Effects without a Cause: The Search for Demographic Anomalies

Mathew E. Hauer *
Department of Sociology, Florida State University
and
Stephanie A. Bohon
Department of Sociology, University of Tennessee - Knoxville

January 10, 2020

Abstract

The proliferation of data, modern computing advances, and powerful statistical algorithms unlock the potential search for hidden, previously un- or understudied demographic anomalies. Here, we demonstrate how demographers can use causal inference techniques to identify anamolies by investigating US state-level fertility and mortality time series since 1999 to uncover the hidden baby booms/busts and mortality plagues?/non-plauges?. We find 38 states exhibited at least one mortality anomaly, totalling more than 318k anomalous deaths and an additional 6 states exhibited at least one mortality non-plague? totalling 164k protective deaths. 12 states exhibited baby busts, totalling more than 240k missing births and an additional 11 states exhibited baby booms totalling more than 134k additional births. These results suggest the widespread detection of demographic anomalies. Our analysis does not examine the *causes* of these anomalies and our results point to important further research on the causes of anomalous demographic behavior.

Keywords: mortality, fertility, causal inference

*Thanks y'all!

1 Introduction

The proliferation of data, advances in high performance ("super") computing, and the development of powerful statistical algorithms mark the era of "Big Data" or data science (Van Der Aalst 2016, Zikopoulos & Eaton 2011), with the potential for researchers to find hidden or understudied social phenomena (Bohon 2018). While the availability of Big Data and high performance computing allows novel exploration of data through causal inference (Bohon 2018, Brodersen et al. 2014, Shiffrin 2016), relatively few studies utilize causal inference techniques in the study of demographic phenomena. However, understanding social phenomena using these advances reveals important insights into society (Angrist et al. 1989, Mas & Moretti 2009) and allows us to better monitor population trends (Nobles & Seltzer 2019, Torche & Shwed 2015).

Causal inference, in its most common usage, is an attempt to uncover the underlying mechanism that results in changes in a phenomenon and have a long history approaches in social sciences (Grimmer 2015). Often, the identification of a casual mechanism requires either a randomized control trial (RCT) or expert knowledge applied to a natural experiment. For population-level analysis, expensive RCTs usually preclude their widespread adoption (West et al. 2008). Using natural experiments provide less certainty about causation and fewer opportunities to replicate work but they provide important and widespread insights such as the determinants of migration, however, provides important and more widespread insights on the determinants of migration after Hurricane Katrina in 2006 (Fussell et al. 2014, Hori et al. 2009), mini baby booms after electrical blackouts (Fetzer et al. 2018), and the highly publicized estimates of excess mortality after Hurricane Maria in Puerto Rico in 2016 (Kishore et al. 2018, Santos-Lozada & Howard 2018). Both RCTs and natural experiments follow traditional, hypothesis testing, inductive scientific paradigms where a question is first posed and scientists generate or find data to answer the question. Such approaches, while extensively utilized and time-tested, are likely to miss important phenomena that might go unnoticed.

Both RCTs and natural experiments follow traditional hypothesis testing–inductive scientific paradigms where a question is first posed and scientists generate or find data to answer the question. Such approaches, while extensively utilized and time-tested, are likely

to miss important phenomena that might go unnoticed. Data scientists are rethinking these approaches and have re-envisioned causal inference as a machine learning technique that allows big data researchers to move from correlations that have been established with good accuracy to making strong conclusions about the underlying mechanisms that produce these correlations (Pearl & Mackenzie 2018).

Abductive modeling is the movement from the inductive to the deductive, and sometimes back and forth, to reach conclusions (Bryant & Raja 2014) or "inferring cause from effect" (Crowder & Carbone 2017). In situations with large data sets, an abductive approach is far superior to deductive hypothesis testing, as p-values with an extremely large number of cases are far from revealing (Head et al. 2015, Nuzzo 2014) and we do not want to make purely inductive inferences from data of questionable generalizability (Ruggles 2014). Owing to the long history of big datasets in demographic research (one of the original "big data" resources (Ruggles 2014)), the rich demographic data available in the United States make the potential revelation of interesting and important demographic phenomena not only possible, but extremely plausible – even if identification of the phenomena occurs without identifying the underlying cause. In other sciences, the identification of effects without causes has led to new theories of galactic migration of planets and other breakthroughs (Gomes et al. 2005).

In this paper, we use modern statistical outlier detection algorithms (Chen & Liu 1993) on nearly twenty years of mortality and fertility data at the US state-level to identify anomalous demographic behavior. In essence, we identify *effects without knowledge of the cause*. We ask two questions regarding demographic anomalies: What are the hidden baby booms/busts and mortality spikes/dips in the United States over the last twenty years? We do not necessarily know the causes of these anomalies but identifying them allows scientists with more detailed knowledge of local population dynamics, state-level policy making, or macro-economics to explain these phenomena post-hoc. From explanations of phenomena that may have previously gone unnoticed, demographers may be able to better forecast populations and provide policy solutions for impending problems.

2 Materials and Methods

2.1 Method

We use a statistical time series outlier detection algorithm (Chen & Liu 1993), implemented in the R programming language (R Core Team 2019) via the tsoutliers package (López-de-Lacalle 2019). This algorithm iteratively uses ARIMA models to 1) produce a counterfactual time series to initially detect an outlier or anomaly, and 2) refit the ARIMA with the outliers removed. Here we briefly summarize and describe the method.

Often, the behavior of a time series can be described and summarized in ARIMA models. If a series of values, y_t^*, is subject to m interventions or outliers at time points $t_1, t_2, , t_m$, then y_t^* can be defined as

$$y_t^* = \sum_{j=1}^{m} \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\phi(B)\alpha(B)}\alpha_t$$

Where $I_t(t_j)$ is an indicator variable with a value of 1 at observation t_j and where the jth outlier arises, $\phi(B)$ is an autoregressive polynomial with all roots outside the unit circle, $\theta(B)$ is a moving average polynomial with all roots outside the unit circle, and $\alpha(B)$ is an autoregressive polynomial with all roots on the unit circle.

We examine three types of outliers at time point t_m: 1) additive outliers (AO), defined as $L_j(B) = 1$; 2) level shift outliers (LS), defined as $L_j(B) = 1/(1-B)$; and 2) temporary change outliers (TC), defined as $L_j(B) = 1/(1-\delta B)$.

Colloquially, additive outliers arise when a single event causes the time series to unexpectedly increase/decrease for a single time period; level shift outliers arise when an event causes the time series to unexpectedly increase/decrease for multiple time periods; and temporary change outliers arise when an event causes the time series to unexpectedly increase/decrease with lingering effects that decay over multiple time periods.

An outlier is detected using a regression equation

$$\pi(B) y_t^* \equiv \hat{e} = \sum_{j=1}^{m} \omega_j \pi(B) L_j(B) I_t(t_j) + \alpha_t$$

where $\pi(B) = \sum_{i=o}^{inf} \pi_i B^i$. The identification of outliers then involves a three step

process to (1) identify all potential outliers, t_j and $L_j(B)$, (2) joint estimates of model parameters and outlier effects are computed to identify potentially spurious outliers, and (3) the outliers and effects are re-estimated without spurious outliers. Of importance here is step 2, which relies on some critical value above which an outlier at time point m is considered spurious. Based on Chien and Liu's (1993) recommendation, we set a critical value of 3.5 which generally minimizes the possibility of Type I errors or false positive outliers. Outliers are reported using a simple t-statistic.

2.2 Data

We search for demographic anomalies using the Center for Disease Control and Prevention's online WONDER monthly fertility (2003-2017) and mortality databases (1999-2016) for all fifty states and the District of Columbia (United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC) 2018, Centers for Disease Control and Prevention, National Center for Health Statistics. 2019). These data sets contain every birth and death record in the United States over the time periods of interest, representing the universe of both mortality and fertility data in the US. These data are considered the "gold standard" of data collections (Mahapatra et al. 2007) and have been considered "complete" since 1968 (Hetzel 2016). We search over each state equivalent's (n=51) mortality (n=228) and fertility (n=180) monthly time series for a total of 20,808 state-months of data.

3 Results

We detect numerous anomalous mortality and fertility events at the US state-level since 1999. A full listing of these anomalies can be found in the **Supplementary Materials**. We begin with a summary of the anomalies we detect and then we highlight highly significant mortality and fertility anomalies across all three types of outliers (Additive Outliers, Level Shift Outliers, and Temporary Change Outliers) with plausible explanations.

3.1 Overall Anomalies

3.2 Mortality

We begin with mortality for New York State (Figure 1). We identify seven anomalies in the mortality time series for New York, all with t-statistics in excess of 3.91, making these anomalies highly significant. The algorithm correctly identifies September 2001 as an additive outlier (2001:09 t=6.40) where mortality in that month was 1,628 higher than anticipated. This mortality event is likely caused by the September 11 tragedy and the detection of this mortality event provides confidence in our detection of other anomalies.

```
## numeric(0)
```

In **Figure 1**, notice the strong level shift (LS) that occurs in February 2004 (2004:02 t=5.59) which prevented almost 870 deaths per month. This shift totals more than 144,000 averted deaths compared to the counter-factual time series and is the single largest mortality protective anomaly among all states. This translates to 4.5% fewer deaths than expected over the time period. What is driving this mortality protection? What policies did NY put into place that might have contributed to this considerable mortality reduction? What environmental conditions may have changed? These are the kinds of questions that arise from our analyses.

Contrast the mortality protection in New York with the enhanced mortality in New Hampshire (**Figure 2**). Here we detect two highly significant level shifts (LS) in the monthly mortality data, first in April 2010 and again in November 2014 (2010:04 t=3.51; 2014:11 t=6.68). These anomalies suggest New Hampshire experienced 9,700 more deaths (+14% more than expected) in a seven-year period beginning in early 2010. This is the single largest percentage mortality increase/decrease we detected among all states. Not coincidentally, NH has the second highest opioid-related mortality in the US (Beetham et al. 2019) and it is reasonable that we detect this epidemic in our results.

```
## numeric(0)
```

In the case of New York and New Hampshire, some of the demographic anomalies have plausible explanations. It seems likely that New York's AO anomaly in September 2001 is

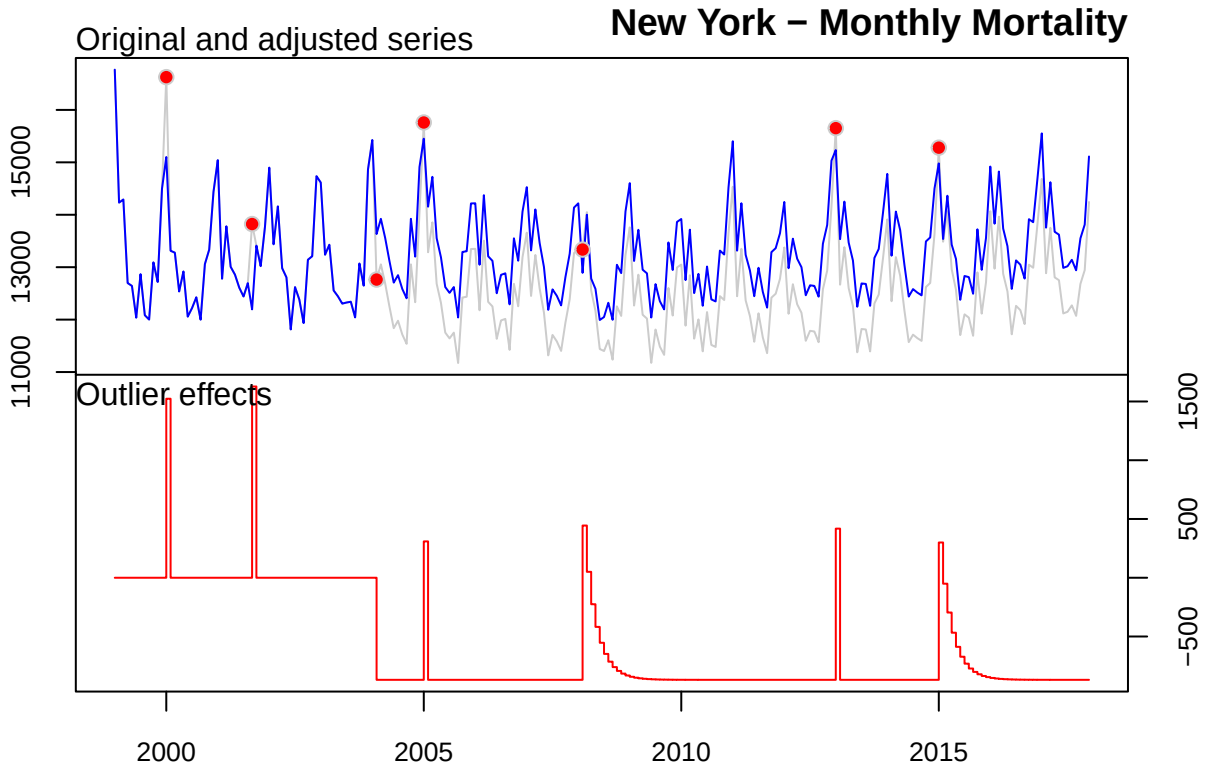

Figure 1: **Anomaly Detection for New York state mortality, 1999-2016.** The top panel contains the original time series (light gray) and the corrected, counter-factual time series in the absence of anomalies (blue). The red dots on the top panel correspond to the onset of detected anomalies. The bottom panel contains the magnitude and type of the outlier. We detect all three types of outliers in New York mortality data during period 1999-2016. We detect additive outliers (AO) in January 2000, September 2001, January 2005, and January 2013; temporary change (TC) outliers in February 2007 and January 2015; and a level shift (LS) starting in February 2004. Outliers range in significance from a low t-statistic of 3.91 in the TC 2015:01 to a high t-statistic of 6.40 in the AO outlier in 2001:09.

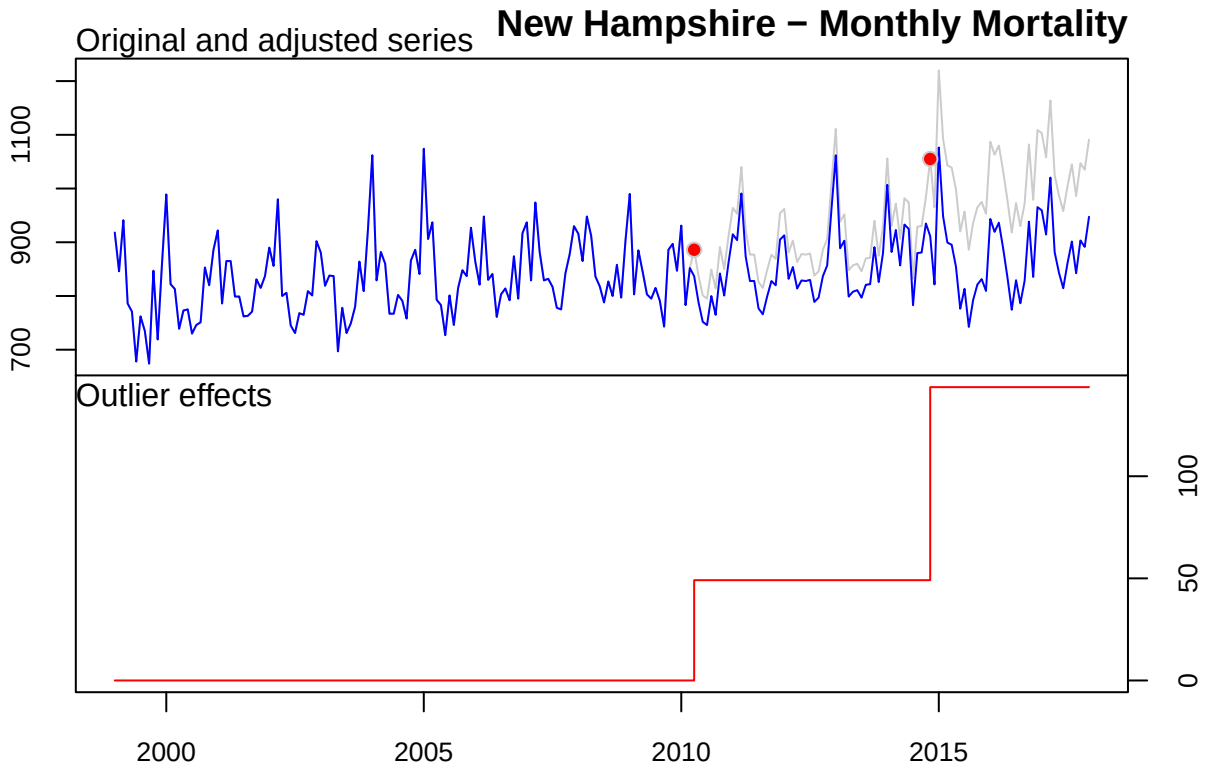

Figure 2: **Anomaly Detection for New Hampshire state mortality, 1999-2016.** Here we detect two outliers, both level shifts in outliers (LS) in April 2010 and again in November 2014. These anomalies suggest New Hampshire experienced approximately 9,700 more deaths than expected since 2010 or 14% more deaths in the state over just seven years.

caused by the 9/11 tragedy and the rise in New Hampshire's mortality starting in 2010 could be linked to the opioid epidemic. However, we detect numerous seemingly unexplainable demographic anomalies in other states. Figure 3 shows two such unexplained anomalies.

References

Angrist, J. D. et al. (1989), *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, number 251, Industrial Relations Section, Princeton University.

Beetham, T., Saloner, B., Wakeman, S. E., Gaye, M. & Barnett, M. L. (2019), 'Access to office-based buprenorphine treatment in areas with high rates of opioid-related mortality: An audit study', *Annals of internal medicine* **171**(1), 1–9.

Bohon, S. A. (2018), 'Demography in the big data revolution: Changing the culture to forge new frontiers', *Population Research and Policy Review* **37**(3), 323–341.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N. & Scott, S. L. (2014), 'Inferring causal impact using Bayesian structural time-series models', *Annals of Applied Statistics* **9**, 247–274.

Bryant, A. & Raja, U. (2014), 'In the realm of big data', *First Monday* **19**(2).

Centers for Disease Control and Prevention, National Center for Health Statistics. (2019), 'Underlying Cause of Death 1999-2017 on CDC WONDER Online Database, released December, 2018. Data are from the Multiple Cause of Death Files, 1999-2017, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program', pp. Accessed at http://wonder.cdc.gov/ucd–icd10.html (data downloaded on 17 August 2019).

Chen, C. & Liu, L.-M. (1993), 'Joint estimation of model parameters and outlier effects in time series', *Journal of the American Statistical Association* **88**(421), 284–297.

Crowder, J. A. & Carbone, J. N. (2017), Cognitive architectures for prognostic health management, *in* S. Ekwaro-Osire, A. C. Gonçalves & F. M. Alemayehu, eds, 'Probabilistic Prognostics and Health Management of Energy Systems', Springer International Publishing, Cham, pp. 91–107.

Fetzer, T., Pardo, O. & Shanghavi, A. (2018), 'More than an urban legend: The short- and long-run effects of unplanned fertility shocks', *Journal of Population Economics* **31**(4), 1125–1176.

Fussell, E., Curtis, K. J. & DeWaard, J. (2014), 'Recovery migration to the City of New Orleans after Hurricane Katrina: A migration systems approach', *Population and Environment* **35**(3), 305–322.

Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. (2005), 'Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets', *Nature* **435**(7041), 466–469.

Grimmer, J. (2015), 'We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together', *PS: Political Science & Politics* **48**(1), 80–83.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. (2015), 'The extent and consequences of p-hacking in science', *PLoS biology* **13**(3), e1002106.

Hetzel, A. (2016), 'US vital statistics system: Major activities and developments, 1950–95.(DHHS publication no.(PHS) 97-1003). hyattsville, MD: National center for health statistics; 1997'.

Hori, M., Schafer, M. J. & Bowman, D. J. (2009), 'Displacement Dynamics in Southern Louisiana After Hurricanes Katrina and Rita', *Population Research and Policy Review* **28**(1), 45–65.

Kishore, N., Marqués, D., Mahmud, A., Kiang, M. V., Rodriguez, I., Fuller, A., Ebner, P., Sorensen, C., Racy, F., Lemery, J. et al. (2018), 'Mortality in puerto rico after hurricane maria', *New England journal of medicine* **379**(2), 162–170.

López-de-Lacalle, J. (2019), *Tsoutliers: Detection of Outliers in Time Series*. R package version 0.6-8.

Mahapatra, P., Shibuya, K., Lopez, A. D., Coullare, F., Notzon, F. C., Rao, C., Szreter, S. et al. (2007), 'Civil registration systems and vital statistics: Successes and missed opportunities', *The Lancet* **370**(9599), 1653–1663.

Mas, A. & Moretti, E. (2009), 'Peers at work', *American Economic Review* **99**(1), 112–45.

Nobles, J. & Seltzer, N. (2019), Finding and Characterizing the Displaced: A method using administrative data, *in* 'Population Association of America Conference, Austin, Texas'.

Nuzzo, R. (2014), 'Scientific method: Statistical errors', *Nature News* **506**(7487), 150.

Pearl, J. & Mackenzie, D. (2018), *The book of why: the new science of cause and effect*, Basic Books.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ruggles, S. (2014), 'Big microdata for population research', *Demography* **51**(1), 287–297.

Santos-Lozada, A. R. & Howard, J. T. (2018), 'Use of death counts from vital statistics to calculate excess deaths in Puerto Rico following Hurricane Maria', *Jama* **320**(14), 1491–1493.

Shiffrin, R. M. (2016), 'Drawing causal inference from big data', *Proceedings of the National Academy of Sciences* **113**(27), 7308–7309.

Torche, F. & Shwed, U. (2015), 'The hidden costs of war: Exposure to armed conflict and birth outcomes', *Sociological Science* **2**, 558–581.

United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC) (2018), 'National center for health statistics (NCHS), division of vital statistics, natality public-use data 2007-2017, on CDC WONDER online database, march 2009', pp. Accessed at http://wonder.cdc.gov/natality–current.html (data downloaded on 19 September 2019).

Van Der Aalst, W. (2016), Data science in action, *in* 'Process Mining', Springer, pp. 3–23.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M. & Mullen, P. D. (2008), 'Alternatives to the Randomized Controlled Trial', *American Journal of Public Health* **98**(8), 1359–1366.

Zikopoulos, P. & Eaton, C. (2011), *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media.