Causal Inference for Mobile Health data: demonstration relating SeeClickFix and crime

Daniel J. Spakowicz[1], Carolyn J. Presley[2,3], Dowin Boatright[4], Ann Greene[5], Marjorie Rosenthal[5,6], Andrew V. Papachristos[7], Mark Gerstein[1,8,9*]

1 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
2 Department of Internal Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA
3 The James Cancer Center, Columbus, OH, USA
4 Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, USA
5 Robert Wood Johnson Foundation Clinical Scholars Program, Yale University School of Medicine, New Haven, CT, USA
6 Department of Pediatrics, Yale University School of Medicine, New Haven, CT, USA
7 Department of Sociology, Yale University, New Haven, CT, USA
8 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
9 Department of Computer Science, Yale University, New Haven, CT, USA

* mark@gersteinlab.org

1-1

2

Abstract

Mobile health data show great promise in supporting personalized medicine. Here we demonstrate the adaptation of an analysis method using Bayesian structural time series for causal inference in these time series data. Lacking sufficient mobile health data, we show that crime data shares many of the features of mobile health data, and is a useful surrogate for causal inference analysis. We demonstrate the process by which an intervention can be evaluated for a causal effect in mobile health data using the introduction of SeeClickFix, a non-violent community reporting tool, as the intervention in crime in New Haven, CT.

1-3

1-2

BETTER

γ

Author summary

Mobile health data are becoming very popular but there are relatively few analytical tools that have been demonstrated to be useful for their analysis. We show how a particular type of analysis, the Bayesian structural time series, is useful in this context. Since there aren't a lot of mobile health data publicly available we demonstrate the analysis using a surrogate that shares many of the mobile health data characteristics: crime data. We study the effect of an intervention, the creation and adoption of a community reporting tool called SeeClickFix, on the crime rate in neighborhoods of New Haven, CT.

Introduction

Mobile health data have great promise for advancing personalized medical care in the coming decades. As more of these data are brought into practice, methods are needed that can infer causal relationships in the context of these within the specific challenges many fluctuating variables . This is particularly true when not all relevant, or potentially relevant, variables are being tracked, as is often the case with longitudinal cohort studies.

For example, a mobile health data source growing in popularity is from activity trackers (e.g. FitBit), which are often correlated to biomarkers of health such as weight, blood pressure, or fasting glucose [1]. Less often captured but important for causal inference in this system are variables such as the number of calories consumed or the macronutrient composition [1]. In addition, each individual in a study is likely to have a distinct mean and variation that is correlated with itself over time, and the extent to which this information is known may vary, e.g. the amount of time that individuals have been tracked or the frequency of data collected.

2-1

These analytical challenges have been tackled in other contexts. Difference-in-difference type approaches have been used to look for causal effects while controlling for latent variables, though these have not been fully adapted to time series typical of mobile health data. The forecasting of time series has a long history with ARIMA models, though these do not handle uncertainly as explictly as may be needed for mobile health data. A flexible method that does not have these shortcommings is the Bayesian structural time series model.

This modeling approach has been used in a variety of fields but not widely adopted for mobile health. For example, Brodersen et al demonstrated its utility in monitoring the effects of marketing campaigns on web page visits [2] and many other fields, including the effects of app releases on sales of smartphones [2], and of a hand hygiene campaign on the rate of hospital-onset bacteremias [1]. Here we demonstrate the utility of these methods using time series data that share many of the characteristics of mobile health data, crime. To simulate an intervention similar to those explored in mobile health experiments we looked at the effect of a social intervention, namely the invention and adoption of a non-violent issue reporting tool called SeeClickFix. This is a smartphone and web application developed in New Haven, Connecticut, where users are able to report issues in their communities such as grafitti or potholes. This publicly available data stream gives real-time resolution of issue reporting and city response, and may speak to the long-standing hypothesis regarding the state of a neighborhood and crime (i.e. the ”broken-window” hypothesis, reviewed in [3]).

Materials and methods

SeeClickFix data

~~The first step in our analysis will be to describe the SeeClickFix posts by neighborhood area in New Haven, CT. SCF posts will be aggregated by neighborhood for descriptive statistics but each individual post will be geocoded and placed on a map figure. We will determine the spatial relationship between areas of highest and lowest SeeClickFix utilization. For each year, we will display the quintiles of use for each area depending on the density of posts. We have 2326 unique users and 9356 anonymous posts. We will perform a social network analysis among the 2326 unique SeeClickFix users. We will measure the interconnectedness of community SeeClickFix activity by calculating the average node degrees and clustering coefficients for each neighborhood.~~

New Haven crime data

To determine the impact of SeeClickFix use on violent and non-violent crimes, we will analyze the differences in crime rates both violent and non-violent crime before and after the introduction of SCF between 2007-2015. The model will be adjusted for income, education, age, gender, and race.

Bayesian structural time series model

Crime rates were modeled using a structural model, defined by the observation equation

$$\mathrm{y_t} = \alpha_t + \beta^T x_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2_\epsilon) \tag{1}$$

Where y_t is the observed crime rate per month time series for $t = 1, ..., n$, that is a function of a vector of d latent state variables $\alpha_t = (\alpha_{1t}, ..., \alpha_{dt})$ and ϵ_t is an i.i.d error term with zero mean and variance σ^2_ϵ. The mean (α) at time t is a function of the mean and slope (δ) at time $t-1$, as is the slope

3-1

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + \delta_{t-1} + \eta_{1t} \qquad && \eta_{1t} \sim N(0, \sigma^2_{\eta_{1t}}) \\ \delta_t &= \delta_{t-1} + \eta_{2t} && \eta_{2t} \sim N(0, \sigma^2_{\eta_{2t}}) \end{aligned} \tag{2}$$

with error terms $\eta_t = (\eta_{1t}, ..., \eta_{mt})$ that are i.i.d. random vectors with mean zero, dispersion matrix $\sum = diag(\sigma^2_1, ..., \sigma^2_m)$. The state equations can then be written as

$$\alpha_t = \underset{(d\times d)}{A}\alpha_{t-1} + \underset{(d\times m)}{B}\eta_t$$

$$\begin{pmatrix} \alpha_t \\ \delta_t \\ \gamma_t \\ ... \\ \gamma_{t-s+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & -1 & ... & -1 \\ ... & & ... & & ... \\ 0 & 0 & 0 & ... & 1 \end{pmatrix} \begin{pmatrix} \alpha_{t-1} \\ \delta_{t-1} \\ \gamma_{t-1} \\ ... \\ \gamma_{t-s} \end{pmatrix} + \begin{pmatrix} \eta_{1t} \\ \eta_{2t} \\ \eta_{3t} \\ ... \\ 0 \end{pmatrix}$$

The prior distribution of α_t is $p(\alpha_t \mid Y_{t-1})$, where Y_{t-1} is a vector of observations $y_1, ..., y_{t-1})'$ for $t = 2, 3....$ The likelihood of α_t is $p(y_t \mid \sigma_t, Y_{t-1})$. Therefore the posterior is given by

$$p(\alpha_t \mid Y_{t-1}, y_t) = \frac{p(\alpha_t \mid Y_{t-1})p(y_t \mid \alpha_t, Y_{t-1})}{p(y_t \mid Y_{t-1})} \tag{3}$$

Posteriors are inferred by (1) simulating draws of the model parameter θ and the state vector z given the observed data y_t in the training period, (2) using the posterior simulation to draw from the posterior predictive distribution over the counterfactual time series $\tilde{y}$ given the observed pre-intervention activity, and then (3) using the posterior predictive samples to compute the posterior distribution of the pointwise impact $y_t - \tilde{y}$.

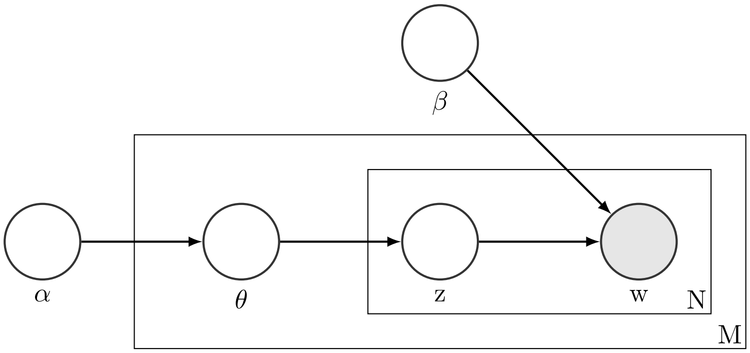
A single alternative neighborhood was used as the covariate to implicitly control for the many variables that may affect crime rates, such as seasonality in temperature and precipitation, as well as social effectors such as changing police chiefs or the hiring additional police officers.

A prior level standard deviation of 0.01 was used

Results and Discussion

Mobile health data share characteristics with both SeeClickFix and crime data in several respects. First, the data are time series with sometimes irregular interval spacing. For

Fig 1. Figure 1. Graphical model of the Bayesian structural time series regression.

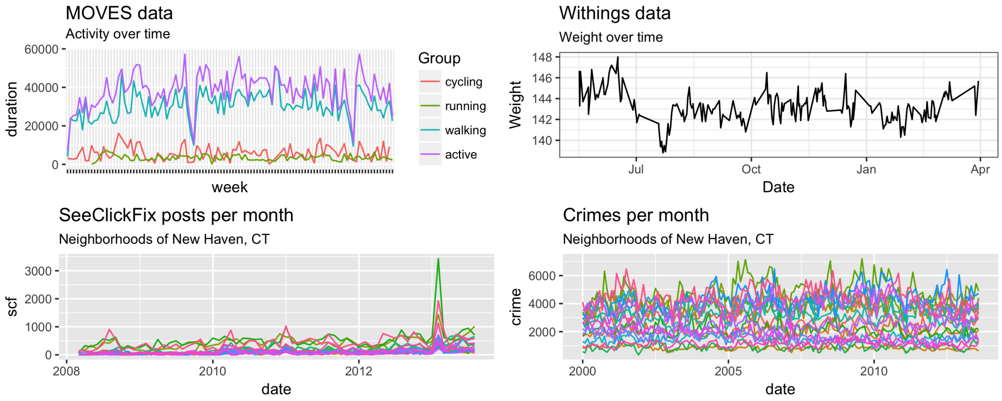


LDA

3^{-1}

example, the Withings weight data show up to four points per day, but mean of just over one data point per day due to several gaps where data are missing (Fig 1B). The crime data show a similar trend when segmented by neighborhoods, where some neighborhoods have many crimes per day and some have on average fewer than one (Fig 1D). As our goal is to determine the long term effects of an intervention; we chose to aggregate to a time unit – days in the case of the Withings data and months in the case of crime data. One could imagine a finer aggregation scale with other types of mobile health data where daily cycles are of interest, such as serum glucose concentration for diabetic patients. Both the mobile health and crime aggregated data

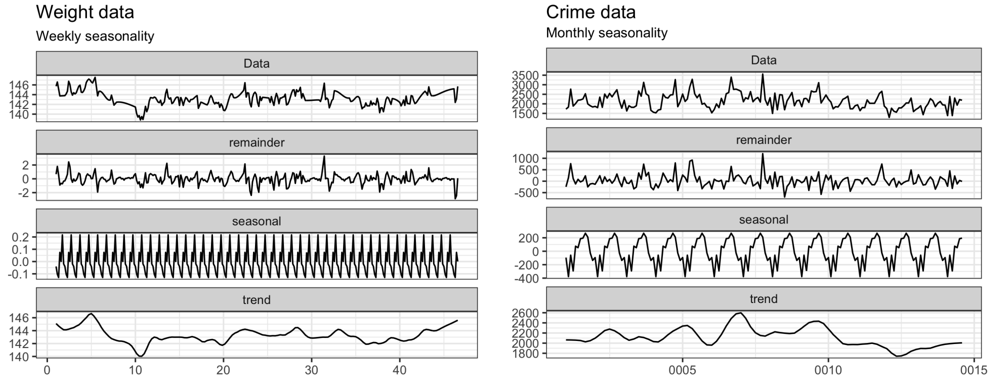
Fig 2. Figure 2. Summary of data types.



show seasonal effect. It is therefore important to use a modeling approach that can accommodate these terms. However, another method of controlling for the seasonality is to use a control time series that shares this feature. In our Bayesian structural time series model we use a time-linked control, i.e. another neighborhood in the same city, as a control for the seasonal variability. This is important because it controls not only for those effects, but also the many latent variables that may affect crime rates such as changes to policing hierarchy. While these are not perfectly controlled by using other neighborhoods (one could imagine one neighborhood getting more attention than another for political reasons, for example) it is better than using time series such as temperature, precipitation and economic indicators (Show in a figure?). A linear-mixed effects model with random effects for each neighborhood shows a significant overall effect of scf, when accounting for such effects as weather and unemployment.

A Poisson logistic regression shows that a few of the neighborhoods show a

Fig 3. Figure 3. Seasonal deconvolution of time series data. Left: Withings weight data with weekly seasonality. Right: Crime data with monthly seasonality.



significant interaction with scf use, but scf alone is not predicted to have an effect. (dwight, newhalville and west river)

Using the bayesian structural time series model, we see that some neighborhoods are predicted to be

Table 1. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				**Heading2**			
cell1row1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
cell1row2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
cell1row3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

LOREM and IPSUM nunc blandit a tortor

3rd level heading

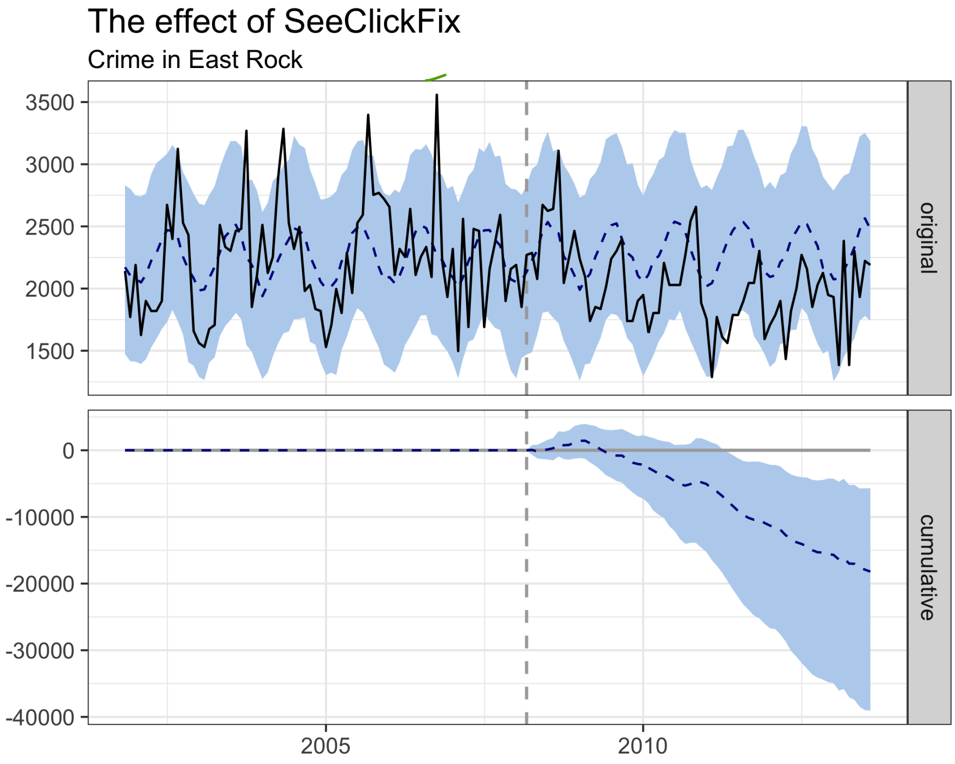
Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id

Fig 4. Figure 4. Causal impact analysis of crime data. Top: Crime data for the New Haven neighborhood ”East Rock” with prediction (blue-dotted line) and confidence intervals. Bottom: Cumulative deviation from the counterfactual prediction of crime in the East Rock neighborhood



massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

In particular, it has advantages over ”difference-in-difference” type causal effect modeling in that it takes into account time-series autocorrelations, and over ARIMA-type forecasts in that the uncertainly can be explicitly defined by priors.

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero [3].

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–950.

2. Ohno S. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.

3. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 2011 Oct;7(10):e1002337.