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 results in an outbreak. Media has a significant effect on the occurrence of suicide, which sometimes leads to "copycat suicide." Copycat suicide is quite prevalent in adolescents or younger adults. In this study, we present an automated detection of copycat suicide among a particular group of people in Los-Angeles. We used the Hidden Markov Model (HMM) to predict the hidden dynamics (copycat or not) in the daily counts 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). Simple mathematical formulation and easy-to-maximize the log-likelihood function let HMM give a reliable fit to the data. The data within each state of suicide followed Poisson distribution, so we considered, Poisson distributed likelihood for the HMM parameters. Since, there was very little information on the hidden states, therefore, we used the Baum-Welch (BW) algorithm to update the HMM parameters. A simulation study was also considered in this study to test the model in a data, where we have the ground truth about the hidden states of suicide. The sensitivity ($\sim 50\%$) and specificity ($\sim 50\% - 70\%$) of both the real data and simulated data indicated that the model fit data moderately well. The automated detection will help the law-enforcement authority, health-care providers and policy makers to predict the copycat state of suicide and work towards decreasing the outbreak of suicide in Los-Angeles. The prediction of the copycat state of suicide well in advance can help to provide individual or community-based counselling support before the suicide takes place. The model built in this study using the data from the U.S. can be tested in other populations (e.g., Canada and Australia) to illustrate the utility of the model.

Keywords

AnyLogic; Baum-Welch Algorithm; Center for Disease Control (CDC) and Prevention; Copycat Suicide; Hidden-Markov Model; National Center for Health Statistics (NCHStats); Poisson Distribution; Simulation Modelling.

1. INTRODUCTION

According to the World Health Organization (WHO) about 800,000 suicides occur worldwide and there are a countless number of people, who attempt suicide[36]. About 1.4% of all deaths are due to suicide [36]. Suicide is one of the leading causes of deaths in the United States [21], which is associated with significant social, economic and health system cost [25]. Suicide attempts sometimes result in hospitalization, or permanent disability, requiring long-term care and loss of income [25]. The total cost associated with per completed suicide and per attempted suicide were estimated to be $\$397,000$ (U.S.) and $\$33,000$ (U.S.), respectively [25]. Suicide can be contagious [14], and sometimes can result in a copycat state [13]. Media has a significant effect on the spread of copycat suicide [32], which is quite common among adolescents and younger adults. The rate of copycat suicide among younger adults and adolescents is increasing over time [11]. 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) [16]. For example, visits to Emergency Departments (ED) due to self-harm or suicide attempt increases following the announcement of celebrity suicides [15]. Stack [32] conducted a meta-analysis of 42 suicide-related studies and found that most of the models in suicide are based on the completers. Stack [32] 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 U.S., Center for Disease Control and Prevention [4] stated that program directors and health authorities should focus on suicide prevention activities to reduce the increasing rate of suicide. To provide preventive interventions Preti & Lentini [29] mentioned the importance of predictive analytics of suicide over exploratory analysis.

2. LITERATURE REVIEW

As mentioned earlier in section 1, suicide is related to the social, economic and health system cost. 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 to "copycat" state. Several studies have been conducted so far to predict suicide and to detect the outbreak of suicide (or to prevent suicide before it happens). Davis & Hardy [8] used a deterministic infection epidemic model to develop a tool (model) for suicide outbreak. Nock & Banaji [23] used Self-Injury Implicit Association Test (SI-IAT) to predict suicide ideations and attempts among adolescents. Cheng *et al.* [5] used Poisson time-series autoregressive analysis to examine if there was an increase in the attempts of suicide after large-scale media coverage of celebrity suicide in mid-Taiwan. Ruiz *et al.* [30] used Indian Buffet Process (IBP) combined with multi-

nomial logit model to capture hidden features that models suicide attempts. Suh *et al.* [35] used exponential modelling to predict the copycat suicide in celebrities in South Korea. Preti & Lentini [29] examined the precision of the forecasting models for suicide in both men and women in Italy. The authors used the Naive method, Drift method and Seasonal Naive method to test the influence of the main trend in the monthly counts of suicide. The authors also used the ARIMA, the Holt-Winters seasonal method, the Error, Trend, Seasonal (ETS) model, and the TBATS model (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components) to take into account seasonality and cyclic effect in the time series data. Preti & Lentini [29] concluded that the models with Mean Absolute Percentage (MAP) of 10% could be used to forecast future trend in suicide.

Most of the literature on suicide or copycat suicide is based on the traditional statistical algorithm. There are some studies, which classify or predict suicide using modern machine learning algorithm [20, 12, 26]. Modai *et al.* [20] used neural network for the prediction of suicide in psychiatric patients. Delgado-Gomez *et al.* [12] used support vector machine, decision trees, Elastic net (Lars-en) along with linear regression and stepwise linear regression to classify suicide attempts among 347 men and women in Madrid, Spain. Poulin *et al.* [26] used machine learning algorithm to predict the risk of suicide by analyzing the text of clinical notes taken from the national sample of U.S. Veterans Administration (VA) medical records.

We can see that only few studies developed model for the prediction of copycat suicide or detected outbreak of suicide. Limited application of modern machine learning algorithm in the prediction of copycat suicide stressed the need for further research. Moreover, the existing studies predict suicide given some events or media coverage [31, 14, 33, 15]. There are limited studies that predict the outbreak or epidemic status of suicide in future, based on the current counts of suicide. In this study, we want to develop a model that will predict whether next state (i.e., day) will be a copycat suicide state or not based on the counts of the suicide of current state (day).

3. 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) [3] Individual records on mortality data are publicly available online in the form of flat files for each year from 1968-2014. Since the interest lies within the counts of suicide and whether it is a 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 suicide-related 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 subdivided to determine the methods used to commit suicide, which help us to draw evidence of copycat effect 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 California 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 firearms is the highest (54.72%) followed by hanging or suffocation (17.08%), various kinds of drugs (13.48%) and jumping (5.23%) in Los-Angeles during 1972-1988.

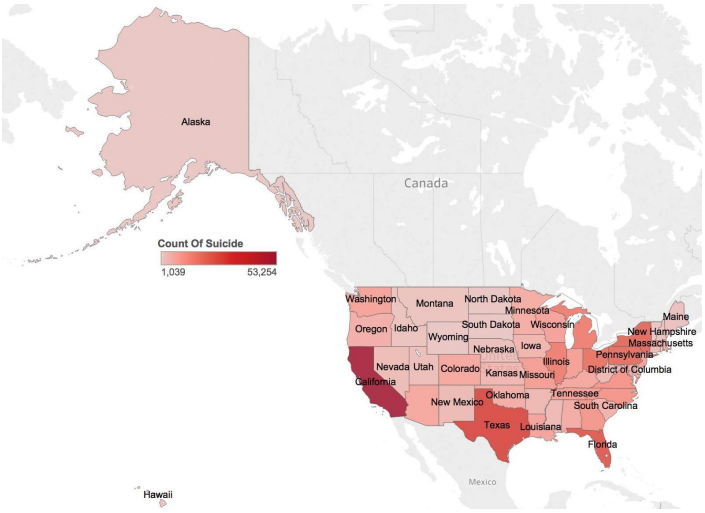


Figure 1: Geo-map of the counts of suicide in different states of US during 1972-1978.

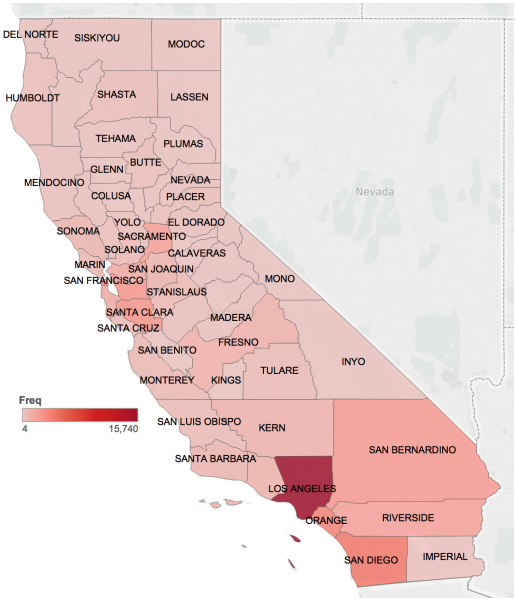


Figure 2: Geo-map of the counts of suicide within counties of California during 1972-1978.

Temporal correlation can be inspected using autocorrelation function (ACF). Since we have information on individual suicide records (i.e., date of occurrence of suicide), we can inspect a plot of autocorrelation function on a daily time series data of the suicide separately for males and females, as given in Figure 3. There exists a linear relationship between the suicides separated by lags. The persistence of high values in ACF plot probably represent a long term positive trend.

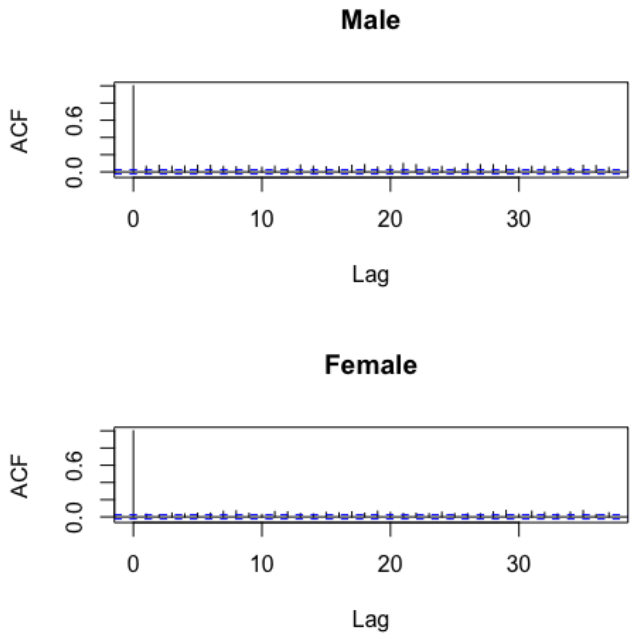
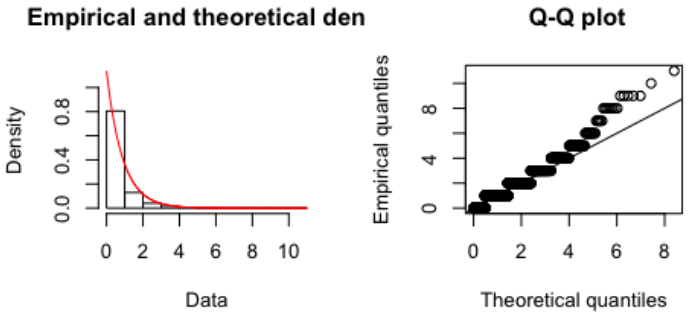


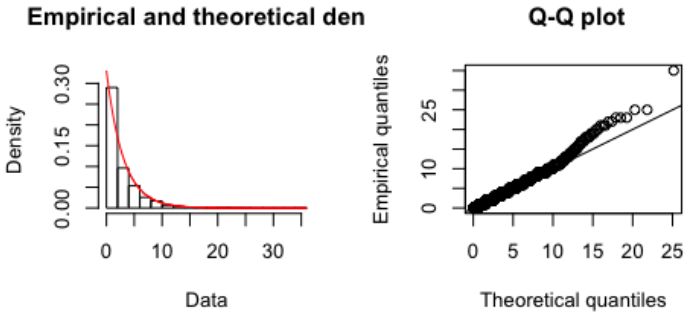
Figure 3: Autocorrelation and 95% confidence interval of daily time series data on suicide in Los-Angeles during 1972-1978.

Some data visualization is done using Q-Q plot and histogram of the days between consecutive suicides. We have inspected the top four methods used to commit suicide in both men and women to look for more copycat effect. The plots are displayed in Figures 4 and 5. The slow decay like exponential distribution in histograms and sudden deviation from the Q-Q plot after a while indicate that there may be a copycat effect in the data of males and females (only those used drugs/medication as means of suicide).

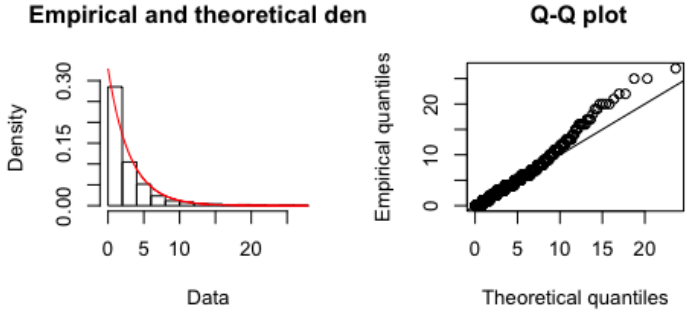
For simplicity, in this project, we will only consider daily counts of suicide of those females, who used drugs to commit suicide. There were a total of 3,600 time series data points for the daily counts if suicide. We will randomly split the data into two different components: (1) training data (67%), and, (2) test data (33%). We do not have labelled sample for copycat state. Therefore, we will hand label the states based on two assumptions: (1) if the number of suicide increases for several consecutive days, those days are considered as copycat state, and, (2) for some consecutive days, if there are counts of suicide present for more than three days, those days are also considered as copycat state. Rest of the counts of suicide in the data will be considered as non-copycat state. We further examine, if the counts of suicide within each state in the study sample follows a Poisson



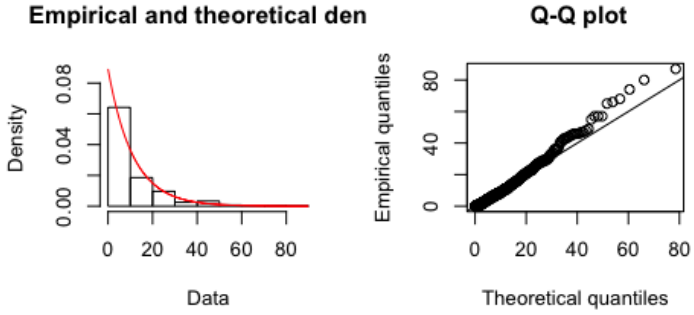
(a) Firearms (Male)



(b) Hanging (Male)

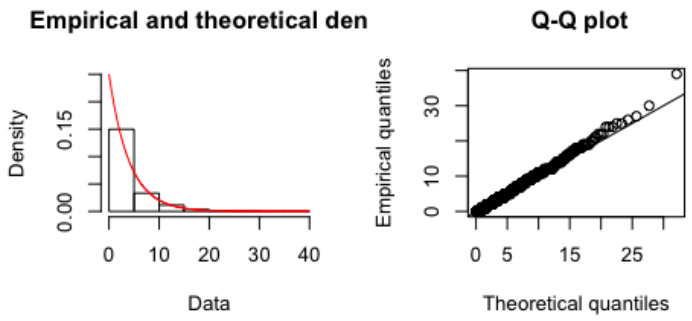


(c) Drugs (Male)

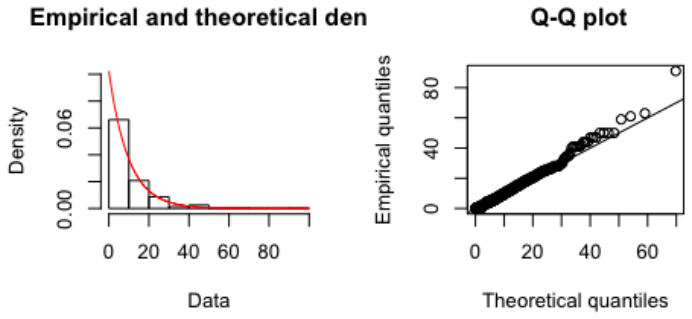


(d) Jumping (Male)

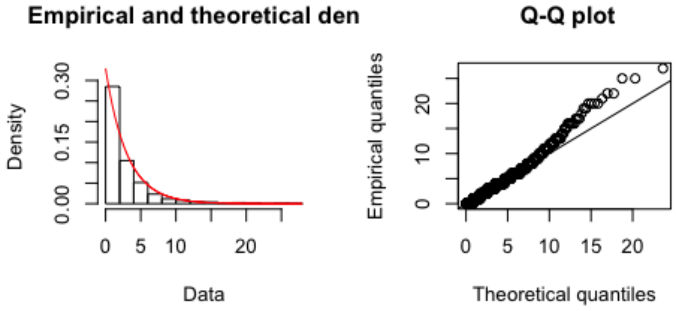
Figure 4: Distribution of the days between successive suicides among males in Los-Angeles during 1972-1988.



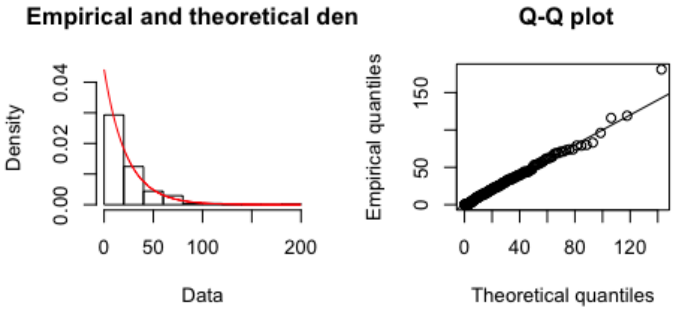
(a) Firearms (Female)



(b) Hanging (Female)

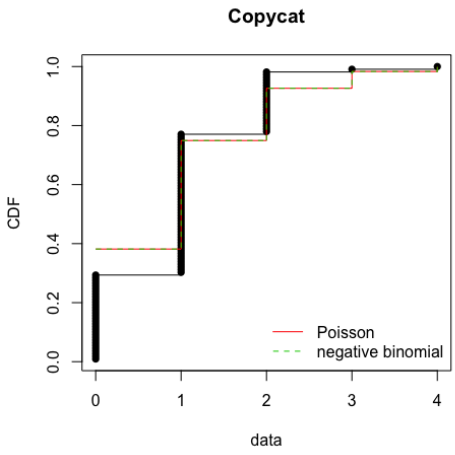


(c) Drugs (Female)



(d) Jumping (Female)

Figure 5: Distribution of the days between successive suicides among females in Los-Angeles during 1972-1988.



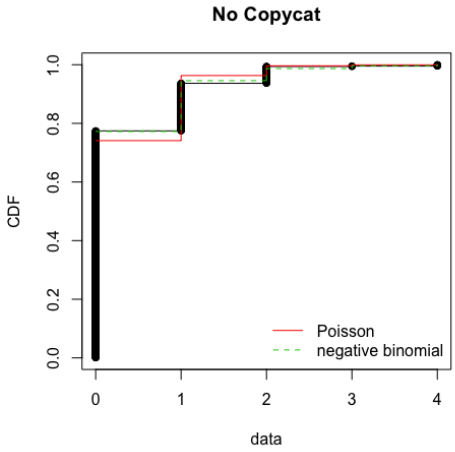


Figure 6: Distribution of the counts of suicide within each hidden state (copycat/no-copycat) for females, who used drugs to commit suicide.

distribution. The CDF (cumulative density function) plots are given in Figure 6. From Figure 6, we can see that both the Poisson distribution and negative binomial distribution fit the data reasonably well. Moreover, Poisson distributed likelihood naturally characterizes arrival process. Therefore, in this study, we will consider Poisson distribution while predicting the hidden states of suicide.

4. ALGORITHM DESCRIPTION

4.1 Hidden Markov Model (HMM)

Billingsley [2] first introduced the concept of HMM. For a given count of suicide in a particular day, an HMM can be used to predict whether it is a copycat suicide state or not. The attractiveness of HMM include their simple mathematical formulation and straightforward computation of likelihood [38]. A graphical representation of the general architecture of HMM is given in Figure 7. Each rounded rectangular and oval shape represents random variables. Let us consider two sequences- $X : x_1, x_2, \ldots, x_n$ and $Y : y_1, y_2, \ldots, y_n$ The random variable $x_t \in X$ is the hidden (copycat/no-copycat) state at time t. $y_t \in Y$ is the observation (counts of suicide)

for day t. The arrows in the diagram indicate conditional dependencies. y_t is conditionally independent of all other observed data (y's) given x_t [10].

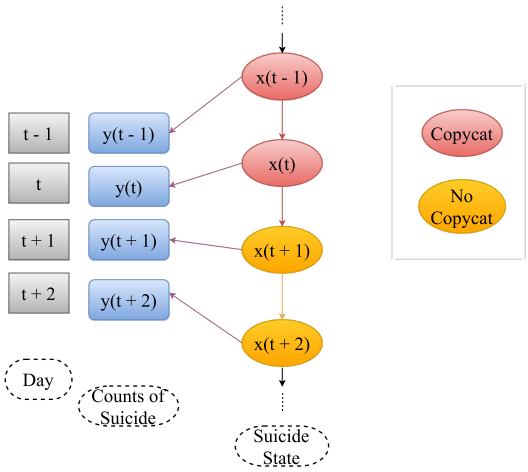


Figure 7: General architecture of an HMM in prediction of suicide state (trellis diagram).

The state space of the suicide state is discrete and the counts of suicide at each day are discrete count data. The HMM has three parameters- initial probabilities, transition probabilities and emission probabilities (or rates). At any moment in time t, the model will be in state x_t. We denote the observed counts of suicide at time t as y_t. Therefore, for a sequence of observed counts of suicide, $Y = y_1, \ldots, y_n$, there will be associated hidden states of suicide, $X = x_1, \ldots, x_n$. The parameters associated with the HMM can be denoted as, $\Theta = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, where $\boldsymbol{\pi}$ is a matrix of initial probability of the states, $\boldsymbol{A}$ is the state transition probability matrix and $\boldsymbol{B}$ is the emission rate vector, which are defined for our study as follows:

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{Copycat} \\ \pi_{No-Copycat} \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

where $a_{ij} = P(x_{t+1} = j | x_t = i)$

The counts of suicide Y is assumed to follow a Poisson distribution within each hidden state. Then $\{(x_t, y_t), t \geq 1$ be a mixture model with

$$P(y_t = Y | x_t = i) = \frac{\exp(-\lambda_i)(\lambda_i)^Y}{Y!} \tag{1}$$

where $i = 1, \ldots, m$ is the number of states, which is 2 in our example. Therefore, the emission rate vector becomes,

$$\begin{aligned} \boldsymbol{B} &= [b_1, b_2] \\ &= [\lambda_{Copycat}, \lambda_{No-Copycat}] \end{aligned}$$

4.2 Parameter Estimation: Baum-Welch (BW) Algorithm

The HMM became widespread since Baum *et al.* [1] presented an algorithm to estimate the parameters of HMM. The algorithm is a special case of Expectation-Maximization (EM), which provides the maximum likelihood estimation (MLE) of the HMM parameters [34]. This algorithm is known as Baum-Welch (BW) algorithm. BW algorithm is used when we do not have information on the hidden states x_t. Then we are in a situation analogous to fitting a mixture model. We mentioned earlier that the observed sequence is conditionally independent of the hidden state sequence. The likelihood function of the Poisson distribution defined in equation 1 is given below:

$$\begin{aligned} P(y_t | x_t = i) &= \prod_{t=1}^{n} P(y_t = Y | x_t = i) \\ b_i(y_t) &= \prod_{t=1}^{n} \frac{\exp(-\lambda_i)(\lambda_i)^Y}{Y!} \end{aligned} \tag{2}$$

The above-mentioned likelihood function does not have a closed form solution when we do not have information on x_t (hidden state). Therefore, we need iterative method to maximize the likelihood to get the maximized model parameters. The training procedure for the above mentioned HMM is conducted using BW algorithm, which consists of iterative re-estimation of the model parameters (Θ) using EM algorithm [37]. At each iteration, the estimation of the model parameters is carried out by use of forward-backward algorithm. Here we make the use of additional variables, that calculates the probabilities of partial observation sequence. The partial probabilities of the observations of forward-backward algorithms are as follows:

$$\begin{aligned} \alpha_t(i) &= P(y_1, y_2, \ldots, y_t, x_t = i | \boldsymbol{\Theta}) \\ \beta_t(i) &= P(y_{t+1}, y_{t+2}, \ldots, y_n, x_t = i | \boldsymbol{\Theta}) \end{aligned}$$

Here $\alpha_t(i)$ and $\beta_t(i)$ are called the forward and backward observations until time t, given state i and model parameters $\boldsymbol{\Theta}$. There are two more statistics, which are useful in simplifying the BW estimation procedure. They are the probability of being in state i at time t and state j at time $t+1$ given the sequence of observations (counts of suicide) Y and model parameters $\boldsymbol{\Theta}$, which is defined as

$$\zeta_t(i, j) = P(x_t = i, x_{t+1} = j | Y, \boldsymbol{\Theta}).$$

We now calculate $\zeta_t(i, j)$ using a product of three terms and dividing them by a normalizing constant, $\alpha_t(i)$, $a_{ij} b_j(y_{t+1})$ and β_{t+1}:

$$\zeta_t(i, j) = \frac{\alpha_t(i) a_{ij} b_i(y_{t+1}) \beta_{t+1}(j)}{P(Y | \boldsymbol{\Theta})}$$

The other statistics of BW algorithm is the probability of being in state i given observation y_t at time t, which is defined as

$$\begin{aligned} \gamma_t(i) &= \sum_{j=1}^{m} \zeta_t(i, j) \\ &= P(x_t = i | y_t, \boldsymbol{\Theta}) \end{aligned}$$

$\gamma_t(i)$ is the number of times we expect to visit state i. Now $\sum_{t=1}^{n-1} \gamma_t(i)$ gives the expected number of outgoing transitions from state i, and $\sum_{t=1}^{n-1} \zeta_t(i,j)$ gives the expected number of times state i and state j are visited consecutively. Terminating at $n-1$ returns the sum $\sum_{t=1}^{n-1} \zeta_t(i,j)$, which is the expected number of transitions made from state i to state j.

We can define the re-estimation formulas for the HMM parameters $\boldsymbol{\Theta'} = (\boldsymbol{\pi'}, \boldsymbol{A'}, \boldsymbol{B'})$ as follows:

1 Initially at $t = 1$, we have $\pi'_i = \gamma_1(i)$, where $i = 1, 2, \ldots, n$.

2 We update $\boldsymbol{A'}$ as

$$a'_{ij} = \frac{\sum_{t=1}^{n-1} \zeta_t(i,j)}{\sum_{t=1}^{n-1} \gamma_t(i)},$$

which is the expected number of transitions from state i to state j divided by the expected number of transitions emitted from state i.

3 We can update $\boldsymbol{B'}$ as

$$b'_j(Y) = \frac{\sum_{t=1, y_t=Y}^{n} \gamma_t(j)}{\sum_{t=1}^{n} \gamma_t(j)},$$

which is the expected number of times state j visited whilst observing Y, divided by the number of times j visited.

Therefore, we re-estimate our model $\boldsymbol{\Theta'} = (\boldsymbol{\pi'}, \boldsymbol{A'}, \boldsymbol{B'})$ re-iteratively, where

$$\begin{aligned} \boldsymbol{\pi'} &= \{\pi'_i\} \\ \boldsymbol{A'} &= \{a'_{ij}\} \\ \boldsymbol{B'} &= \{b'_j(Y)\} \end{aligned}$$

We will check that at each iteration if $P(Y|\boldsymbol{\Theta'}) > P(Y|\boldsymbol{\Theta})$. This will help us to obtain a model that that maximizes the likelihood of producing the given observation sequence Y. The HMM parameters will converge, when the iteration termination condition will meet (i.e., when two results are almost identical to each other).

5. RESULTS

5.1 Parameter Estimates

To start the HMM, we got the initial values for transition probabilities, initial probabilities and emission rates from the hand-labelled data we have. The training data yielded the following results as initial values for HMM:

Initial Probabilities

$$\begin{matrix} Copycat \\ No-Copycat \end{matrix} \begin{pmatrix} 0.045 \\ 0.955 \end{pmatrix}$$

Transition Probability Matrix

		Next State	
		Copycat	No-Copycat
Current State	Copycat	0.055	0.945
	No-Copycat	0.044	0.956

Emission Rates For initial values for emission rates, we have considered the mean of the counts of suicide within copycat and non-copycat states.

$$\boldsymbol{B} = [0.966, 0.300]$$

After training the HMM using Baum-Welch algorithm, we have found the following estimated parameters of HMM:

Transition Probability Matrix Emission Rates

		Next State	
		Copycat	No-Copycat
Current State	Copycat	0.456	0.544
	No-Copycat	0.642	0.358

$$\boldsymbol{B} = \left[0.608, 7.383 \times 10^{-05}\right]$$

The emission rate for non-copycat suicide state is very smaller than the copycat suicide state.

5.2 Simulation Study

The data was also analyzed using agent-based simulation modelling. We used AnyLogic software to generate data that gives label for copycat and non-copycat suicide states. In this model, an agent (person) is considered to be in a scale-free ring network, where the initial speed is considered 10 ms^{-1}. Each person can move into no-acute suicidal ideation state and then into the acute suicidal ideation state. From there, the person can either go back to the non-acute state of suicidal ideation or can commit suicide. A population size of 1000 was considered with the following parameters:

- Hazard rate of starting the copycat suicide period = 0.001/day
- Hazard rate of starting the copycat suicide period = 0.01/day
- Daily hazard for successful suicide from suicidal ideation = 0.05
- Annual hazard rate of suicidal ideation given no copycat effects = 0.05/year
- Daily hazard for departing acute suicidal ideation = 0.1/day
- Probability of suicide notice triggering suicidal ideation = 0.05

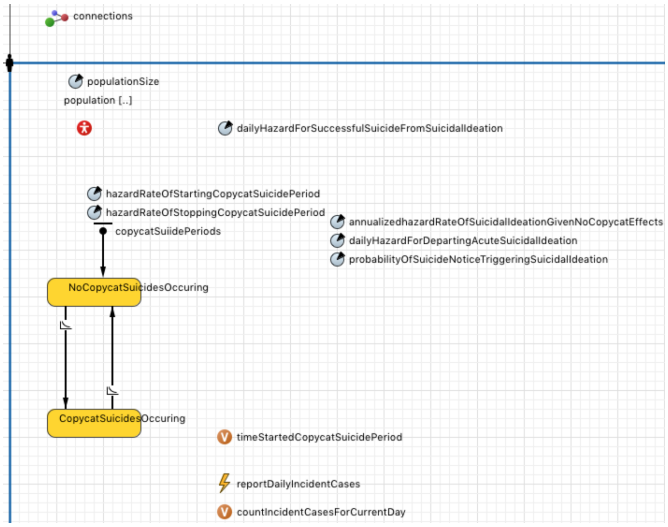
Using the above-mentioned parameters, we can then get some simulated data of the counts of suicide and associated suicide states (copycat/non-copycat). We will use this data to validate the model trained using real training data. Some snapshot of the model are given in Figure 8.

After using the trained HMM, we have predicted the labels of suicide states in simulated data and the test data.

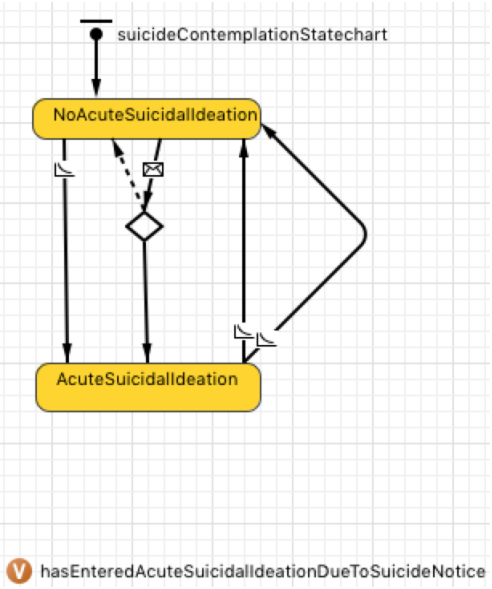
5.3 Model Adequacy Checking

The general structure of a confusion matrix and the formula for sensitivity and specificity are given below:

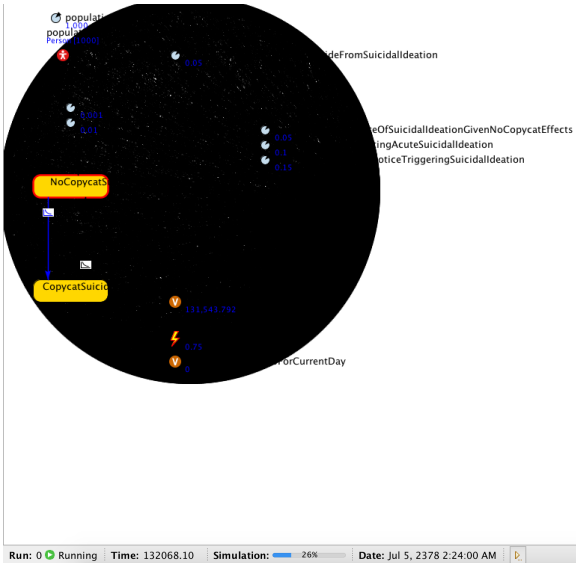
$$\begin{aligned} Sensitivity &= \frac{TP}{TP + FN} \\ Specificity &= \frac{TN}{TN + FP} \end{aligned}$$



(a) Main window



(b) Person window



(c) Model Run

Figure 8: Agent-based simulation model in AnyLogic to generate the daily counts of suicide and associated hidden state of suicide.

		Observed	
		Copycat	No-Copycat
Predicted	Copycat	TP	FP
	No-Copycat	FN	TN

Table 1: Confusion matrix for two states variable.

Here TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

The confusion matrices along with sensitivity and specificity for both real (test) data and the simulated data are given below:

		Observed	
		Copycat	No-Copycat
Predicted	Copycat	590	18
	No-Copycat	540	57

Table 2: Confusion matrix for the real data (test).

$$Sensitivity = 0.52$$
$$Specificity = 0.76$$

		Observed	
		No-Copycat	Copycat
Predicted	No-Copycat	224222	21649
	Copycat	232969	21161

Table 3: Confusion matrix for the simulated data.

$$Sensitivity = 0.49$$
$$Specificity = 0.49$$

The results of sensitivity and specificity indicate that the mode predicted the simulated data reasonably well, since the values are around 50%. The specificity (correctly predicting non-copycat state) for the real (test) data is high (76%), which is not our interest. The aim of this study is to predict the copycat suicide state correctly. Since, the sensitivity is 50%, therefore, the models built in this study can moderately predict the states of the suicide.

6. DISCUSSION

6.1 Discussion of the Study Results

Suicide is preventable, which is facilitated by identifying a population at increased risk [22]. In this study, we have applied machine learning algorithm to predict the outbreak of suicide (i.e., the copycat state of suicide). Chew & McCleary [6] mentioned that aggregated suicide data by age and sex showed that suicide is equally likely to occur at any time (day, week, or month). The suicidal effect is limited to a sub-population, when a particular time is preferred (day, week, or month) [6]. In our study, females, who committed suicide in Los-Angeles using drugs/medication exhibited a copycat pattern in the successive suicidal events. There are

three phases of this study. In phase 1, HMM was applied to predict the copycat/non-copycat state of the suicide counts in Los-Angeles among females, who used medication or some drugs to commit suicide. Baum-Welch (BW) algorithm was used to estimate the model parameters of HMM.

In phase 2, the AnyLogic software was used to generate some simulated data on the counts of suicide for each day (model unit) and associated hidden states (whether it was a copycat or not). This simulated data help us to get a data, where we have information about the hidden states of the suicide.

In phase 3, some cross-validation was carried out using sensitivity and specificity analysis. The values for sensitivity and specificity indicate that the model was able to predict the copycat suicide states moderately well. There may be some reasons for such predictive ability. Several authors [6, 27, 28, 9, 29] have shown that seasonality impacts in the occurrence of suicide. Suicides, which were committed using violent methods exhibit seasonal effect compared to suicides, which were performed by non-violent methods [27].

We did not adjust the data for seasonality effect, which may have a high impact on the counts of suicide. HMMs are designed to work with quite temporally fine-grained data. Therefore, aggregating suicide counts in individual days are useful while applying HMM.

6.2 Knowledge Gained in this Study

This study helped us to learn how a model can be trained to learn the past events and their hidden phenomena, which then predicts the future phenomena based on the events. When we have complete data for the features associated with the study, the model requires a supervised learning. In this study, we did not have the information of the hidden features, therefore, we had to use the unsupervised learning. In unsupervised learning, the model itself learned the data based on limited information of the hidden features. The machine learning algorithm was useful in predicting complex real-world problem which seem to be impractical, but not impossible. For example, in this study, we do not have information on the hidden features, still the model predicted the states moderately well.

Although the model showed moderate level of performance, therefore, we cannot disregard the model. We can try several ways to improve the model. For example, we can apply mixture HMM for training the model. We can also consider multivariate HMM, where we will take into account several other variables, which maybe associated with the copycat suicide state. We may need to employ different machine learning algorithm (e.g., artificial neural network), which will work more efficiently to predict the hidden states of suicide. The model in this study needs more learning features based on some evidence and ground truth before it predicts the data.

7. FUTURE WORK

There are some trend, seasonality and cyclic variation in the daily data of suicide incidence. 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 counts/percentage). For this, we need population numbers within each county. Recent suicide data (2004-2010) suggests that rural counties have a higher burden of suicide incidence compared to urban counties in U.S. Therefore, we will have to find out rural counties with a high incidence rate of suicide within a specific period (days/weeks).

States like New York, Texas, California are very diversified in terms of people. Therefore, a given statistical fluctuation in the overall rate seems less likely to be copycat suicide there.

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. In case, where we do not have information on copycat suicide states, we will use social media data. For example, Twitter or Facebook to get information on suicide related posts or news in small counties or communities. Advanced functional programming (e.g., Spark or Scala) help us to do social data mining, which will help us to identify copycat suicide periods.

As mentioned in the section 4, there will be two hidden states: copycat and no-copycat. We will use a phased approach. In phase 1, we will use a supervised learning approach to arrive at a best-fitting HMM (i.e., to estimate an age/sex-specific Poisson rate of arrival for both the copycat and non-copycat 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 time point, we will use a vector of recent (detrended) counts of suicides in that particular age/sex group, stratified by recent time-points. 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 women), 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 copycat state, 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 copycat 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).

We may also consider the negative binomial distribution since the suicide count data followed negative binomial distribution pretty well (refer to Figure 6). The algorithm in this study should be tested on other real-world population for validation. More advanced method for validation will be used. For example, N-fold validation, leave-One-Out

Cross-Validation (LOOCV), Receiver Operating Characteristic (ROC) curve.

8. SUMMARY

Suicide is a major public health issue, which is associated with economic and human cost [12]. Media has a significant effect in the occurrence of suicide, which sometimes leads to "copycat" effect. News of celebrity suicide or known person(s) in a community can lead to copycat state. Copycat suicide are quite prevalent in adolescents or younger adults. Every completed suicides and suicide attempts are associated with health system cost, or social or economic cost. CDC recommended taking necessary actions to reduce the high burden of suicide in U.S. [4].

In this study, we present an automated detection of copycat suicide among a particular group of people in U.S. We used the Hidden Markov Model (HMM) to predict the hidden dynamics (copycat or not) in the daily counts of suicide in U.S.

Individual death records were obtained from The National Center for Health Statistics (NCHStats), which was publicly available online. Only those data were considered, where the deaths occurred due to suicide. Los-Angeles (LA) in the state of California has the highest counts of suicide. The daily counts of deaths in form of time-series data was obtained using statistical software package `R`. Women living in LA, who used medication/drugs to commit suicide exhibited copycat pattern in the data. The data were hand-labelled to define copycat states based on some pattern visualized in the data. Baum-Welch algorithm (BW) was used to estimate the parameters of Poisson Hidden Markov Model (PHMM).

The sensitivity (50%) and specificity (76%) indicate that the model fitted the data reasonably well. A simulation study was also considered in this study to test the model in a data, where we have the ground truth about the hidden states of suicide. The sensitivity (49%) and specificity (49%) of the simulated data indicated that the model fit data moderately well.

The automated detection of copycat suicide state will help the law-enforcement authority, health-care providers, and policy makers to predict the copycat state of suicide and work towards decreasing the outbreak of suicide in Los-Angeles. It can also help to provide early community-based on individual-based counselling support where we predict that it might be a copycat state of suicide. Time series components (e.g., trend, seasonality and cyclic variation) should be considered while training the model.

References

[1] Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41(1)**,164–171.

[2] Billingsley, P. (1961). Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, 12–40.

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

[4] 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.

[5] Cheng, A. T., Hawton, K., Chen, T. H., Yen, A. M., Chen, C. Y., Chen, L. C., & Teng, P. R. (2007). The influence of media coverage of a celebrity suicide on subsequent suicide attempts. *Journal of Clinical Psychiatry*, **68(6)**, 862–866.

[6] Chew, K.S.Y., & McCleary, R. (1994). A life course theory of suicide risk. *Suicide and Life-Threatening Behavior*, **24(3)**, 234–244.

[7] Chris, T. (2016). *Performance modelling with adaptive hidden Markov models and discriminatory processor sharing queues* (Doctoral dissertation, Imperial College London).

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

[9] Dixon, P. G., Sinyor, M., Schaffer, A., Levitt, A., Haney, C. R., Ellis, K. N., & Sheridan, S. C. (2014). A suicide epidemic model.Association of weekly suicide rates with temperature anomalies in two different climate types. *International journal of environmental research and public health*, **11(11)**, 11627–11644.

[10] Everitt, B. S. (1981). *Finite mixture distributions*. John Wiley & Sons, Ltd.

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

[12] Delgado-Gomez, D., Blasco-Fontecilla, H., Sukno, F., Ramos-Plasencia, M. S., & Baca-Garcia, E. (2012). Suicide attempters classification: Toward predictive models of suicidal behavior. *Neurocomputing*, **92**, 3–8.

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

[14] Gould, M. (2001). Suicide and the Media. *Annals of the New York Academy of Sciences*, **932**, 200-221.

[15] Jeong, J., Shin, S. D., Kim, H., Hong, Y. C., Hwang, S. S., & Lee, E. J. (2012). The effects of celebrity suicide on copycat suicide attempt: a multi-center observational study, *Social Psychiatry and Psychiatric Epidemiology*, **47**, 957âĂŞ965.

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

[17] 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.

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

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

[20] Modai, I., Greenstain, S., Solomish, I., & Mendel, S. (1998). Prediction of suicide in psychiatric patients by neural networks. *European psychiatry*. 147–147.

[21] 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.

[22] Murphy, G. E. (1983). On Suicide Prediction and Prevention. *Archives of General Psychiatry*, **40(3)**, 343–344.

[23] Nock, M. K., & Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of consulting and clinical psychology*, **75(5)**, 707–715.

[24] 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.

[25] Palmer, C. S., Halpern, M. T., & Hatziandreu, E. J. (1995). The Cost of Suicide and Suicide Attempts in the United States, *Clinical Neuropharmacology*, **18**, S25–S33.

[26] Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Flashman, L., & McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes, *PloS one*, **9(1)**, e857333.

[27] Preti, A. (2002). Seasonal variation and meteotropism in suicide: Clinical relevance of findings and implications for research. *Acta Neuropsychiatrica*. **14(1)**, 17–28.

[28] Preti, A. (2011). Animal model and neurobiology of suicide. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. **35(4)**, 818–830.

[29] Preti, A., & Lentini, G. (2016). Forecast models for suicide: time-series analysis with data from Italy. *Chronobiology international*. **33(9)**, 1235–1246.

[30] 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).

[31] Stack, S. (2000). Media impacts on suicide: A quantitative review of 293 findings, *Social Science Quarterly*, **57**, 957–971.

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

[33] Stack, S. (2005). Suicide in the media: A quantitative review of studies based on nonfictional stories, *Suicide and Life-Threatening Behavior*, **35(2)**, 121–133.

[34] Scott, S. L. (1998). *Bayesian methods and extensions for the two state Markov modulated Poisson process* (Doctoral dissertation, Harvard University Cambridge, Massachusetts).

[35] Suh, S., Chang, Y., & Kim, N. (2015). Quantitative exponential modelling of copycat suicides: association with mass media effect in South Korea, *Epidemiology and Psychiatric Sciences*, **24**, 150–157.

[36] WHO - Suicide Data. - Homepage. http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/. December 07, 2016.

[37] Xydas, D., Spencer, M.C., Downes, J. H., Hammond, M. W., Becerra, V. M., Warwick, K., Whalley, B. J., & Nasuto, S. J. (2010). Application of Poisson-based hidden Markov models to in vitro neuronal data. In *Cybernetic Intelligent Systems (CIS), 2010 IEEE 9th International Conference on* (pp. 1-6). IEEE.

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