West Nile virus forecast model submission form
Email completed form to vbd-predict@cdc.gov

Team name
NCSA health disparities

Team leader

Name	Institution	Email
Dr. Weihao Ge	National Center for Supercomputing Applications at University of Illinois U-C	wge2@illinois.edu

Other team members

Name	Institution	Email
Sparsh Agarwal	University of Illinois, Urbana-Champaign	sparsha2@illinois.edu
Joshua Allen		jallen17@illinois.edu
Vinu Prasad Bhambore		vpb2@illinois.edu
Matas Lauzadis		matasl2@illinois.edu
Shubham Rawlani		rawlani3@illinois.edu
Wayne Wan		guangya2@illinois.edu
Dr. Luidmila S. Mainzer		lmainzer@illinois.edu
William Brown		wmbrown@illinois.edu
Bo Li		lbo@illinois.edu
Rebecca Smith		rlsdvm@illinois.edu
Chris Stone		cstone@illinois.edu
John Uelmen		uelmen@illinois.edu

Model description
Provide a brief summary of the model methods with sufficient detail for another modeler to understand the approach being applied. If multiple models are used, describe each model and how they were combined.

This research work aims at predicting the total number of neuroinvasive diseases (e.g. meningitis, encephalitis, or myelitis) due to prevalent West Nile Virus (WNV) for each county in contiguous United States for upcoming years. This prediction will allow the Department of Public Health of respective areas to proactively react when and where the WNV transmission occurs.

The experiment involves 19 years of historical data ranging from 2000 to 2018 which includes total number of deaths due to neuroinvasive disease, weather data (temperature, precipitation, humidity, etc.) and socio-economic data (income, population, race, etc.) for each of 3109 counties in the United States. To improve the accuracy, the model also includes spatial relation with respect to the total number of cases by considering neighboring county average deaths due to neuroinvasive disease.

Prediction is accomplished by developing a Long Short Term Memory (LSTM) recurrent neural network. LSTM is robust and well-suited to classifying and processing based on time series data. To include the temporal effect, the input parameter involved the past 5 years data to predict the current year. For each prior year used, the features included were weather

(temperature, precipitation, humidity, and gini index of precipitation) and spatial variable (neighboring county average wnv cases). The target variable was a 15-element vector, and each element describes the probability of WNV cases and points to a bin. Each bin was defined by an inclusive minimum and a non-inclusive maximum number of cases of the disease.

A huge amount of collective and collaborative effort was made in developing this prediction model and as a result the model has an accuracy of 73% on unseen data that means the model can predict 73 times correctly out of 100 observations. However, while the current precision is relatively low, it outperforms the trivial prediction which has higher accuracy but misses all high numbered cases. Such a problem is caused by highly unbalanced data that most samples have 0 counts, which is typical in modeling epidemic diseases. Therefore, we would rather sacrifice accuracy so that valid warnings of potential outbreaks could be predicted.

Variables

List each variable used and its temporal relationship to the forecast. If multiple models are used, specify which enter into each model.

1. Total WNV cases
2. Neighboring counties averaged WNV cases
3. Gini index for precipitation
4. Average yearly temperature
5. Average yearly precipitation
6. Average yearly humidity
7. County type (Rural/Urban nominal variable)
8. Outbreak inclusion (Nominal variable indicating years before and after WNV outbreak)

Computational resources

Describe the programming languages and software tools that were used to write and execute the forecasts.

For the scope of this research, there were several resources put into use -

- In general, **Python 3.6** as a programming tool for data preparation, cleaning, manipulating and building model was employed along with its packages - Numpy, Pandas, Scikit-Learn, Keras, Scipy - and **Jupyter Notebook** was used as an IDE.
- **CyberGIS-Jupyter** was used for the preparation of weather data.
- **HAL (Hardware-Accelerated Learning)** Cluster was used for running the model. (This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.)

Publications

Note whether the model was derived from previously published work and, if so, provide references.

Participation agreement

By submitting these forecasts, the team agrees to abide by the project rules and data use agreements.

Team lead name	Date
Dr. Weihao Ge	04/30/2020