Predicting Incidence of West Nile Virus

Biostatistics M.S. Oral Exam

Christopher Aden

May 20, 2015

Introduction

West Nile Virus (WNv):

- Transmitted by infected mosquitoes
- Symptoms: Fever, aching, fatigue, vomiting. Rarely: meningitis, encephalitis, death
- Occurs most frequently during peak mosquito seasons (late Spring–early Fall)

Chicago's West Nile Virus Problem:

- Experienced first case in 2002.
- Established surveillance and management programs to control outbreaks.
- Started monitoring/controlling mosquito populations in 2004.

The Data

Surveillance Data (Traps)

- Data from geo-tagged mosquito traps.
- Presence of WNv, species and counts.
- Location: Latitude-Longitude, Street, Block
- Time: Day, Month, Year

Management Data (Sprays)

- Time (D-M-Y) and Location (Lat-Lon) of all insecticide sprays.

NOAA Atmospheric and Weather Data (Weather)

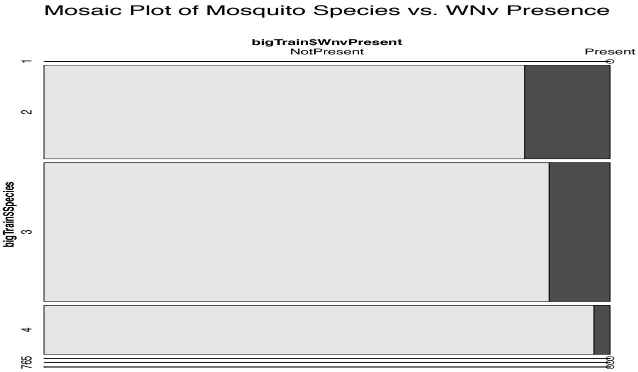
- Collected at Chicago's two airports.
- Time (D-M-Y)
- Temperature (Max, Min, Avg), Atmospheric Pressure
- Precipitation, Dew Point, Wind Speed, etc

Main Questions

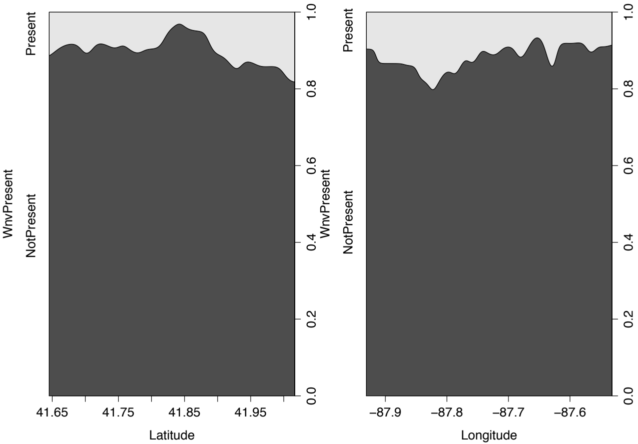
- Controlling for location effects, is spraying an effective method of reducing the incidence of West Nile virus?
- What weather conditions influence WNv incidence and how?
- Can we develop a tool that has high sensitivity to detect West Nile virus, while still maintaining specificity?

Effect of Mosquito Species on WNv Incidence

Figure: 1: *C. Erraticus*, 2: *C. Pipiens*, 3: *C. Pipiens* or *C. Restuans*, 4: *C. Restuans*, 5: *C. Salinarius*, 6: *C. Tarsal*, 7: *C. Territans*

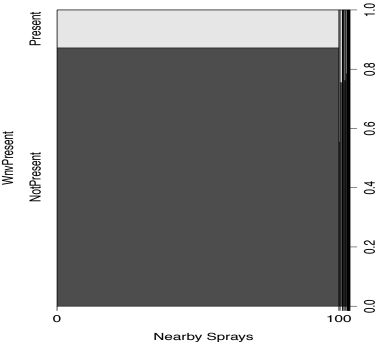


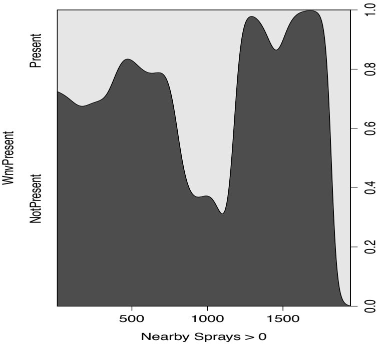
Location Effect



Spray Effect

- Compute Haversine distance for all spray and trap pairs.
  $$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$
- For each trap, count how many sprays conducted within two miles and within two weeks of surveillance.





Confirmatory Analysis

Full Logistic with weather, spray and trap predictors

- Poor model: Too complicated, strong collinearity, no significance
- Ridge regression helps with collinearity, but not enough! Need to "zero out" terms

ElasticNet

- Combination of Lasso and Ridge Regression, minimizes $\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right]$
- Like lasso, shrinks coefficients exactly to zero–produces simple models
- Like ridge, doesn't just take one correlated predictor and leave the rest

Generalized Additive Model (GAM)

- Allows non-linear terms
- *Far* better at modeling location effect.

Generalized Linear Mixed Model (GLMM)

- Mosquitoes in the same trap are correlated! Account for it!

Final Logistic Regression on Held-Out Data ($n = 11,049$)

	Estimate	SE	z-value	p-value	VIF
(Intercept)	-470.5712	44.66	-10.54	$\ll 0.0001$	-
isSpecies1247TRUE	-0.1289	0.06	-2.05	0.0405	1.04
Longitude	-3.7336	0.60	-6.21	$\ll 0.0001$	4.08
Year2013	0.6037	0.11	5.70	$\ll 0.0001$	1.89
Month7	2.3112	0.42	5.51	$\ll 0.0001$	2.07
Month8	4.6864	0.42	11.28	$\ll 0.0001$	
Month9	4.4322	0.42	10.66	$\ll 0.0001$	
Tmax	0.0712	0.01	8.34	$\ll 0.0001$	3.97
HeatDegreeDay	0.0781	0.02	3.19	0.0014	2.95
StationPressure	3.9241	0.70	5.60	$\ll 0.0001$	4.79
ResultSpeed	0.0980	0.02	5.74	$\ll 0.0001$	2.98
anyNearbySpraysTRUE	0.5445	0.11	5.06	$\ll 0.0001$	1.08
Latitude:Longitude	-0.0042	0.005	-0.87	0.3859	4.04

Central Goal of Prediction

Can we develop a tool that has high sensitivity to detect West Nile virus, while still maintaining specificity?

- On the whole data set (without Spray data)?
- For 2011 and 2013 (with Spray and weather data)?
- For 2011 and 2013 (with weather data only)?

Tools

GLM Logistic regression, with no model selection.

GLM-Net GLM followed by optimally-tuned ElasticNets.

RandomForest CART-based ensemble, taking a random subset of data *and* random subset of the *predictors* for each tree. Trees then averaged to form classifier.

Adaptive Boosting (AdaBoost) Ensemble that builds really bad trees ("weak learners"), then tunes subsequent trees to perform better on the samples that the previous trees misclassified.

Gradient Boosting Machines Class of ensembles that includes AdaBoost. Allow for more general loss functions. Optimizes with gradient descent. Faster and more memory-efficient than AdaBoost.

Models have fantastic accuracy, utterly useless

- Unbalanced classes–can achieve great accuracy by always predicting to common class
- Classification Accuracy and AUC are bad metrics for unbalanced responses.

GLM-Net, SprayYears, All vars		
(In %)	True −	True +
Predict −	86.10	12.50
Predict +	0.5	0.9

Solutions

- Need to balance high specificity (easy) with high sensitivity (very hard). Punish classifiers that "cheat".

Balanced Accuracy (BA)

- Mean of sensitivity and specificity
- Weights desire to correctly classify WNv-negative and WNv-positive cases by their incidence.

F_1 Score

- $F_1 = 2 \cdot (\mathrm{PPV} \cdot \mathrm{Sens}) \,/\, (\mathrm{PPV} + \mathrm{Sens})$
- Harmonic mean of Positive Predictive Value ($P(\text{WNv}+ | \text{Guess}+)$) and sensitivity
- Tries to achieve high sensitivity, controlled by true WNv prevalence.

Optimizing for F_1 and BA, instead...

Confusion Matrices for F_1-based optimization (subset of table)

Method	Data	True Neg	False Neg	False Pos	True Pos	F_1
GLM	AllYears, Weather	89.10	10.50	0.10	0.30	0.05
ElasticNet	AllYears, Weather	89.10	10.50	0.10	0.20	0.04
RandomForest	**AllYears, Weather**	88.30	4.00	1.00	6.80	**0.73**
GLM	SprayYears, Weather	86.10	12.50	0.50	0.90	0.12
RandomForest	**SprayYears, Weather**	85.20	3.40	1.40	10.00	**0.81**
AdaBoost	SprayYears, Weather	84.60	8.80	2.00	4.70	0.47
GLM	SprayWeather	86.10	12.40	0.50	1.00	0.13
RandomForest	**SprayWeather**	85.10	3.30	1.50	10.10	**0.81**
AdaBoost	SprayWeather	84.50	8.70	2.10	4.70	0.47

Confusion Matrices for *BA*-based optimization (subset of table)

Method	Data	True Neg	False Neg	False Pos	True Pos	BA
GLM	AllYears, Weather	89.10	10.50	0.10	0.30	0.51
ElasticNet	AllYears, Weather	89.10	10.50	0.10	0.20	0.51
RandomForest	**AllYears, Weather**	88.20	3.90	1.00	6.80	**0.81**
GLM	SprayYears, Weather	86.10	12.50	0.50	0.90	0.53
ElasticNet	SprayYears, Weather	86.00	12.50	0.50	0.90	0.53
RandomForest	**SprayYears, Weather**	85.00	3.20	1.50	10.20	**0.87**
AdaBoost	SprayYears, Weather	84.60	8.80	2.00	4.60	0.66
RandomForest	**SprayWeather**	84.90	3.20	1.70	10.20	**0.87**
GradientBoosting	SprayWeather	84.50	5.00	2.10	8.40	0.80
AdaBoost	SprayWeather	84.50	9.00	2.10	4.40	0.65

Concluding Remarks

Regularization separates the wheat from the chaff

- Found very significant model
- Parsimonious and intuitive to explain
- Previous weather findings reinforced
- Clear location effect

Strong Spraying Effect, Interpretation Unclear

- Need better tools (spatiotemporal analysis)

Accuracy is a bad choice with uneven frequencies

- GLMs are garbage at classification
- F_1 and Balanced Accuracy give good sensitivity, decent specificity
- "Best" classifier is dependent on desired properties (No Free Lunch)

Parallel Programming Is *Vital* to Predictive Modeling

New Statistics department compute server - poisson.ucdavis.edu

Nehad Ismail May 13 (7 days ago)

Dear Statistics Faculty and Students, The department has a new compute server...

Christopher Aden <christopher.b.aden@gmail.com> May 17 (3 days ago)

to Nehad

Nehad,

Do you have a faster one? I ran out of cores on this one ;). Thanks for saving me the AWS EC2 fees!

2. cbaden@poisson:~/MS Exam/src (ssh)

```
USER       PID %CPU %MEM PRI  NI STAT _CPU_Time_ Elapsed_Time  Started CMD
cbaden   44989  132  0.1   0  19 SN     00:04:02        03:03 16:49:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45535  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45536  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45537 99.3  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45538  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45539  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45540  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45541  101  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45542  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45543  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45544  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45545  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45546  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45547  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45548  100  0.1   0  19 RN     00:00:03        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45549 99.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45550 99.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45551 99.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45552 98.3  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45553 99.0  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45554 98.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45555 98.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45556 98.6  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
cbaden   45557 98.3  0.1   0  19 RN     00:00:02        00:03 16:52:17 /usr/local/R-3.2.0/lib64/R/bin/exec/R —no-save
```