Ebola Prevalence Prediction in Africa

Mei Duanmu

November 23, 2014

Objective

- Use Spatial Features to train and predict prevalence
- Spatial Features: Longitude, Latitude
- Label Feature: Prevalence(define by total cases)

Data

- Training Data: Longitude, Latitude, Label for 61 locations across 8 countries in Africa.
- Testing Data: Longitude, Latitude for 159478 locations across 8 countries

Label Method

- Three Prevalence Levels for regions in dataset2:
- High: total cases rank top 25%
- Medium: total cases rank between top and bottom 25%
- Low: total cases rank bottom 25%
- Merge dataset 1 and 2 by region name. Label location prevalence level the same with region it belongs to.

Classification Method

- randomForest package in R
- Accuracy/Error rate: OOB estimate of error rate: 49.18%
- Add features population in training: Error rate decreases to 39.34%

Prediction

- High prevalence location: 8559 (GN:1639, LR:2731, SL:4189)
- Median prevalence location: 66872 (GN:5955,LR:16783,NG:29754,SL:3442,ML:49,CD:117,CI:10772)
- Low prevlanece location: 84047 (GN:3932, LR:79, NG:21108,SL:319,ML:14830 CD:34243, CI:53, SN:9483)

High Prevalence	Medium Prevalence	Low Prevalence
8559	66872	84047

Prevalence Analysis

- Based on the given locations in dataset1:
- High prevalence locations concentrate in country GN, LR and SL.
- Low prevalence locations concentrate in country NG, ML,CD,SN

In the future (Added after presentation)

- Add the features.
- Plot ROC curve for measure classification performance
- Compare classification accuracy with other classification methods(e.g. SVM).
- Map the results by shiny-googleCharts or shiny-leaflet.