Big Data, Simulation and Causation

Timothy H. Savage, PhD[1] and Huy T. Vo, PhD[1,2]
[1] Center for Urban Science and Progress, New York University, New York, New York, United States
[2] Department of Computer Science, City College of New York, New York, New York, United States

Big data typically arises non-experimentally. It is the digital exhaust of human activity. As a result, correlation cannot be interpreted as a causal treatment effect. Using data on taxi rides in New York City, an increasingly popular big dataset, we present an approach that permits causal interpretation. Using a random forest learner and all taxi journeys in the year prior to Hurricane Sandy, we develop Markov transition probabilities that capture the likelihood of a particular drop off location conditional on the pickup location and other features of the journey. The fitted forest is then used to simulate an alternative set of transition probabilities using randomization over the pickup location together with random sampling with replacement of the other features of the journey. These simulated transition probabilities can be compared to observed transition probabilities with differences being equivalent to a treatment effect. Our use of the NYC taxi data serves as a proof of concept. This methodology can easily be extended to a number of different contexts.

Goal: To use a well-understood ML algorithm together with an increasingly popular big data source [1, 2, 3, 4] to explore an important topic and further an interdisciplinary examination of problems confronting machine learning with big data.

Causality and Prediction: There are often more police officers per capita in high crime precincts, which does not imply that increasing the number of police officers per capita increases crime. It has long been understood that to predict the causal impact of some “treatment,” in this example more police officers, researchers must compare the outcome, crime, with the treatment intervention to what would have happened in its absence, which is unobservable [5]. It has been shown that observed differences can be decomposed into an average treatment effect, more police officers, and selection bias, non-random assignment of police officers [6]. Moreover, as Hal Varian of Google notes, “A good predictive model can be better than a randomly-chosen control group, which is usually thought to be the gold standard.” [7]

Alternatives: Randomized control trials (RCT), natural experiments, and techniques such as instrumental variables (IV), regression discontinuity (RD), and inverse propensity score weighting (IPSW).

Data and Methods: All NYC yellow cab journeys in the year prior to Hurricane Sandy (~173 million). Pickup and drop off longitudes and latitudes are geocoded to the Taxi and Limousine Commission map with 28 possible locations. [8] We model the journeys as Markov transitions: the probability of dropping off in Chelsea conditional on features of the journey. To address scale, we use a random forest classifier.

```
   dow  tod  passenger_count  pickup_zone  dropoff_zone  weight
0    4   10                6           27             8       7
1    6    9                4            4             4     271
2    6   22                4           22             6       1
3    6   22                1           27            25      18
4    3    8                4           24            28       1
```

A Sunday Without Sandy: Empirical exploration of treatment effects exclusively approaches learning and prediction from the perspective of evaluating the impact of an intervention: the effect of more police per capita on crime or the effect of an advertising campaign on sales. Our use case instead *removes* an intervention: Hurricane Sandy. Using the fitted random forest, we simulate one billion journeys the Sunday Sandy hit NYC using randomizing over the pickup location and sampling with replacement for the other features. Comparing the simulated Markov transition probabilities to the observed transition probabilities measures the impact of Sandy, we find substantial differences. Moreover, asymptotic theory tells us that there exists a stable distribution of drop off locations, the differences of which may also be compared. We find that, by drop off location, the ~420,000 journeys were impacted by Sandy.

```
change in magnitude
[-169 -1114 -3410 -1571 3324 -162 -3238 -737 -3438 2822 1980 2544 -1680 -2030
5756 -1201 -591 -3353 -69 -1257 1249 100 2030 4388 917 -239 -81 386 -1151]
```

Extensions: Our goal was a proof of concept using a standard ML algorithm and a popular big dataset. We have not addressed issues regarding feature selection or model selection (in this instance, tree pruning).

References:

[1] Savage, T.H. and Vo, H.T., 2013. "Yellow Cabs as Red Corpuscles," IEEE Workshop on Big Data and Smarter Cities.

[2] Doraiswamy, H., Vo, H.T., Silva, C., and Freire, J., 2016. "A GPU-Based Index to Support Interactive Spatio-Temporal Queries over Historical Data," *Proceedings of ICDE 2016*, International Conference on Data Engineering (forthcoming).

[3] Ota, M., Vo, H.T., Freire, J., and Silva, C., 2015. "A Scalable Approach for Data-Driven Taxi Ride-Sharing Simulation," *Proceedings of IEEE BigData 2015*.

[4] Poco, J., Doraiswamy, H., Vo, H.T., Comba, J., Freire, J., and Silva, C., 2015. "Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions," *Computer Graphics Forum* (in proceedings of EuroVis 2015), 34(3): 161-170.

[5] Rubin, D.B., 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66(5): 688-701.

[6] Angrist, J.D. and Piscke, J.S., 2009. *Mostly Harmless Econometrics*, Princeton University.

[7] Varian, H., 2014. "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2): 3-28.

[8] http://www.nyc.gov/html/tlc/downloads/pdf/passenger_info_map.pdf.