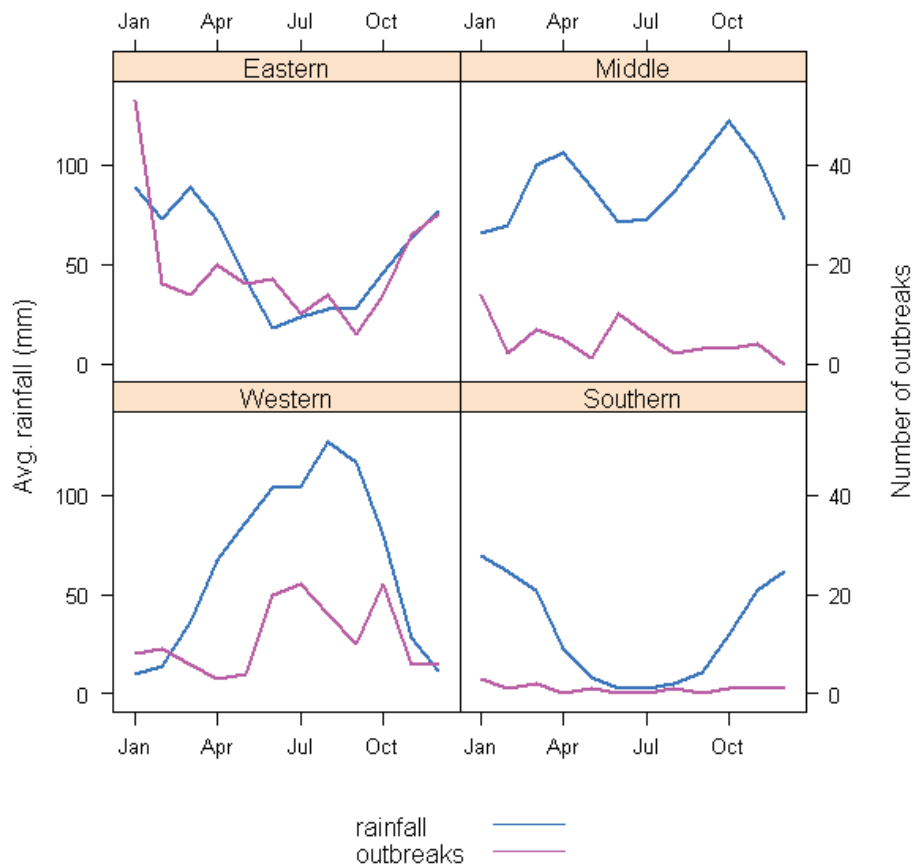
May 1995



Rainfall Monthly Average by Region

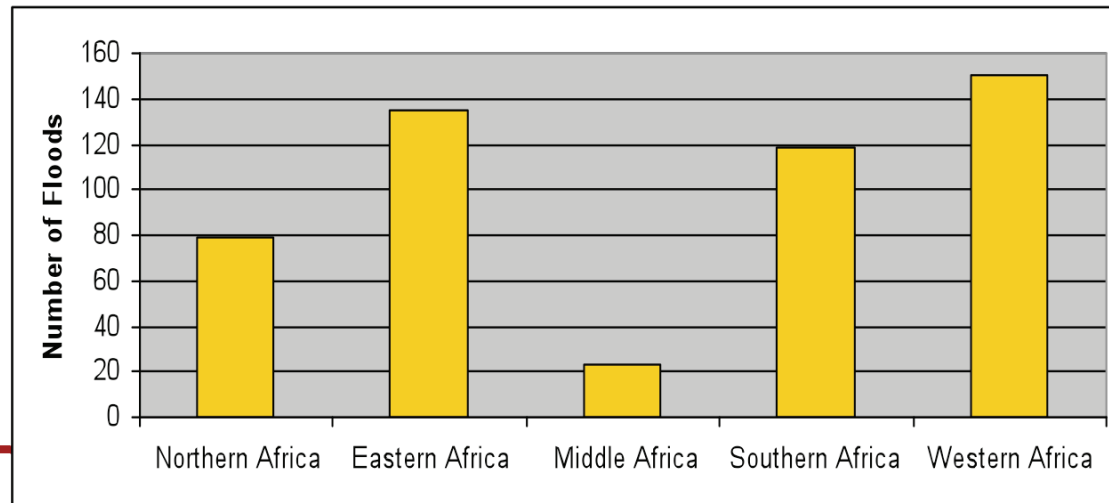


Rainfall and Outbreaks



Flood Data

- **Computation of flooding:**
 - **Dartmouth Flood Observatory**
 - **Data about all major floods in the world since 1985**
 - **The data set is difficult to automatically process**
 - Large number of free text fields
 - Names of provinces are spelled in many different ways
 - **We processed the data in order to extract: country, province (sometimes missing), start and end dates, severity of flood**
 - **1995-2005: 506 floods in Africa**



Mixed Effects Logistic Regression

$$\Pr(y_{i(t)} = 1) = \text{logit}^{-1}\{\beta_0 + \beta_{\text{pop}} \cdot \text{pop}_{i(t)} + \beta_{\text{urban}} \cdot \text{urban}_{i(t)} + \beta_{\text{neighbor}} \cdot \text{neighbor}_{i(t-1)} \\ + \beta_{\text{flood}} \cdot \text{flood}_i + \beta_{\text{GNI}} \cdot \text{GNI}_{j[i](t)} + \alpha_{j[i]\text{country}} + \alpha_{k[i]\text{region}} + \beta_{k[i]\text{rain}} \cdot \text{rain}_{i(t-1)}\}$$

where $\Pr(y_{i(t)} = 1)$ is probability of outbreak in province i starting at month t

$j[i]$ = country of province i

$k[i]$ = region of province i

pop = population size

urban = % urban constitution

neighbor = outbreak in neighboring province

flood = decile of flood risk

GNI= Gross National Income/capita

rain = rainfall

Mixed Effects Logistic Regression: Effect Estimates

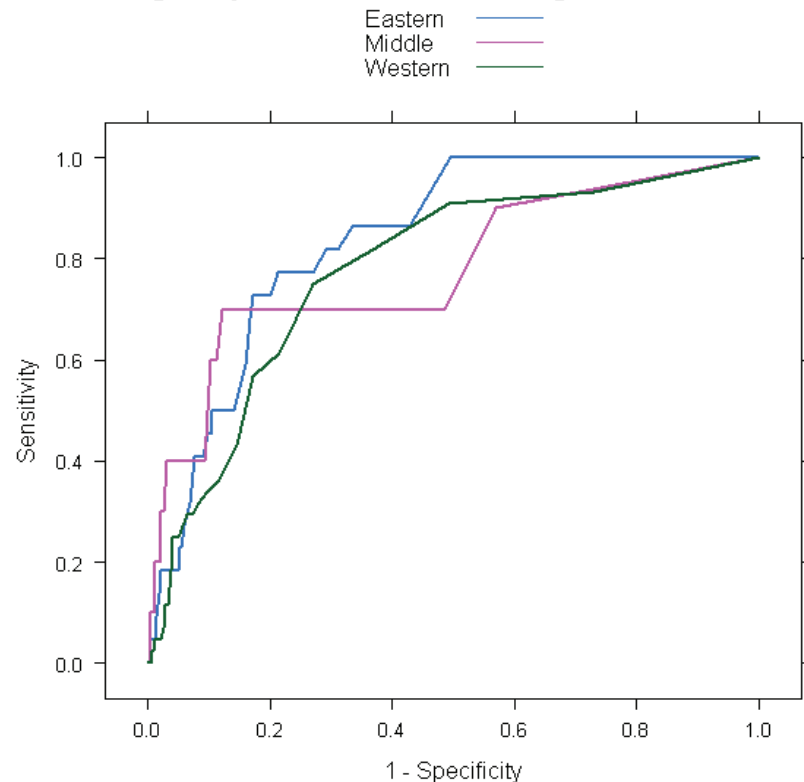
Variable	OR	95% CI
Population size / 1,000,000	1.3	1.2, 1.4
% urban constitution / 10	1.2	1.1, 1.3
Outbreak in neighbor province (1/0)	1.9	1.4, 2.5
Flood risk decile (0=no risk, 1,..., 10)	1.1	1.0, 1.2
Gross National Income per cap / 1,000	0.8	0.6, 1.0
Rain previous month (mm) / 100		
Eastern Africa	1.2	1.0, 1.5
Western Africa	2.0	1.5, 2.7
Middle Africa	0.8	0.5, 1.2
Southern Africa	1.3	0.6, 2.8

Country-specific intercepts not shown.

Logistic Regression Prediction Results for 2005

- Model developed using 1997-2004 data
- Model tested on 2005 data:
 - Desired sensitivity can be obtained by moving on the ROC curve
 - No curve for Southern Africa since no outbreaks in 2005.

ROC: logistic prediction for 2005 using 1997-2004 model



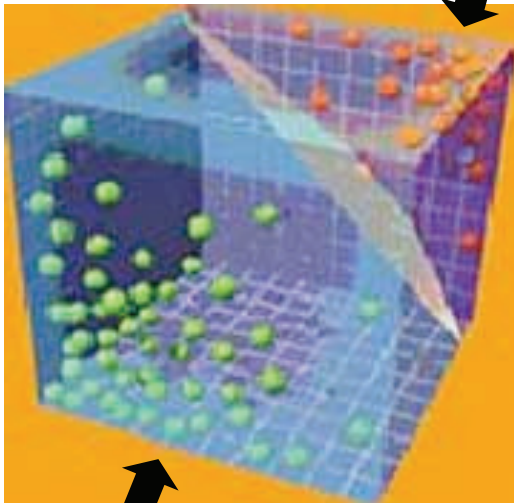
Predicting 2005 Outbreaks using 1997-2004 Model

Region	# Outbreaks in 2005	Sensitivity in 2005 (%)	Specificity in 2005 (%)
Eastern	22	86	62
Western	44	82	65
Central	10	70	83
Southern	0	--	93
All	76	82	70

Machine Learning for Cholera Outbreak Prediction

- Support Vector Machine (SVM):
 - Very powerful classifier
 - Especially well suited for highly dimensional data

Outbreak



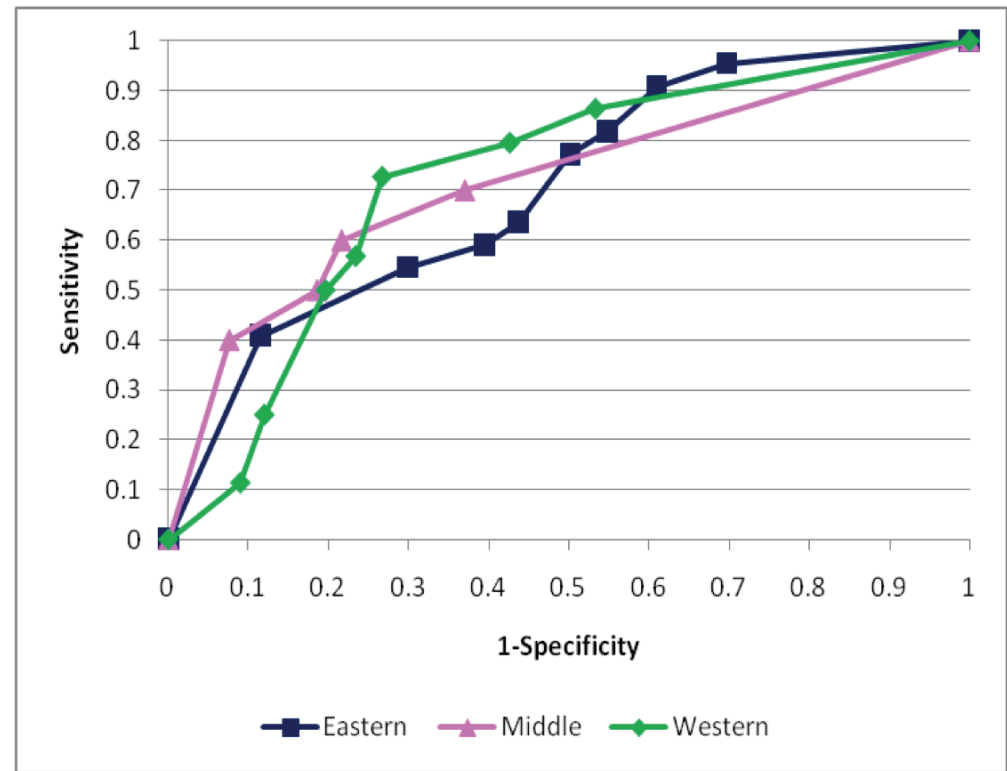
No Outbreak

- Predictors used by SVM:
 - Rainfall (t-1)
 - Flood (t-1, t-2)
 - Flood severity
 - SOI (t-1)
 - Sea Surface Temperature (t-1)
 - Population density
 - Percent urban constitution
 - Flood susceptibility
 - Percent improved drinking water
 - Percent improved sanitation
 - GNI
 - Life expectancy at birth
 - Outbreak in neighboring province (t-1)
- SVM with Radial Basis Function kernel used

SVM Prediction Results for 2005

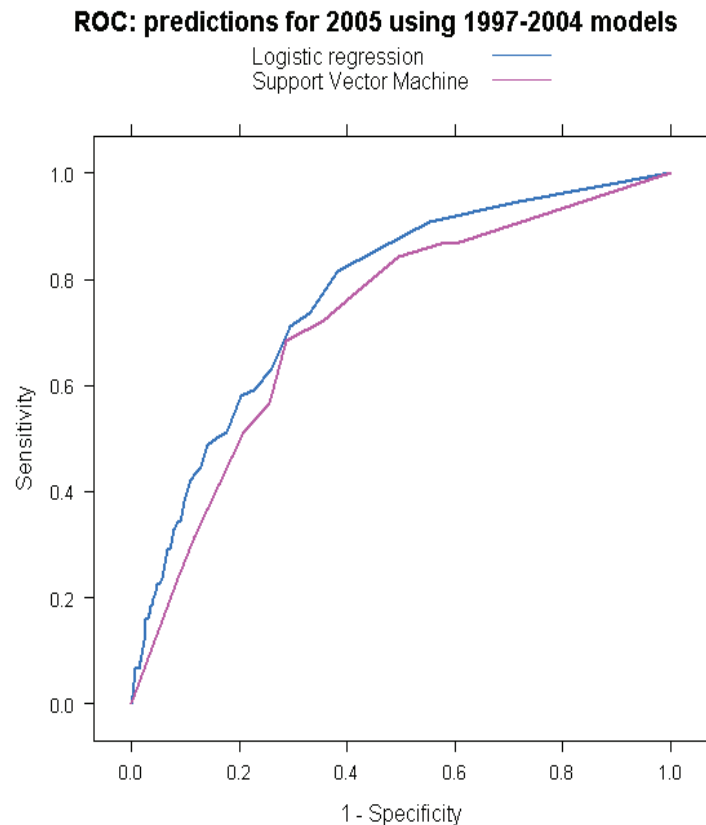
- SVM trained on 1997-2004 data.
- Outbreak predictions for 2005.
- One model developed (covering Eastern, Western, Middle and Southern Africa).
- Desired Sensitivity can be obtained by moving on the ROC curve.
- No curve for Southern Africa since no outbreaks in 2005.

ROC for SVM predictions for 2005



Comparison of Logistic Regression and SVM Results

- Similar predictive accuracy by both methods, with a slight advantage for logistic regression.
- Logistic Regression has certain separate parameters for Eastern, Western, Middle and Southern Africa.
- SVM has only one model covering the whole four regions of the continent. Developing separate SVM models for different African regions, should improve SVM accuracy.



Conclusions

- Large effort in identification of data sources, data acquisition and data preprocessing
- Results:
 - Mixed Effects Logistic Regression
 - Specific parameters per sub-region
 - Support Vector Machines
 - One model only
- In order to obtain higher Sensitivity / Specificity combination we will develop separate models for Eastern, Western, Middle and Southern Africa