Hidden Markov Model for the Prediction of Copycat Suicide

Rifat Zahan
Department of Computer Science
University of Saskatchewan
Saskatoon, SK
rifat.zahan@usask.ca

ABSTRACT

Suicide is one of the leading causes of deaths, and the rate of suicide is increasing worldwide. Suicide can be contagious and result in an outbreak, especially in adolescents or younger adults. In this study, we present an automated detection of copycat suicide among particular age groups in U.S. In this study, we use the Hidden Markov Model (HMM) to distinguish the hidden dynamics (copycat or not) in the incidence of suicide. There are about 8,45,178 suicides the U.S. during 1972-1988. The data is obtained from The National Center for Health Statistics (NCHStats). In Statistics, very small data leads to unreliable fit to the real data, and very big data leads to overestimation. HMM can be used in such instances for a reliable fit. The automated detection will help the law-enforcement authority, health-care providers and policy makers to predict the copycat state and work towards decreasing the outbreak of suicide in the U.S. The model built in this study using the data from the U.S. can be tested in other populations to illustrate the utility of the model.

Keywords

Center for Disease Control (CDC); Copycat Suicide; Hidden-Markov Model; National Center for Health Statistics (NCHStats).

1. MOTIVATION

Suicide is one of the main causes of deaths in The United States (Morabito *et al.*, 2015), which is associated with significant social, economic and health system cost. Suicide can be contagious and sometimes can result in a copycat state (Gould *et al.*, 2003). Media has a significant effect on the spread of copycat suicide (Stack, 2003), which is quite common among adolescents and younger adults, that is increasing over time (Garland & Zigler, 1993). In the literature, two types of clusters have been specified regarding suicide: point clusters (related to local events/phenomena) and mass clusters (related to media) (Joiner, 1999). Stack (2003) conducted a meta-analysis of 42 suicide-related studies and found that most of the models in suicide are based on the completers. Stack (2003) also mentioned that finding the "copycat effect" based on completed suicide studies are less likely than studies based on attempts of suicide and stressed the need to conduct further research in this area. Given the high rate of suicide in the US, Center for Disease Control and Prevention (CDC, 2007) stated that program directors and health authorities should focus on suicide prevention activities to reduce the increasing rate of suicide.

To deliver early-stage person-centred or community-based suicide prevention strategies and counselling support, the decision-makers need to know, when the suicide is transitioning in "copycat" state. Several studies have been conducted so far to detect the outbreak of suicide (to prevent suicide before it happens). Davis & Hardy (1986) used deterministic infection epidemic model to develop a tool (model) for suicide outbreak. Ruiz *et al.* (2012) used Indian Buffet Process (IBP) combined with multinomial-logit model to capture hidden features that models suicide attempts.

In this study, we want to develop a model that will predict whether next state will be a copycat suicide state or not based on the data of current state.

2. DATASET OVERVIEW

The data on mortality due to multiple causes in The United States are obtained from the website of Center for Disease Control and Prevention (CDC) (CDC, 2016) Individual records on mortality data are publicly available online in the form of flat files for each years from 1968-2014. Since the interest lies within the counts of suicide and whether it is copycat or not, we have retrieved the information only relating to deaths due to suicide. The date of deaths are not provided during 1968-1971 and 1989-2014. Therefore, we are using only the data during 1972-1988.

After parsing the data using a Python script (from flat file to CSV), we have retrieved information for each individual suicide: year, month, date, state and county of occurrence (in 5-digit FIPS code), sex, International Disease Code (ICD8 during 1972-1978 and ICD9 during 1979-1988) to identify the suicide cases. According to the data documentation, ICD code between E950-E959 are considered as suicide cases. ICD codes are also extended to identify the methods of suicide, which can help us to draw evidence of copycat suicide within county-level.

The geo-map displayed in Figure 1 indicates that the highest number of suicide is observed in California (53,252), followed by Texas (28,712) and Florida (25,214). We further inspect into California state within county-level to look for any pattern on suicide. The geo-map in Figure 2 shows that the counts of suicide is higher in Los Angeles (15,741), which is about 30% of the total occurrence of suicide in this state. Males are more likely to commit suicide (74.11%) than compared to females (25.89%). Among the different means of suicide, the involvement of gunfire is the highest (62.71%) followed by hanging or suffocation (15.21%) and different kinds of drugs (7.25%) in the US during 1972-1988.

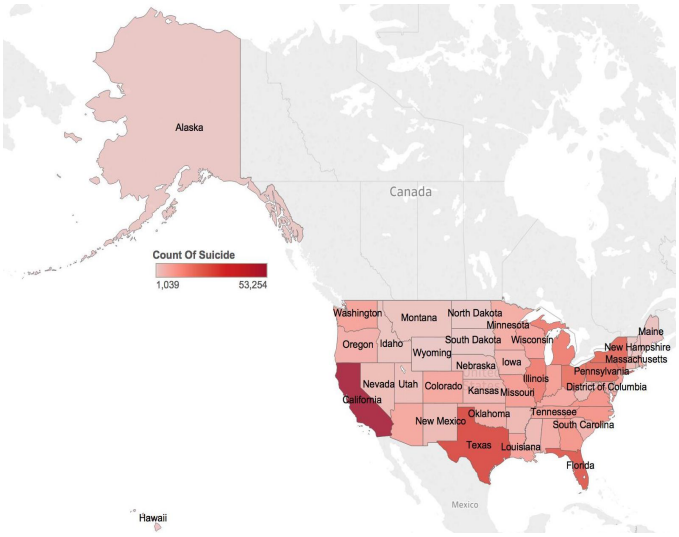

Figure 1: Geo-map of the counts of suicide in different states of US during 1972-1978.

Temporal correlation can be inspected using autocorrelation function. Since we have information on individual death records (i.e., date of occurrence of suicide), we can inspect a plot of autocorrelation function on a quarterly time series data of the suicide, as given in Figure 3. The slow decay of the plot and the significance until lag 3.25 suggests that it may be an evidence of a long-memory process of suicide (serial dependence). The same dependence is observed in separately males and females.

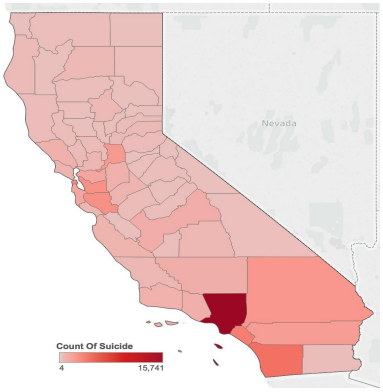

Figure 2: Geo-map of the counts of suicide within counties of California during 1972-1978.

Some data visualization is done using lag-plot. That is, we have plotted the data at quarter t versus quarter $t+1$ in Figure 4. We can see that for males, in the first few years, the counts of suicide in one-quarter does not depend on the previous quarter. But after certain years, there is some dependency visible. For females, the dependency observed throughout the study period. For both males and females, the dependency of the suicide counts between two consecutive states are increasing as time increases.

According to O'Carrol *et al.* (1989), if suicide attempts or group suicides (or both) occurs in a particular space and time beyond the normally expected numbers are termed as

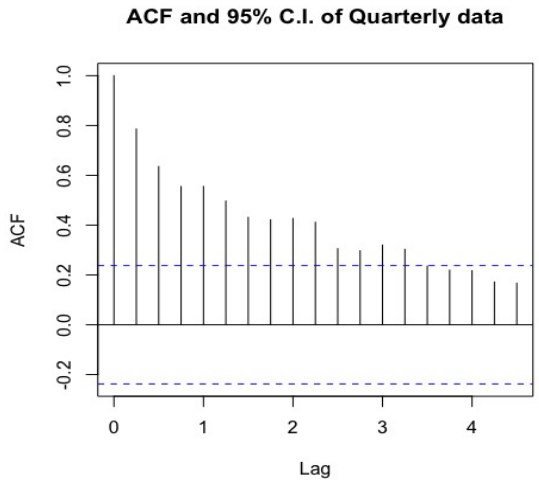

Figure 3: Autocorrelation and 95% confidence interval of quarterly time series data on suicide in Los Angeles during 1972-1978.

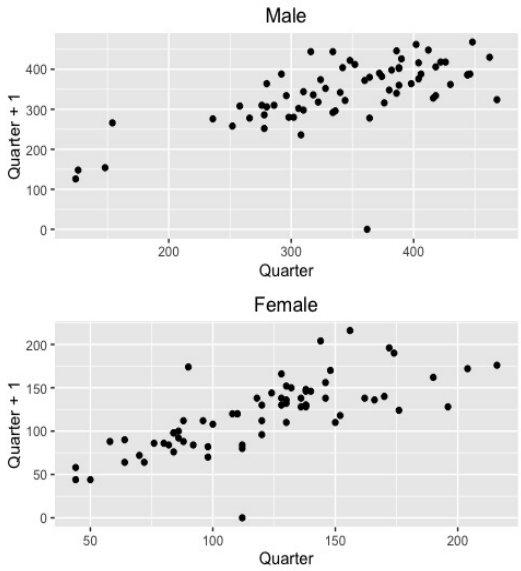

Figure 4: Lag-plot of quarterly time-series data of Los Angeles for both the sexes during 1972-1988.

suicide cluster. We have identified cluster (i.e., The United States –> California–> Los Angeles) with high suicide percentage. Several authors (O'Carrol *et al.*, 1989; Mark, 1997; Messoudi, 2009) used different threshold amount to term one cluster as a suicide cluster. We have set up a threshold value that if the counts of suicide in one quarter is 6% higher than the average counts of the suicide of that particular year, then we will define the suicides occurring that quarter as an outbreak (say, copycat). Based on this threshold amount, we have found the density plot as in Figure 5.

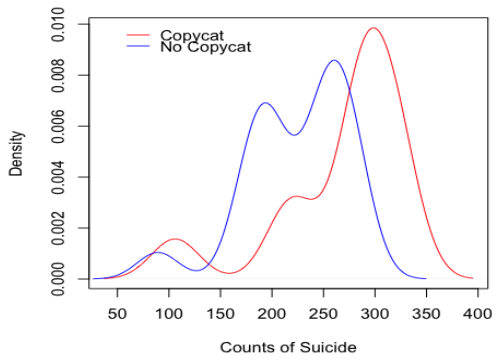

Figure 5: Distribution of quarterly data based on copycat or no-copycat suicide states for Los Angeles during 1972-1988.

3. ALGORITHM DESCRIPTION

For a given count of suicide in a particular quarter, an HMM can be used to predict whether it is a copycat suicide or not. The attractiveness of HMM include their simple mathematical formulation and straightforward computation of likelihood (Zucchini & MacDonald, 2009). A draft trellis diagram is presented in Figure 6 . For example, there were 300 suicides in the first quarter in a particular year and 400 on the second quarter. The first quarter was not in copycat state. Given this count and states of suicide, the second quarter is a copycat state.

As seen in Figure 5, the density curve for non-copycat state is bimodal, whereas, the copycat density plot is unimodal. Given this density plot, it is difficult to decide which distribution does it follow. Due to the fact that it is a data of counts, these two data sets can follow either Poisson or Negative Binomial Distribution. We further inspected the empirical and theoretical distribution function for both copycat and non-copycat data sets. `fitdist` function in `R` package named, `fitdistrplus` is used to fit both Poisson and Negative Binomial distribution. The plots of of the CDFs as presented in Figures 7 and 8 indicates that the data of copycat suicide states follow a Poisson distribution and the non-copycat suicide state follows a Negative Binomial Distribution.

4. FUTURE PLAN

There are some trend, seasonality and cyclic variation in the data set. Therefore, we will have to adjust for these components to get the clear picture of the copycat effect. We will focus on counties with high incidence rate (not the

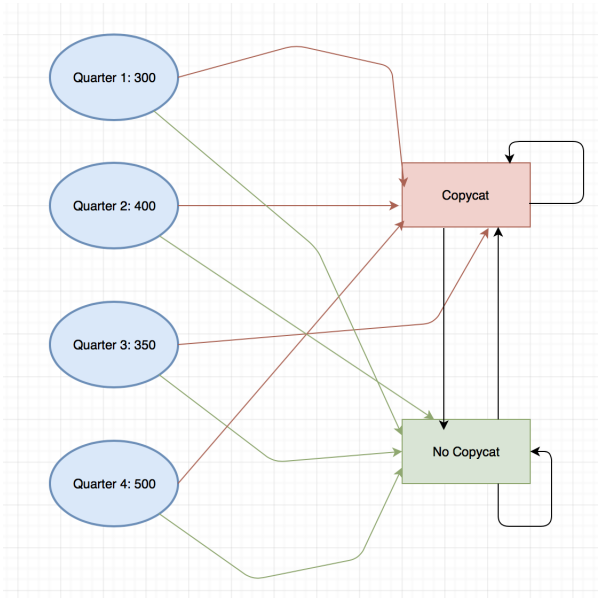

Figure 6: Distribution of quarterly data based on copycat or no-copycat suicide states for Los Angeles during 1972-1988.

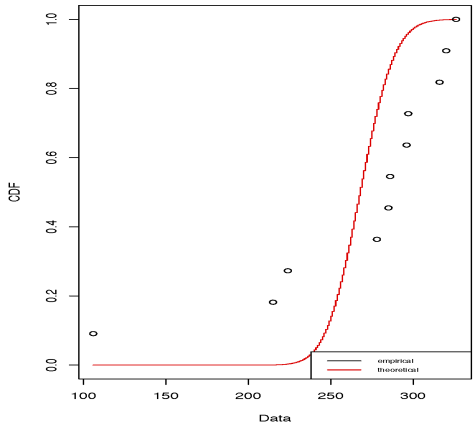

Figure 7: Empirical and theoretical CDF of the data in copycat state during 1972-1988 in Los Angeles.

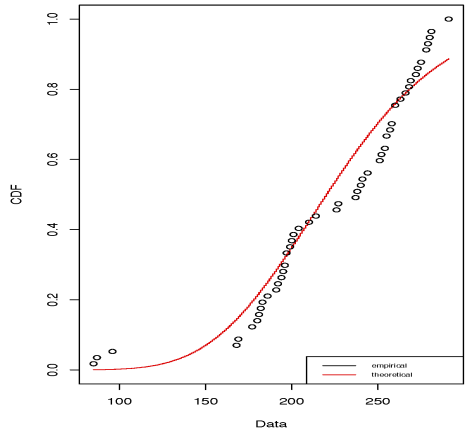

Figure 8: Empirical and theoretical CDF of the data in non-copycat state during 1972-1988 in Los Angeles.

counts/percentage). For this, we need population numbers within counties. Recent suicide data sets (2004-2010) suggests that rural counties have a higher burden of suicide compared to urban counties. Therefore, we will have to find out rural counties with high incidence rate of suicide within specific time period (days/weeks). States like New York, Texas, California is very diversified in terms of people. Thus, A given statistical fluctuation in the overall rate seems less likely to be copycat suicide there. Quarterly view is far too aggregated, and mostly dominated by seasonality. To examine the copycat effects, weekly data are better than quarterly data. Weekly data reveals a better picture of the ripple of a copycat effect than quarterly or monthly data. HMMs are designed to work with quite temporally fine-grained data. Age/sex group is another important factor while examining the copycat effect. Therefore, we will aggregate the data in terms of age group, sex and the type of method used for suicide in a week. These should help to understand the effect of copycat suicide.

As mentioned in the section 3, there will be two hidden states: copycat (C) and no-copycat (NC). We will use a phased approach. In phase 1, we will use an unsupervised learning to arrive at a best-fitting HMM (i.e., to estimate an age/sex-specific Poisson rate of arrival for both the C and NC states and the transition matrix) that best explains the data. Here the HMM would be for a particular age/sex group, and the observation for a given timeslot would be a single number (the number of suicides occurring in that age/sex group for that timeslot). Note that while the Poisson HMM would actually be estimating a coefficient for the Poisson rate of arrival for a given state; that coefficient would then be multiplied by a seasonal rate to yield the Poisson rate of arrival.

In phase 2, we will pursue a more fulsome strategy. The HMM would again be for a given age/sex group. For the observation in the HMM for each successive timepoint, we will use a vector of recent (detrended) counts of suicides in that particular age/sex group, stratified by recent timepoints. This will allow us to explicitly capture the existence of recent suicides in the age/sex group (not merely have that be implicit in the current state). For the phase 2 model, in the NC state, there would be a simple coefficient for the Poisson rate of arrivals of suicides of a given age/gender group (e.g., young men), independent of recent suicides (information on which is captured in the observation vector); this reflects the fact that there are no copycat effects postulated while in that state. In the C state (when copycat effects obtain), the coefficient for the (Poisson) arrival rate of suicides in a given age/gender group depends on the recent suicides in that group (captured in the observation vector). Initially, for simplicity, this dependence in C of the current arrival rate might just be dichotomous (e.g., if there are recent suicides in the past month, we assume a higher coefficient, and thus a higher rate, than otherwise). Later we might look at linear dependence (i.e., that as the number of recent suicides in that age/sex-specific group rises, it is postulated to drive up the of suicide in that same group proportionally).

5. METRICS OF SUCCESS

To investigate the accuracy of the model, i.e., whether it can accurately predict the two hidden states (copycat/no-copycat), we can use the *confusion matrix*. We have labelled the data using a threshold amount explained in section 2. We will split the data in such a way that 33% of the copycat and no copycat observations are in training data and the rest will remain in test data. After implementing the model on test data, we then we will calculate the confusion matrix to check for accuracy. Sometimes confusion matrix itself is not enough to infer about a model accuracy. Therefore, we will also use Receiver Operating Characteristic (ROC) curve by calculating the *sensitivity* and *specificity* to check the model performance.

References

CDC - National Center for Health Statistics. - Homepage. http://www.cdc.gov/nchs/. September 19, 2016

Center for Disease Control and Prevention (2007). Suicide trends among youths and young adults aged 10-24 years–United States, 1990-2004. *MMWR: Morbidity and mortality weekly report*, **56(35)**, 905–908.

Davis, B. R.,, & Hardy, R. J. (1986). A suicide epidemic model. *Social Biology*, **33**, 3–4.

Garland, A. F., & Zigler, E. (1993). Adolescent suicide prevention: Current research and social policy implications. *American Psychologist*, **48(2)**, 169–82.

Gould, M., Jamieson, P., & Romer, D. (2003). Media contagion and suicide among the young. *American Behavioral Scientist*, **46(9)**, 1269–1284.

Joiner, T. E. (1999). The clustering and contagion of suicide, *Current Directions in Psychological Science*, **8(3)**, 89–92.

Kessler, R. C., Berglund, P., Borges, G., Nock, M., & Wang, P. S. (2005). Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003. *Journal of American Medical Association*, **293(20)**, 2487–2495.

Mark, J. G. W. (1997). Cry of pain : understanding suicide and self-harm, *London : Penguin Books*.

Mesoudi, A. (2009). The Cultural Dynamics of Copycat Suicide, *PLoS ONE*, **4(9)**: e7252, doi:10.1371/journal.pone.0007252.

Morabito, P. N., Cook, A. V., Homan, C. M., & Long, M. E. (2015). Aagent-Based Models of Copycat Suicide, *In Social Computing, Behavioral-Cultural Modeling, and Prediction: 8th International Conference, SBP 2015, Washington, DC, USA, March 31-April 3, 2015. Proceedings*, (Vol. 9021, p. 369) Springer.

O'Carroll, P. W., Mercy, J. A., & Steward, J. A. (1989). CDC Recommendations for a Community Plan for the Prevention and Containment of Suicide Clusters, *Morbidity and Mortality Weekly Report*, **37(6)**, 1–12.

Ruiz, F., Valera, I., Blanco, C., & Perez-Cruz, F. (2012). Bayesian nonparametric modeling of suicide attempts. In *Advances in Neural Information Processing Systems*. (pp. 1853–1861).

Stack, S. (2003). Media coverage as a risk factor in suicide, *Journal of Epidemiology and Community Health*, **57**, 238–240.

Zucchini, W., & MacDonald, I. L. (2009). Hidden Markov models for time series: an introduction using R. (Vol. 150). CRC press.