Social media data and computational models of mobility: A review for demography [Extended abstract]

Momin M. Malik & Jürgen Pfeffer

18 March 2016

Abstract

From our research with social media data, we have several key things about which to warn demographers exploring the possibilities of social media data: unreliable data access, idiosyncratic behaviors, platform effects, and unrepresentative samples. Each of these contribute to making social media data potentially useless for studying anything other than social media. But there are still opportunities in the unprecedented scale and granularity of social media's spatial and temporal data: it gives us a 'test bed' to ask, if we were to have perfect information about movement, how would we characterize migration? While we do not aim to present a definitive answer, we review several popular computational and statistical models in widespread use that have potential application in demography. We also review some important statistical issues, including the difference between explanation (or information) and prediction, an increasingly important but (outside of statistics) underappreciated distinction in modeling.

Introduction: Social Media Data

From early hopes and promises of social media data revolutionizing social science (Lazer et al., 2009; Golder and Macy, 2012), researchers have come to appreciate many complications (Tufekci, 2014; Ruths and Pfeffer, 2014). We have focused our own work on understanding the nature of social media data, and the challenges involved in its use. First, in data collection, Morstatter et al. (2013) demonstrates that the data made available for free through Twitter's Streaming API returns a nonrandom sample of data matching submitted queries. This distorts relative frequencies, which can relate to key research questions. Second, behavior on social media platforms is governed by emergent norms that may be peculiar to those platforms (Malik and Pfeffer, 2016b). Third, social media platforms are not neutral public utilities, but private companies with their own interests; they design platform features to constrain guide users towards desirable behaviors, often successfully (Malik and Pfeffer, 2016a). Lastly, and of particular importance to demography, is that social media users are not representative of larger populations (Malik et al., 2015). These concerns add up to suggest that it is risky to use social media data for studying anything other than social media.

In our study on the geographic representativeness of geotagged tweets (Malik et al., 2015), we compared demographic information from the US Census at the level of block groups to the number of geotagged tweets placed in those block groups over a period of time. Unsurprisingly, we demonstrated that geotag tweet users are unevenly distributed across demographic characteristics of block groups. But in the process, one obstacle we ran into was that of uniquely locating users who had tweets in more than one block group. We used a majority rule (they were located wherever the majority of their geotagged tweets were), but we found cases of this obviously being ineffective, for example in how many people were placed in the block groups of international airports. On this point especially, we received feedback from demographers that, instead of forcing social media data into the limitations of Census data (requiring people to be uniquely identified in one location), we could use social media data to do things not possible with Census data. One of the key opportunities is in using this 'mobility' data to address migration.

Models of Mobility

Out of the three parts of the demographic balancing equation, births and deaths are relatively unambiguous, as they are just points in space and time. Migration, however, is more complicated. Migration is reduced to immigration and emigration, but these are simplifications that cannot account for phenomena like multiple domiciles, migrant labor, commuting, and other ways in which people are identified with more than one current place.

Important to note is that there are a wide class of models of mobility (Bai and Helmy, 2006) that are *simulation* models rather than data models. Simulation models generate data for various purposes, whereas in statistical models, as Fisher put it in 1922 (Fisher, 1922), "briefly, and in its most concrete form, the object of statistical methods is the reduction of data." The simplest example of a simulation model is the 'random walk,' where a path is simulated by moving to a certain distance away from the current position in a random direction. The purpose of simulation models is to generate 'realistic' behavior, assessed in various ways, often for the purpose of testing systems (e.g., seeing how a system for data management will handle incoming data before actually deploying it). Simulation modeling can also be used for conceptual exploration; in such cases, the argument is that if a simple simulation setup can generate realistic behavior (Gilbert and Troitzsch, 2005), then there is a case that the simulation terms represent the underlying causal process of the real-world system, but we will not cover such cases. We are interested only in models we can use to describe and analyze data.

The ideal complete information would be to infer a person's path through space from individual geographic points (i.e., samples of that path). In signal processing and computer science, this is generally a 'solved' problem. Across multiple tracking applications, for example mouse trackpads and GPS, paths are created from point samples using an algorithm called the *Kalman filter* (Faragher, 2012), which is also celebrated as being used in the first manned spaceflight. It works by taking the 'state' of a system at a given time, where the state includes the position, speed, and direction, and making a prediction for the state at the next point in time (extrapolating the new position and velocity from the current position and the velocity). Once the new state is observed (or a previously made observation is fed into the algorithm), the algorithm corrects for however much it was off, and then uses the corrected values to make future predictions. By making predictions between actual observations, the algorithm fills in values to create a smooth path of motion.

However, while this is useful for visualization and for tracking purposes, it is likely not very useful for demography. What is needed is a reduction of data, and preferably, an abstraction that can capture some fundamental features of movement across a population. For this, one relevant model is that of a *transition matrix*. If there are n states, where here a 'state' would be for example a geographic location, the transition matrix is an $n \times n$ table where the entry in row i and column j represents the probability of an observation moving from state i to state j. When given data, we can use the fraction of times observations moved from i to j instead of from i to some other state as a stand-in for probability. By looking at which entries of the table are large, we can characterize motion patterns. The drawback is that a finite number of places need to be prespecified, but this is an example of a data reduction technique that may have use for demographic representation. Usefully, transition matrices can be multiplied with one another, with the resulting matrix also being a valid transition matrices; this can be used to calculate, for example, the predicted location after several movements.

Here, it is important to note a distinction in statistical models: there are *predictive* models and *explanatory* models (Breiman, 2001; Shmueli, 2010). In statistics 'predicted values' is a technical term that is synonymous with 'fitted values,' and represents what we would guess if we were to make a prediction, but is not itself a 'prediction' about unrealized outcomes. Counterintuitively, uninterpretable black-box models can predict better (i.e., fit very well to the data) than models that try to capture the underlying causal processes Shmueli (2010). And, even more counterintuitively, there are conditions under which models that are wrong actually predict better than the model used to generate the data (Wu et al., 2007). In machine learning, the objective is usually to find a model that fits extremely well; but this means that, for social scientific purposes, machine learning models are seldom useful because there is no guarantee they have captured anything about the underlying process. Both the Kalman filter and transition matrices are examples of predictive models, in that they are used to generate predictions rather than to understand via modeling the underlying processes. Depending on the application, however, prediction alone is good enough; smoothing out a path of travel is one example, where there are no causal processes of particular interest.

Conclusion

While social media data and computational models have much to offer demography, there are also subtle methodological and statistical issues that will need to be addressed carefully. Conversations and collaborations between computational modelers and demographers will be critical for bringing out the possibilities of social media data for the study and understanding of the structure of human populations.

References

Bai, F. and Helmy, A. (2006). A survey of mobility models in wireless adhoc networks. In Safwat, A., editor, *Wireless Ad-Hoc and Sensor Networks*, pages 1–30. Springer.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Faragher, R. (2012). Understanding the basis of the Kalman Filter via a simple and intuitive derivation. *IEEE Signal Processing Magazine*, 29(5):128–132.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.

Gilbert, N. and Troitzsch, K. G. (2005). *Simulation for the Social Scientist*. Open University Press.

Golder, S. and Macy, M. (2012). Social science with social media. *ASA footnotes*, 40(1):7.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–723.

Malik, M. M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*, ICWSM-15 SPSM, pages 18–27.

Malik, M. M. and Pfeffer, J. (2016a). Identifying platform effects in social media data. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, ICWSM-16.

Malik, M. M. and Pfeffer, J. (2016b). A macroscopic analysis of new content in Twitter. *Digital Journalism*.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM-13, pages 400–408.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM-14, pages 505–514.

Wu, S., Harris, T. J., and Mcauley, K. B. (2007). The use of simplified or misspecified models: Linear case. *The Canadian Journal of Chemical Engineering*, 85(4):386–398.