Suicide Prediction Analysis with Generalized Additive Model

Jun Shen*
School of Economics
Univeristy of Edinburgh
Edinburgh, United Kingdom
jun.shen@ed.ac.uk

Shihui Zhao*
Department of Statistics
Univeristy of Toronto
Toronto, Canada
shihui.zhao@mail.utoronto.ca

Mingzi Ye*
Department of Economics
Southern Methodist University
Dallas, United States
mingziy@smu.edu

***Abstract*—The issue of suicide rate has become increasingly rigorous and has received extensive attention in contemporary society. To explore the pattern of suicide rate variation and intrinsic incentive for suicide, the research was conducted by analyzing the data involved multifaceted factors since suicide is rarely caused by only one factor. Establishing statistical models like the "Generalized Additive Model" provides a conceptual framework for the study and is instrumental in predicting the suicide rate in 2017. Based on the past data from 1985 to 2016, the prediction of suicide rate in 2017 is explicitly presented. Multiple invisible results are revealed, including the negative linear relationship between GDP and suicide rate, the positive correlation with age, and the uneven distribution worldwide. These results presented with functions and graphs show how financial, cultural, or regional factors might affect the suicide rate in different areas. The prediction is still affected by factors challenging to estimate, such as socio-economic relations. However, the results are sufficient to help understand the suicide issue by analyzing its origin and distribution.

***Keywords*—Generalized Additive Model, Data Visualization, Suicide Prediction, GDP

I. Introduction

Suicide has been a serious global public health issue throughout the lifespan. According to statistics from the World Health Organization, nearly 800,000 people die each year from suicide, which means one person dies of suicide every 40 seconds. In recent years, the research on the relationship between suicide rate and the social-economic situation has gradually become the central issue of academic attention.

Psychiatric disorders and socioeconomic elements have proved to be the most closely related to suicide. Durkheim (1897) postulates that economic prosperity is epitomized by urbanization and industrialization, usually accompanied by a grimmer situation of social anomie. Therefore, global suicide rates data and economic development indicators from 1987 to 2016 collected by World Health Organization [1] were dissected. The suicide data in this research were classified by 101 countries, gender, age group, population, Human Development Index (HDI), Gross Domestic Product (GDP), Region, and GDP per capita.

The direct visualization methods can be applied as the impact of economic factors on vicissitudes in suicide rates will significantly discrepant between countries with different development levels. The result was unexpected: the improving of life standard fails to mitigate the rate of suicide. The nonlinear trend is nonnegligible. Furthermore, several regression models are constructed in the research to demonstrate the impact of social factors on suicide rates further.

II. Missing Value

In the research, the Human Development Index (HDI) [2] is missing from the database. As a synthetic index of average achievement, including three substantial dimensions of human development: lifespan, education level, and living standard, HDI contributes to four-level partitions for analyzing human development (Mahbub ul Haq, 1990). Higher life expectancy, education level, or gross national income per capita results in a higher HDI score. According to the Human Development Report Office of the United Nations Development Program (UNDP), the annual variation of HDI in a particular country is exceptionally subtle. Hence, an algorithm to estimate missing values in $\mathcal{H}$ can be given as follows:

Suppose $\mathcal{H} = [\mathbf{H}_\mathrm{i}]$ is the HDI for all 101 countries, where $1 \leq i \leq 101$ and $\mathbf{H}_i = [h_{ij}]$ is the country i HDI from 1986 to 2016, where $1 \leq j \leq 30$.

* Equal contribution

978-1-7281-3885-5/19/$31.00 ©2019 IEEE
IEEE UEMCON 2019, Columbia University, New York, USA

Algorithm 1 Missing values generation

```
procedure    ▷ Estimate missing values in $\mathcal{H}$
    for i from 1 to 101 do
        While missing values exist in $\mathbf{H}_i$
        Suppose $k$ is the first none NA value in $\mathbf{H}_i$
        if $k$ exists then
            $h_{i1}, ..., h_{ik-1} \leftarrow h_{ik}$
        else
            $h_{i1}, ..., h_{i30} \leftarrow h_{2016}$ where $h_{2016}$ is the country $i$ HDI in 2016
            return $\mathbf{H}_i$ (end inner loop)
        end if
        Suppose $l$ is the next none NA value in $\mathbf{H}_i$
        if $l$ exists then
            $h_{ik+1}, ..., h_{il-1} \leftarrow \frac{h_{ik}+h_{il}}{2}$
        else
            $h_{ik+1}, ..., h_{i30} \leftarrow h_k$
            return $\mathbf{H}_i$ (end inner loop)
        end if
    end for
end procedure
```

III. Data Visualization

Figure 1 represents the worldwide proportional suicide rate, with color and blank areas representing varying degrees of suicide rate and missing data, respectively. It can be noticed that suicide is a grievous social issue in Europe, especially during the period from the year 1989 to 2009, followed by America and Oceania are less significant.

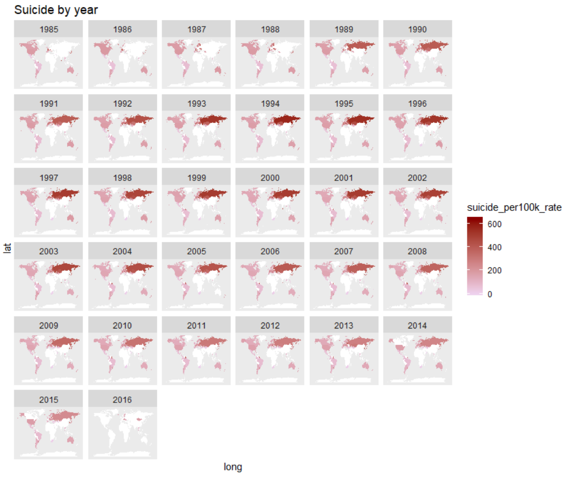

Figure 1. Suicide Rate by Year

Besides this, graphics classified by continents can be created to compare each other. As is exhibited in figure 2, the male suicide rate is higher than that of women. Despite the fact that suicide rates are almost invariant in other countries, male suicide rates tend to undulate by time-varying in Central and Eastern Europe. By collecting and investigating historical events, like the rebuilding of world war II and the disintegration of the Soviet Union, the research will draw the causality between the geopolitical factors and suicide rate.

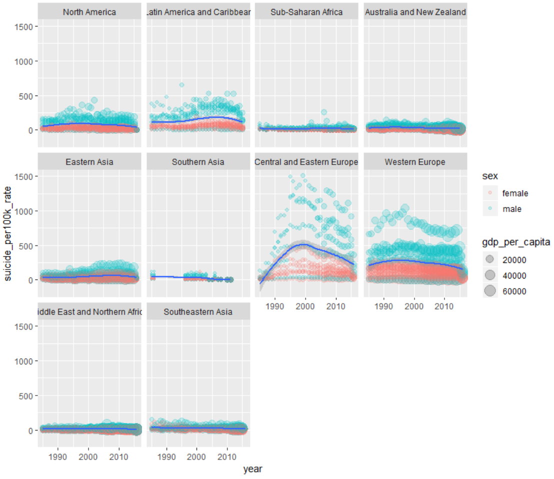

Figure 2. Suicide Rate by continent

Figure 3 reveals the suicide rate in Central and Eastern European countries. The Red label, firebrick label, and blue label represent countries in the Soviet Union, Warsaw Pact, and Yugoslavia separately. In the Soviet Union, one of the countries with the highest suicide rate, approximately 1 million people committed suicide from 1989 to 1991. The reasons for the high suicide rates in the Warsaw Pact organization and Yugoslavia may involve both political factors and economic factors. As the historical records, the cold war, as well as policy glasnost and perestroika, corresponded with the abnormally high suicide rate. After the Revolutions of 1989 in Eastern Europe, the living standards were generally plummeted, resulting in an aggravation of people's family burden and a vicious rise in suicide rates. In addition, some researchers indicated that the high suicide rate is also closely related to Russian alcoholism.

Figure 4 demonstrates the effect of GDP on the suicide rate, presenting a trend of increasing primitively and decreasing afterward. The magnitude of the effect increases with age, and the change in men is more pronounced than that of women.

The impact of GDP on suicide rates is inextricable results from social pressures. The areas with extremely low and high GDP have low suicide rates. In low GDP areas, struggling with food and clothing problems, people are less likely to tackle numerous complicated social issues such as unemployment and failure marriages. On the other hand, most of the regions with higher GDP are developed countries where citizens are entitled to participate in sophisticated welfare and a higher lifespan. Hence, they are also less likely to face

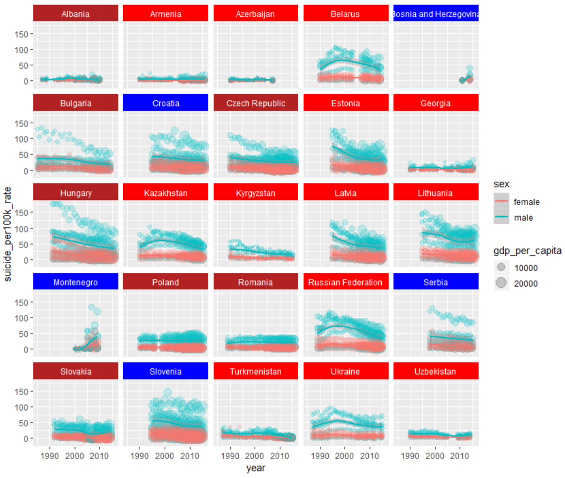

Figure 3. Suicide Rate in Central and Eastern Europe

severe social pressures, which may lead to fewer suicides.

The regions with medium-level GDP are in a high-speed period of economic development. Meanwhile, the economic society in this development period may be characterized by high price commodities, high unemployment, and high instability. For example, in South Korea, most young people might face problems such as the inability to afford housing and the faster replacement of social labor structures. High housing prices have increased the economic burden, thus leading to mental issues for many young people. Simultaneously, the incredibly high unemployment rate caused by rapid replacement also plays an essential role in the high suicide rate.

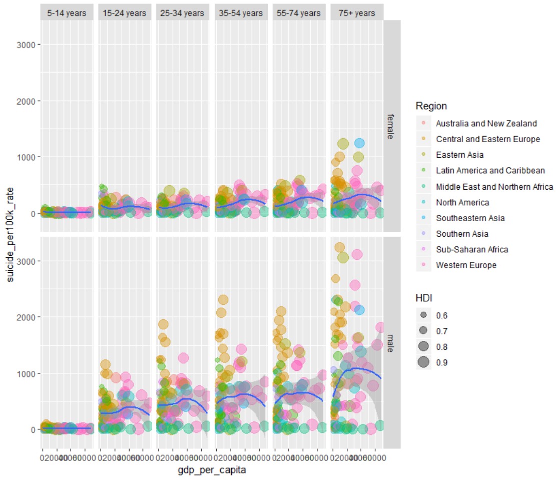

Figure 4. Suicide Rate by GDP and Age

IV. Methodology

The linear regression model can be built with the response variable $\mathbf{Y}$ (‘**suicide per 100k rate**’) and mean structure $\boldsymbol{\mu}$

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{r}.$$

where $\boldsymbol{\mu} = [\mu_i], i = 1, ..., n$ and n is the number of observations. To be more specific,

$$\begin{aligned} \mathbf{Y} \quad &= \quad \mathbf{year} + \mathbf{country} + \mathbf{sex} + \mathbf{age} + \\ &\quad\quad \mathbf{population} + \mathbf{GDPperCapita} + \\ &\quad\quad \mathbf{GDP} + \mathbf{Region} + \mathbf{HDI} + \mathbf{r} \qquad (0) \end{aligned}$$

where $r_i \overset{iid}{\sim} N(0, \sigma^2)$. In model(0), ‘year’, ‘population’, ‘GDPperCapita’ and ‘GDP’ are numerical variables, while the rest are categorical variables (which would be turned into dummy variables in computation). Due to the possibility of multicollinearity, “step wise BIC” method [3] [4]

$$\text{BIC} = log(n)k - 2log(\hat{l})$$

could be implemented to select variables, where k is the estimated parameters and $\hat{l}$ is maximized likelihood function.

After the procedure, explanatory variable ‘population’, ‘GDP’ and ‘Region’ are dropped. Thus, the model is rewritten as

$$\begin{aligned} \mathbf{Y} \quad &= \quad \mathbf{year} + \mathbf{country} + \mathbf{sex} + \mathbf{age} + \\ &\quad\quad \mathbf{GDPperCapita} + \mathbf{HDI} + \mathbf{r} \qquad (1) \end{aligned}$$

and the residual sum of squares is

$$\text{RSS}_1 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

where $\mathbf{X}$ is the column space of independent variables. By the Ordinary Least Square(OLS) method, the R^2 of our fit is 0.5205, which indicates 52.05% of the variability of $\mathbf{Y}$ that is predictable from the independent variables. Some residual tests can be performed here:

1) Shapiro test [5]: Shapiro test is a test of normality. The p-value of the test is less than $2.2e^{-16}$ which indicates residuals do not follow normal distribution.
2) Wald–Wolfowitz runs test [6]: Run test is a test that checks randomness. Since p-value is less than $2.2e^{-16}$ which means there is evidence against null hypothesis, residuals are not independent.
3) Breusch–Pagan test [7]: Breusch–Pagan test is used to test for heteroskedasticity. Due to the 0 p-value, heteroskedasticity is present.

All three residual tests reject the null hypothesis, indicating that the linear model is invalid. Hence, we need to modify model(1) to improve the performance. What’s more, in model(1), the continuous independent variables are restricted and cannot be set to specific parametric form. To be more general, the model could be

$$\begin{aligned}\mathbf{Y} &= s(\mathbf{year}) + \mathbf{country} + \mathbf{sex} + \mathbf{age} + \\ & s(\mathbf{GDPperCapita}) + s(\mathbf{HDI}) + \mathbf{r} \qquad (2)\end{aligned}$$

where $s(\cdot)$ is an unknown parametric form, which can be polynomial, a smooth spline, and etc. This model is widely known as Generalized additive model (GAM) [8]. Its penalized residual sum of squares is defined as:

$$\text{RSS}_2 = (\mathbf{Y} - \mathbf{N}\boldsymbol{\gamma})^T(\mathbf{Y} - \mathbf{N}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}_N\boldsymbol{\gamma}$$

where $\mathbf{N} = [N_{ij}]$ is an $n \times n$ matrix whose (i, j) element $\mathbf{N}_{ij} = N_j(\mathbf{x}_i)$ is the jth natural cubic spline basis function [9] evaluated by $\mathbf{X}$ column space; $\boldsymbol{\Omega}_N = [\omega_{ij}]$ is an $n \times n$ matrix whose (i, j) element is $\omega_{ij} = \int(\frac{\partial^2 N_i(t)}{\partial year^2} + \frac{\partial^2 N_i(t)}{\partial GDPperCapita^2} + \frac{\partial^2 N_i(t)}{\partial HDI^2})(\frac{\partial^2 N_j(t)}{\partial year^2} + \frac{\partial^2 N_j(t)}{\partial GDPperCapita^2} + \frac{\partial^2 N_j(t)}{\partial HDI^2})dt.$ The solution is easily found to be

$$\hat{\gamma} = (\mathbf{N}^T\mathbf{N} + \lambda\boldsymbol{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{Y}.$$

Smoothing parameter λ is related to degree of freedom df_2 and can be optimized to pursuit minimal Generalized Cross Validation (GCV)

$$\text{GCV} = \frac{nD}{(n - df_2)^2}$$

criterion, where D is the deviance. After fitting "mgcv" package [10] [11], we can get the $R^2 = 0.521$ which slightly increases and GCV $= 173$. To compare these two models, "anova"(Analysis of variance) [12] table can be performed here

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(df_2 - df_1)}{\text{RSS}_2/(n - df_2)}$$

and $F \sim F(df_2 - df_1, n - df_2)$. Since the p-value is close to zero, which means model(2) is not reducible and more proper than model(1).

In model(1) and model(2), we assume $y_i \sim N(\mu_i, \sigma^2)$, nevertheless, y_i represents 'suicide per 100k' which is non-negative. A more appropriate assumption would be $y_i \sim Gamma(\mu_i, \phi)$. Thus, our model should be

$$\begin{aligned}log(\boldsymbol{\mu}) &= s(\mathbf{year}) + \mathbf{country} + \mathbf{sex} + \\ & \mathbf{age} + s(\mathbf{GDPperCapita}) + \\ & s(\mathbf{HDI}) \qquad (3)\end{aligned}$$

where $r_i \overset{iid}{\sim} Gamma(0, \phi)$. All important features are listed at the following table:

	R^2	BIC	GCV	Deviance
model(1)	0.5205	144389.7	173.75	4795488
model(2)	0.5209	222315.3	172.99	4769898
model(3)	0.7054	120459.2	3.71	102289

we notice the GCV decreases significantly from 173 to 3.71 and R^2 increases sharply from 0.521 to 0.705! Also, the deviance drops from 4769898 to 102289. All these symbols indicate model(3) is overwhelmingly the best of all.

V. Prediction

According to the data collected from the year 1985 to 2016, the suicide rate can be predicted in 2017 based on model(3). As illustrated in Figure 5 about the worldwide suicide per100k rate in 2017, the suicide rate in Central Europe is hugely higher than other continents.

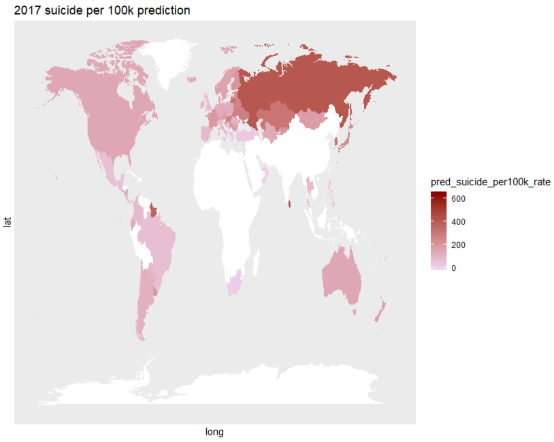

Figure 5. 2017 suicide prediction

Figure 6 reveals the relevance between "GDP" and predicted suicide rate. Sri Lanka saw a significant surge in the suicide rate, reaching above 400 suicide per100k rate and a noticeable rise in the north of South American countries like Guyana and Suriname. By contrast, a downward tendency was found in Southeast Asian nations in 2017, with suicide per100k rate decreasing by 200 approximately. Nonetheless, the majority of Central European countries still stand out as having the highest level among others.

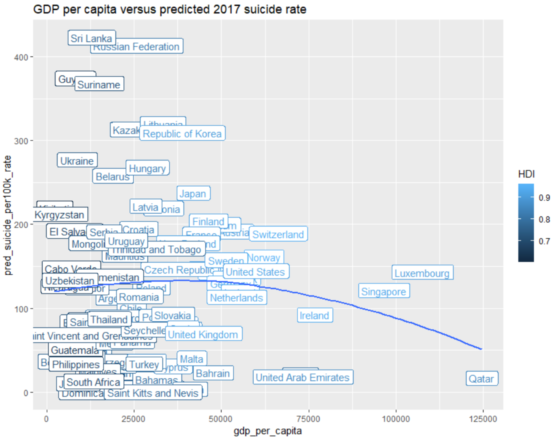

Figure 6. suicide rate versus GDP per capita

Figure 7 explicates suicide per100k rate versus GDP per capita classified by gender and age.

The line graphs compare suicide per100k rate of males with females by progressively increasing age groups based on GDP per capita. There is an upward trend initially and then downward for each graph. The graph indicates that the male suicide rate is around twice that of females, reflecting the consistency with data collected in this research. When it comes to the striking dissimilarity for age groups, the suicide rate increases with age and reaches a peak when people over 75 years old.

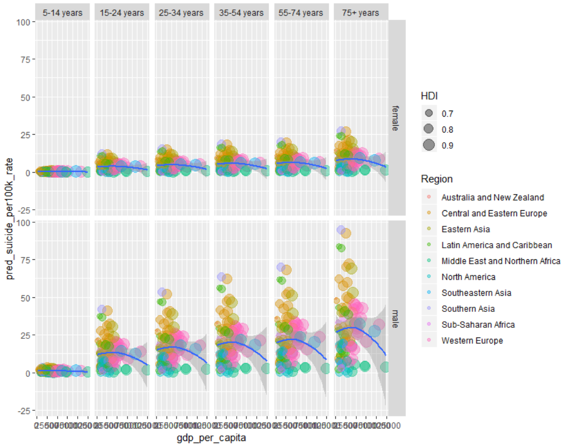

Figure 7. suicide rate by age and gender

VI. Discussion

In the research of the relationship between suicide rate and socio-economics, although some accomplishments have been achieved in the field of sociology and development economics, boundness and faultiness still cannot be ignored, such as the deficiency of detailed comparative studies and horizontal comparisons, especially the dynamic study of suicide rate with the development of socio-economic relations.

By analyzing the correlation between suicide rate and economic condition in various countries via statistical analysis, it is clear that developed areas such as Central and Eastern Europe are more likely to have a higher suicide rate compared to the developing countries. Also, suicides are more common in the elder or male groups.

The research results will further elaborate on the pattern of suicide rate variation with social and economic status. Meanwhile, the research will provide the scientific rationale for suicide prognosis and prevention interventions by describing how the suicide rate is related to different genders and age groups. Socially, the suicide intervention training programs such as Applied Suicide Intervention Skills Training (ASIST) to assess people's risk factors scientifically should be encouraged to establish more and open to the public. (Gould, Cross, Pisani, Munfakh & Kleinman, 2013) [13]. The knowledge about suicide needs to be popularized and attract the public's attention. Personally, seeking help actively when faced with a dilemma would contribute to alleviating melancholy. Sharing one's feelings with families or friends is an effective approach to cope with stress. To ease the pressing problem in worldwide society and help people who have suicide tendencies find alternatives to their problems is the ultimate goal of this suicide prediction analysis.

However, it does not mean the statistical analysis of the suicide rate is faultless.

> "All models are wrong, but some are useful" – George Box [14]

To this topic, the model we build could serve as general guidance to the government health organization. However, specific actions might differ according to different cases.

References

[1] Rusty. (2018) Suicide rates overview 1985 to 2016. [Online]. Available: https://wiki.math.uwaterloo.ca/statwiki/index.php?title=stat841f11

[2] Max.R. (2014) Human development index(hdi). [Online]. Available: https://ourworldindata.org/human-development-index

[3] N. Draper and H. Smith, *Applied Regression Analysis, 2d Edition*. New York: John Wiley & Sons, Inc., 1981.

[4] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 1978.

[5] Y. W. NM Razali, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of Statistical Modelling: Theory and Applications*, 2011.

[6] Bradley, *Distribution-Free Statistical Tests*. Prentice-Hall, 1968.

[7] Econometrica, "A simple test for heteroskedasticity and random coefficient variation," *R package version, 2017*, 1979.

[8] R. J. Hastie, T. J.; Tibshirani, "Generalized additive models," *Chapman & Hall/CRC*, 1990.

[9] C. H. Reinsch, "Smoothing by spline functions," *Numerische Mathematik*, 1967.

[10] S. Wood. (2019) "mgcv: Mixed gam computation vehicle with automatic smoothness estimation". [Online]. Available: https://CRAN.R-project.org/package=mgcv

[11] ——. mgcv: Gams in r. [Online]. Available: https://people.maths.bris.ac.uk/~sw15190/mgcv/tampere/mgcv.pdf

[12] H. Scheffe, *The analysis of variance*. A Wiley-Interscience Publication JOHN WILEY & SONS, INC, 1999.

[13] P. M. Gould, Cross and Kleinman. (2013) Impact of applied suicide intervention skills training (asist) on national suicide prevention lifeline counselor. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3838495/

[14] G. Box, "Science and statistics," *Journal of the American Statistical Association*, 1976.