Developing winter weather risk indexes: a machine learning approach

David N. Beede, US Dept of Commerce (presenter) and Jeffrey C. Chen, BEA

Finding

We used machine learning (ML) and big weather data to develop experimental winter weather risk indexes for two rare events: school closures (2.2 percent probability on days with snow forecasts) and traffic crash fatalities (0.4 percent). Such indexes may help people translate weather forecasts into information to make better travel decisions and alternative teleworking and child-care arrangements. The algorithms used to predict school closures performed well (goodness-of-fit = 0.87-0.90); less so for fatalities (0.73-0.75). We plan to use state data on all police-reported traffic crashes to develop better travel risk indexes and estimates of weather information's value in school closure decision-making.

Introduction

Winter weather is costly in the United States:

- 1,400 die annually in winter weather-related traffic crashes
- In 2014-15, 78,000 schools closed due to snow/ice conditions, associated with:
  - 42 million child-days of instruction
  - 2.7 million teacher and staff workdays
  - > 30 million work-days for parents

Households and businesses could make better roadway travel choices, and school district superintendents could better assess the cost tradeoffs of school closure decision-making, if there were widely available and reliable tools that could help assess the potential risks associated with severe winter weather. This research is a first step in producing such tools.

Data

Dependent variables

- *Unscheduled K-12 School Closures in the United States, 2011-2016*. The Centers for Disease Control and Prevention (CDC) collected these data (Wong, Shi, & Gao, et al., 2014).
- *Fatality Analysis Reporting System* (FARS). Incident-level data on all roadway motor vehicle traffic crashes resulting in a fatality within 30 days of the accident (National Highway Traffic Safety Administration, 2016).

Independent variables

- *U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) data.* We use public-school district-level estimates of overall, school-age, and school-age in poverty populations (Bureau of the Census, 2018a).
- *National Oceanic and Atmospheric Administration (NOAA) Global Forecast System (GFS).* GFS is a weather forecast model covering the U.S. in 2.5 km to 10+ km grid points. We used the first 24-hour forecast snow and rain accumulations and the 12-hour minimum temperature forecasts available between 00:00 and 03:00 UTC (National Weather Service, 2016).
- Spatial (geocoordinates, state dummies) and temporal (month, day of week, holiday) variables.

Data processing

- Unit of observation is school district-day (e.g., Montgomery County [Maryland] Public School District on January 7, 2015)
- FARS data matched to school districts using geocoordinates
- GFS data averaged by county (using county shape-files)
- Limited data to public school district-wide closures due to snow/ice (95 percent of closures)
- Limited data to district-days with snow forecast during peak winter weather months (Oct–Apr)

Methods

Supervised ML methods (to predict which school-district days have traffic fatalities and/or unscheduled school closures)

- General linear model (GLM). In this case we used logistic regression.
- Generalized additive model: $E[Y|X_1,\ldots,X_k] = \alpha + f_1(X_1)+\ldots+f_k(X_k)$
- Two decision tree ensemble methods (combining many small and weak decision trees in a "wisdom of crowds" approach): AdaBoost and XGBoost

We used 10-fold cross-validation on training data (2012-2014) to assess model fit while avoiding over-fitting the data and tested on 2015 data.

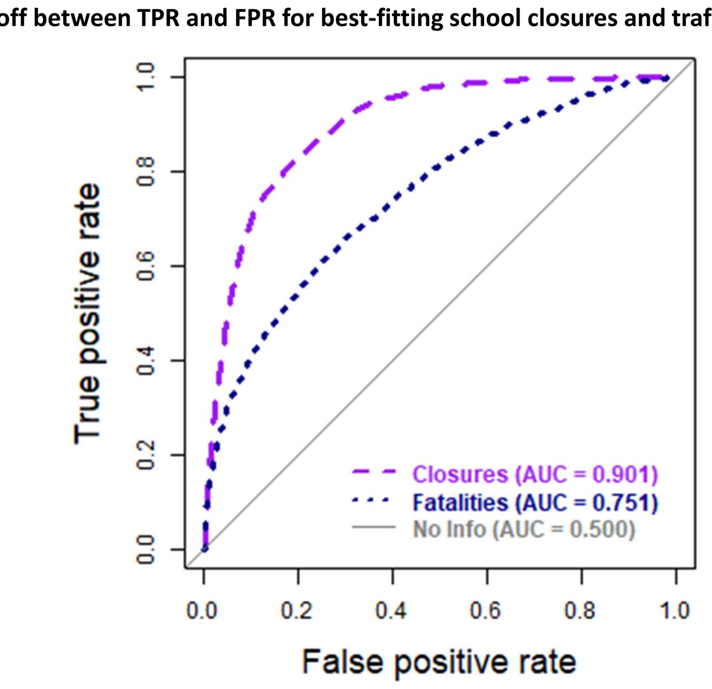
Unsupervised ML methods to create risk indexes:

- We used k-means clustering, which finds the *k* clusters of estimated event probabilities minimizing within-cluster variance
- We set *k* = 5 and labeled the group with the lowest probabilities as one and the highest probability group as five.

Results

Of our four base algorithms, extreme gradient boosting (XGBoost) fit the data best (Chen and Guestrin, 2016). However, the XGBoost had better success predicting school closures than traffic fatalities. We compare goodness-of-fit measures for our traffic fatalities and school closures models based on the area under curves (AUC) showing the tradeoff between the true positive rate and the false positive rate at each probability cut-off between zero and one (see Figure 1). The AUCs for the fatalities and closures models are 0.75 and 0.90, respectively, suggesting that the school closure risk index may be more useful in making decisions than the fatalities index.

Figure 1: Tradeoff between TPR and FPR for best-fitting school closures and traffic fatalities models

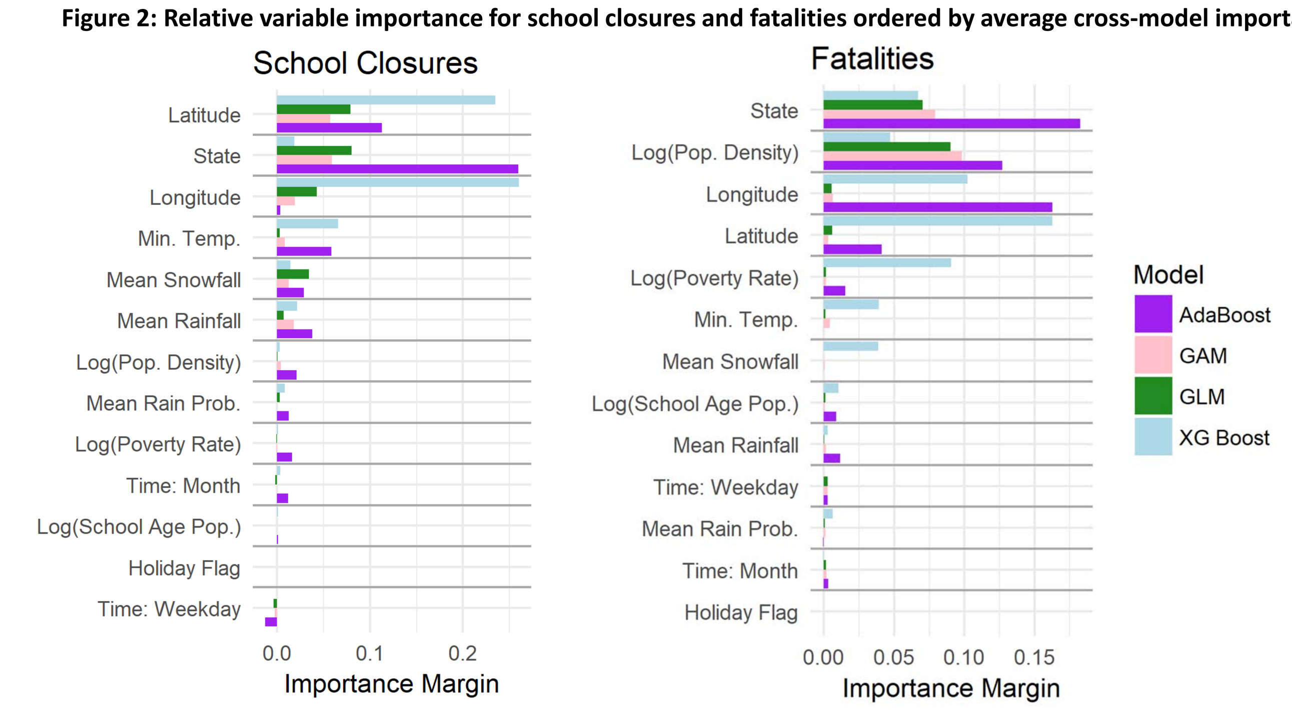


Although our goal is to predict rather than explain events, it is interesting to identify the variables relatively more important in affecting model accuracy. Each model is estimated on the training sample, then the AUC is calculated based on the test sample but omitting each independent variable X_j (denoted by AUC_{-j}) and compared against the AUC of the fully specified model (denoted by AUC_{full}):

$$RVI_j = AUC_{full} - AUC_{-j}$$

The relative variable importance (RVI) statistic measures the contribution of each variable to the overall model performance (summarized in Figure 2).

Figure 2: Relative variable importance for school closures and fatalities ordered by average cross-model importance

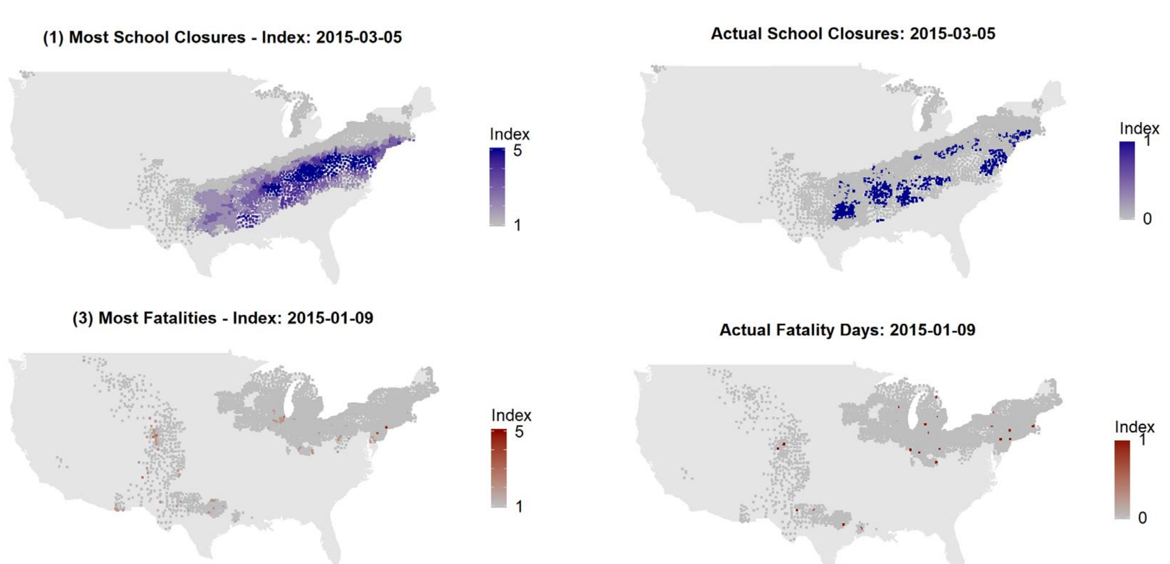


In both the additive and tree ensemble models, geographic variables have the greatest RVI though for different reasons. While *state* is treated as a dummy matrix in all models, *longitude* and *latitude* coordinates are treated as continuous values by additive models (e.g., GLM and GAM), thus capturing linear relationships on a macro-continental scale that are associated with gradual regional differences from north to south and east to west. These differences can be due to differences in long-term weather patterns, socioeconomic characteristics, or both. Tree ensemble models (e.g., XGBoost and AdaBoost) convert coordinates into a series of binary features that effectively create bounding boxes and regions in the data rather than continuous relationships. This helps identify discrete pockets of discontinuous behavior such as how community clusters respond to winter weather. While the RVI measures do not provide the sign of associations, latitude is likely negatively associated with fatalities, consistent with more northern areas tending to have higher long-term average snowfalls and thus more prepared for adverse winter weather.

To put our findings in context, we show in Figure 3 risk index values and actual outcomes for days in 2015 (our test data set) with the most school closures and traffic fatalities. The school closure index performs favorably with respect to TPR, predicting 94.3 percent of actual closures. These findings can be seen in Figure 3 by the noting that most of the areas with actual school closures are within the darkly shaded areas in the corresponding maps on the left. But the predictions' high TPR comes at the expense of low FPR. On the day with the most closures, FPR is 47 percent, which in implementation would translate into many districts receiving a warning even though they ultimately will not close. This can be seen in Figure 3 by the relatively large gaps in actual school closures in the right map compared to the darkly shaded areas in the index map. Clusters tend to indicate similarity in behavior or treatment, thus school closures may be coordinated within regions or perhaps the actual observed weather was far more severe in certain pockets. We plan to develop a formal model of school closure decision-making that accounts for spatial autocorrelation to capture clustered behavior.

Unlike school closures, the rarity of fatalities poses a significant challenge in producing an accurate risk index. Given that less than 20 weather-related traffic-induced fatalities may occur in a given day, TPR is very low (that is, the index is only able to predict 8.8 percent of actual fatalities).

Figure 3: Comparison of Predicted versus Actual Outcomes for School District-Days in 2015 with the Most Actual School Closures and Fatalities.



Notes: Light grey indicates no data or non-snow days; Dark grey indicates Risk Tier = 1 (left panel) or no events (right panel). Darker shades of blue or red indicate higher index values (left panel); blue or red indicate actual school closures or traffic fatalities, respectively (right panel).

Acknowledgements

We thank the Community Interventions for Infection Control Unit (CI-ICU) of the Division of Global Migration and Quarantine (DGMQ) at CDC for providing access to their *Unplanned K-12 School Closures in the United States, 2011-2016* data set. The paper on which this poster is based benefited from comments and suggestions by Sandra Cooke-Hull, Ed Kearns, Christopher Lauer, Sabrina Montes, Regina Powers, Robert Rubinovitz, and Sally Thompson. All errors are the authors' responsibility. Views and conclusions are the authors' and do not necessarily reflect those of any U.S. federal government agency.

References

Bureau of the Census. (2018a). Small area income and poverty estimates (SAIPE) Program. Retrieved September 18, 2017 from https://www.census.gov/programs-surveys/saipe.html.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY: Association for Computing Machinery, 785-794.

National Highway Traffic Safety Administration. (2016). Fatality analysis reporting system (FARS) analytical user's Manual, 1975-2015 (Report No. DOT HS 812 315). Washington, DC:U.S. Department of Transportation.

National Weather Service. (2016). The global forecast system (GFS) - global spectral model (GSM) (GSM version 13.0.2). Retrieved September 28, 2017 from http://www.emc.ncep.noaa.gov/GFS/doc.php.

Wong, K.K., Shi, J., Gao, H., Zheteyeva, Y.A., Lane, K., Copeland, D., Hendricks, J., McMurray, L., Sliger, K., Rainey, J.J., & Uzicanin, A. (2014). Why is school closed today? Unplanned K-12 school closures in the United States, 2011-2013. PLoS ONE, 9, 1-15.

Contact: David Beede (dbeede@doc.gov; 202-482-1226) or Jeff Chen (Jeffrey.Chen@bea.gov; 301-278-9751)

DEPARTMENT OF COMMERCE · UNITED STATES OF AMERICA

bea Bureau of Economic Analysis · U.S. DEPARTMENT OF COMMERCE