STATISTICAL AND MACHINE LEARNING MODELLING OF SUICIDE RATE WITH RESPECT TO PROVINCES IN TURKEY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SÜLEYMAN ERDOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

JULY 2022

Approval of the thesis:

STATISTICAL AND MACHINE LEARNING MODELLING OF SUICIDE RATE WITH RESPECT TO PROVINCES IN TURKEY

submitted by **SÜLEYMAN ERDOĞAN** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Özlem İlk Dağ
Head of Department, **Statistics** _____

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Prof. Dr. Ceylan Talu Yozgatlıgil
Statistics, METU _____

Prof. Dr. Özlem İlk Dağ
Statistics, METU _____

Assoc. Prof. Dr. Rukiye Dağalp
Statistics,Ankara University _____

Date: 18.08.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Süleyman Erdoğan

Signature :

ABSTRACT

STATISTICAL AND MACHINE LEARNING MODELLING OF SUICIDE RATE WITH RESPECT TO PROVINCES IN TURKEY

Erdoğan, Süleyman
M.S., Department of Statistics
Supervisor: Prof. Dr. Özlem İlk Dağ

July 2022, 75 pages

There are many factors that can affect the individual's mental health and thus lead them to commit suicide. This study, focuses on the social and economic factors that may contribute to the suicide rates in all provinces of Turkey. Possible effects of these factors are studied in an 8 year period between 2012-2019 using standard longitudinal data modelling methods and hybrid modelling methods. Standard longitudinal data models include; fixed effect models,random effect models and transition models whereas hybrid models include Mixed Effect Regression Tree(MERT) models,Mixed Effect Random Forest(MERF) models,Random Effect - Expectation Maximization Trees(RE-EM Tree) models. The overall results suggest that the divorce and healthcare accessibility of the provinces have significant relation with the suicide rates of provinces. Another observed result was the hybrid models overall performed better than the standard longitudinal data models.

Keywords: Longitudinal Data, Fixed Effect Models, Random Effects Models, Transition Models, Hybrid Models

ÖZ

TÜRKİYE'DEKİ İNTİHAR HIZININ İLLER BAZINDA İSTATİSTİKSEL VE MAKİNE ÖĞRENMESİ MODELLEMESİ

Erdoğan, Süleyman
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi: Prof. Dr. Özlem İlk Dağ

Temmuz 2022 , 75 sayfa

Ruhsal sağlığı etkileyen ve kişiyi intihara sürükleyen birçok faktör vardır. Bu çalışma, Türkiye'deki bütün illerdeki intihar oranı ile bağlantılı olabilecek sosyal ve ekonomik faktörleri incelemektedir. Bu olası faktörler, 2012-2019 yıllarını kapsayan 8 yıllık bir dönemde standart uzunlamasına veri modeli metodları ve hibrit model metodları kullanılarak analiz edilmiştir. Standart uzunlamasına veri modelleri; Sabit Etkiler Modeli, Rassal Etkiler Modeli ve Geçiş Modelini kapsamaktayken, hibrit modeller karışık etki regresyon ağacı(MERT) modeli, karışık etki rastgele orman(MERF) modeli ve rassal etki - beklenti maksimizasyonu ağacı(RE-EM Tree) modellerini kapsamaktadır. Modellerin sonucunda, faktörler arasından boşanma ve illerdeki sağlık imkanlarının, illerdeki intihar oranı ile önemli bir ilişkisi olduğu görülmüştür. Bir diğer gözlemlenen sonuçta ise hibrit modellerin, standart uzunlamasına veri modellerinden daha iyi sonuç verdiği gözlemlenmiştir.

Anahtar Kelimeler: Uzunlamasına Veri, Sabit Etkiler Modeli, Rassal Etkiler Modeli,

Geçiş Modeli, Hibrit Modeli

To My Parents

ACKNOWLEDGMENTS

First of all I would like to express my deepest and the most sincere gratitude to my supervisor Prof. Dr. Özlem İlk Dağ for her guidance and knowledge through this ordeal. I am especially grateful for her patience during the pandemic struggles that I had.

I would also like to thank my friends from the department, Caner and Zeynep. We started our Masters journey together and learned a lot from each other. For me, meeting and getting to know them was a privilege and a very valuable experience that I gained during my time as a student.

Finally I would like to express my gratitude to my parents. Without their support during my studies, I would've never managed to get where I am today. I love them both deeply and forever grateful for everything they have done.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vi

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES xii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1 INTRODUCTION 1

2 LITERATURE REVIEW 3

3 METHODOLOGY 9

3.1 Longitudinal Data Models 9

3.1.1 Random Effects Models 9

3.1.2 Fixed Effects Models 10

3.1.3 Transition Models 11

3.2 Hybrid Machine Learning Models 12

3.2.1 Mixed Effect Regression Trees (MERT) 12

3.2.2 Mixed Effect Random Forest (MERF) 14

3.2.3 Regression Expectation Maximization Trees (RE-EM Trees) . 15
3.3 Comparison Metrics 16
4 RESULTS 19
4.1 Exploratory Analysis 19
4.2 Standard Longitudinal Data Models 24
4.2.1 Random Effect Models 25
4.2.2 Fixed Effect Model 28
4.2.3 Transition Models 31
4.3 Hybrid Methods 33
4.3.1 MERT Model 34
4.3.2 MERF Model 36
4.3.3 RE-EM Trees Model 40
4.4 Model Comparisons and Discussion 42
5 CONCLUSION 47
REFERENCES 49
APPENDICES 50
A CORRELATIONS BETWEEN NUMERIC VARIABLES 51
B NORMALITY TEST TABLES FOR RESIDUALS 53
C R CODES 57

LIST OF TABLES

TABLES

Table 4.1 Descriptive Statistics 21

Table 4.2 Shapiro Wilk Normality Test Results of each year 22

Table 4.3 Shapiro Wilk Normality Test Results after transformation 23

Table 4.4 Temporal Correlations of Suicide Rate 24

Table 4.5 VIF values of the each variable in the random effect model 25

Table 4.6 Results of the Random Effect Model with significant variables . . . 25

Table 4.7 Results of the Final Random Effect Model 26

Table 4.8 Results of the Fixed Effect Model for unstructured correlation structure 28

Table 4.9 Results of the Fixed Effect Model for AR(1) Correlation Structure . 28

Table 4.10 Results of the Fixed Effect Model for Exchangeable Correlation Structure 29

Table 4.11 Results of the Final Fixed Effect Model 29

Table 4.12 Results of the Transition Model 32

Table 4.13 Variable Importance of the Regression Tree in MERT Model 34

Table 4.14 Comparison of the three MERT Models 34

Table 4.15 Importance of variables in the Random Forest part of the MERF Model 37

Table 4.16 Comparison of the three MERF Models 37

Table 4.17 Variable Importance of the Regression Tree in RE-EM Tree Model . 40

Table 4.18 Comparison of the three REEM-Tree Models 41

Table 4.19 Comparison Metrics for all models 42

Table 4.20 Correlations between the predicted suicide rate and actual suicide rate values for each model 43

Table 4.21 Comparison of Actual Values with Predicted Values for Bayburt Province 44

Table 4.22 Comparison of Actual Values with Predicted Values for Trabzon Province 44

Table 4.23 Comparison of Actual Values with Predicted Values for Konya Province 45

Table 4.24 Comparison of Actual Values with Predicted Values for Istanbul Province 45

Table A.1 Correlations between numeric variables 51

Table B.1 Shapiro Wilk Normality Test Results for Random Effect Model Residuals 53

Table B.2 Shapiro Wilk Normality Test Results for Fixed Effect Model Residuals 53

Table B.3 Shapiro Wilk Normality Test Results for Transition Model Residuals 54

Table B.4 Shapiro Wilk Normality Test Results for MERT Model Residuals . . 54

Table B.5 Shapiro Wilk Normality Test Results for MERF Model Residuals . . 55

Table B.6 Shapiro Wilk Normality Test Results for RE-EM Tree Model Residuals 55

LIST OF FIGURES

FIGURES

Figure 4.1 Boxplots of suicide rates of each year 22

Figure 4.2 Box-plots of Standardized Residuals of Random Effect Model in each year 27

Figure 4.3 Random effect model fitted values vs true values 27

Figure 4.4 Box-plots of Standardized Residuals of Fixed Effects Model in each year 30

Figure 4.5 Fixed effect model fitted values vs true values 31

Figure 4.6 Box-plots of Standardized Residuals of Transition Model in each year 32

Figure 4.7 Transition model fitted values vs true values 33

Figure 4.8 Box-plots of Standardized Residuals of MERT Model in each year 35

Figure 4.9 MERT model fitted values vs true values 36

Figure 4.10 Mean square error (MSE) vs number of trees in the random forest part of MERF model 38

Figure 4.11 Box-plots of Standardized Residuals of MERF Model in each year 39

Figure 4.12 MERF model fitted values vs true values 39

Figure 4.13 Box-plots of Standardized residuals of RE-EM Tree model over time 41

Figure 4.14 RE-EM Tree model fitted values vs true values 42

LIST OF ABBREVIATIONS

MAE	Mean Absolute Error
MERF	Mixed Effect Random Forest
MERT	Mixed Effect Regression Tree
RE-EM Tree	Random Effect - Expectation Maximization Tree
RESD	Random Effect Standard Deviation
RMSE	Root Mean Squared Error

CHAPTER 1

INTRODUCTION

Suicide, which is defined as death caused by self harm with the intent to die, has been of interest to several different fields of science throughout the history. Majority of this interest comes from the fields of psychology and psychiatry. Both of these fields explore the underlying mental health conditions that can lead to suicide. Some of the other fields that research suicide from different perspectives include biology and philosophy. One of the examples of the biological research on suicide includes the research on the cellular suicide (Steller, 1995). While not fully adhering to the definition mentioned above, cellular suicide, also known as apoptosas, is the programmed self destruction of the cells.Topic of suicide has been a point of debate in the field of Philosophy for centuries. Those discussions mainly centered around the suicide as an ethical issue (Kelly & Dale, 2011). In this thesis we are exploring suicide from statistical perspective by modelling the factors that may contribute to it.

It is crucial to understand the reasons behind the suicide in order to prevent it. While individual reasons for suicide vary significantly, external factors can contribute to the mental health problems that can lead to suicide. Sociologist Emile Durkheim, in his famous book(Durkheim, 1951), pioneered the idea that the societal factors can contribute to the suicidal inclination rather than the individual reasons alone. Durkheim also in his book, defined suicide in different categories based on the reasons behind the suicide. Majority of these reasons include individual's reaction to societal changes and expectations. Since we are exploring suicide from a statistical perspective; we are interested in the factors that can be quantified by data. Rather than exploring individual data for suicide however, we are taking a broader approach in this thesis by researching the factors related to the suicide rate of provinces. Specifically we are an-

alyzing the socio-economic data of all 81 provinces of Turkey, such as GDP per capita and divorce rate. After deciding the related factors to use as predictors in our model, statistical modelling of suicide rates of all 81 provinces in Turkey are constructed to determine the significance of these factors in predicting the suicide rate.

In the model annual suicide rates of all 81 provinces of Turkey between years 2012 and 2019 are used as the dependent variable. Since multiple subjects are observed across several different time points, longitudinal data models are used for statistical modelling. While there are few studies about statistical analysis of suicide numbers in Turkey, none of them include the longitudinal data modelling of all 81 provinces. This study allows us to see the effects of yearly socio-economic changes on the suicide rate with respect to each provinces of Turkey. Discovering the related factors that have significant effect on the suicide rate would be important to address those issues and to help decrease suicide numbers in the future.

To determine the optimal model, well known longitudinal models such as the fixed effect, random effect and transition models were tested. Afterwards hybrid machine learning models were also tested for comparison. Those techniques include: mixed effects regression tree (MERT),mixed effects random forest (MERF) and random effects expectation maximization trees (RE-EM Trees). Main premise behind these approaches is to replace linear estimation of the fixed component in the longitudinal models with trees or tree-based machine learning algorithms. For all of the statistical modelling we used several different R packages which we will look into more detail in the later sections.

This thesis study is organized in five chapters. In the first chapter we gave a brief introduction to the topic and introduced the methods that we will use to explore the given topic. Second chapter will include the review of the literature for studies on longitudinal data modelling of suicide around the world. In the third chapter we will give more detailed explanations about the three mentioned longitudinal data models as well as the three machine learning methods. After these explanations in Chapter 3, we will give exploratory analysis as well as the results of each model. In the fifth and final chapter we will summarize the results we obtained from the models and give our final conclusions as well as investigate possible future studies.

CHAPTER 2

LITERATURE REVIEW

In this section we examined some of the studies in the literature that researched the factors related to the number of suicides and suicide rates around the world. Since the scope of this thesis is based on the provinces, the studies that are included here research the related factors based on provinces, states and countries rather than individuals. Majority of these models utilized the fixed effect model where the distribution of the dependent variable are either normal or negative binomial. The reason for that is while suicide numbers are discrete, suicide rate is continuous which necessitates a different distribution when modelling.

Minoiu and Andrés (2008) explored the effects of public spending on health and welfare as well as several other social and economic factors on the suicide rates of US states between 1982 and 1997. These factors include: State income, public health expenditure(PHE), public welfare expenditure(PWE), divorce rate, gini index, migrant population, unemployment rate, dummy variable to indicate whether the state is particularly mountainous, population density and average yearly number of days with sunshine. First four variables mentioned are used in one-period lagged form which is cited as the main reason for the preference of Generalized Method of Moments estimator(GMM) for modelling over the classic fixed effect models. Following results were obtained: Divorce rate, mountain state dummy and PWE were significant for overall, male and female suicide rate models. Among these variables; divorce rate and mountain state dummy had positive relation with suicide rate whereas PWE had negative relation. Population density was significant and negatively associated for male and overall suicide rate only.

Milner et al. (2010) investigated the effects of socio economic factors on the male and

female suicide rate of 35 countries between 1980 and 2006. A least-square dummy variable (LSDV) fixed-effect model was used in the analysis and following variables were included as the predictor variables in the model: Divorce rate, percentage of employed population in agriculture(EA), international migrant population, rural population, unemployment rate, female population in labor force, fertility rate, public health expenditure(PHE), number of elder population(people over 65 years old) and GDP per capita. Among these variables; unemployment rate, elder population, PHE and female labour force were significant for both gender suicide rate models. Unemployment and elder population had positive relation whereas female labour force and PHE had negative relation with respective suicide rates. The only other significant variable in either of the models was fertility rate which was only significant and had negative relation in the male suicide rate model.

Yamamura (2010) analyzed the effects of socioeconomic factors on the suicide rate of Japanese prefectures. Specifically, the research focused on the male, female and overall suicide rates of 47 prefecture of Japan consisting of years between 1988 and 2001. Factors that were included were: Income growth rate, income per capita, unemployment rate, population turnover within prefecture, immigrants from other prefectures, number of public baths, number of marriages, divorce rate, population, number of members in households, alcohol consumption and birth rate. Three fixed effect longitudinal data models were used for total, male and female suicide rates with the following results: Income growth, alcohol consumption, population, population turnover, divorce rate, marriages were significant for all three models where the growth rate and marriages had decreasing effect on suicide rate while the other variables had increasing effect. Immigrant variable was significant and positively associated with overall and female suicide rate models only. Public baths and birth rate were significant only for the female suicide rate model where public baths had decreasing while birth rate had increasing effect on the suicide rate.

Barth et al. (2011) investigated the socioeconomic factors that could affect the suicide rate of 18 countries in a 25 year period between 1983 and 2007. Those countries were Austria, Belgium, Denmark, France, Finland, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, the United Kingdom and the United States. Fixed effects and transition longitudinal data

models were used with three separate dependent variables which include overall, male and female suicide rates.Independent variables that were used in the models included: GDP per capita, unemployment rate, divorce rate and labour force ratio. In the fixed effect models: GDP was significant only for the overall suicide rate model and had negative relation with the dependent variable. Divorce rate was significant only for the female suicide rate model and had positive relation with the dependent variable. In the transition models: GDP and labour force ratio were significant for the overall suicide rate model and had negative and positive effect on the dependent variable respectively.

Ross et al. (2012) analyzed the effects of mental health expenditures on the suicide rate of US states between 1997 and 2005. For the analysis; suicide rate of two sexes and two age groups have been selected as dependent variable. These four variables were suicide rates of; all male, all female, males between 25-64 and females between 25-64. Predictor variables that were used income, mental health expenditures(MHE), public welfare expenditure(PWE), public health expenditures(PHE), unemployment rate, population density, divorce rate, migration rate, non white(Caucasian) population, dummy variable to indicate whether the state is particularly mountainous called mountain dummy and number of sunny days. Fixed effects and transition longitudinal data models were used with four different dependent variables. Following results are obtained from the fixed effect models: PHE was significant and had negative relation in the both male suicide rate models. MHE was significant and had negative relation in the both female suicide rate models. From the transition models following results were obtained: mountain state dummy was significant and had positive relation with dependent variable in all four models. Population density was significant and had positive relation with all male suicide rate model only.

Kõlves et al. (2013) explored the effects of socioeconomic factors on the male and female suicide rates of 13 Eastern European countries between 1990 and 2008. Main focus of this paper was to explore the effects of the collapse of Soviet Union on the social and economic factors of these countries and how those factors collectively affected the suicide rate of both genders. Fixed effects longitudinal data model was used where the dependent variables were the male and female suicide rates respectively. Predictor variables include: unemployment rate, GDP per capita, alcohol consump-

tion, crude birth rate, general practitioners(GPs) per 100, 000 people, divorce rate, female labour participation and gini index. From the two models following results are obtained: Unemployment rate and GDP per capita were significant in both models but had positive and negative relation with dependent variable respectively. GPs and birth rate were significant for only male suicide rate model and both of them had negative relation. Gini index and alcohol consumption were significant for only female suicide rate model but had positive and negative relations respectively.

Breuer (2014) explored the relation between economic factors and suicide rate of 275 regions from 29 European countries over a 12 year period between 1999 and 2010. These countries include 25 of the EU-27 countries with the exception of Croatia and Denmark, as well as Iceland, Norway, Switzerland and United Kingdom. This study explores suicide rate with respect to both, gender and age where alongside overall suicide rate, male and female suicide rate of population over and under 65 years old are observed separately. Predictor variables used in the models include: unemployment rate, fertility rates, gender and age specific life expectancy, heating degree days divided by 365 which is used as a proxy for weather, GDP per capita and annual growth rate of real Gross Value Added(GVA) which is used as a proxy for economic growth. Several models were constructed for different age groups and genders. From those models following results are obtained: Life expectancy was significantly and negatively associated with suicide rate for all genders and age groups. Unemployment was significant for overall and male suicide rates only and positively associated in both cases. Economic growth was also significant for only overall and male suicide rates but negatively associated in both cases. GDP per capita was significant only for overall suicide rate and had negative effect on it. When the same models were constructed for people over and under 65 years old following changes occurred: For the people under 65, previously insignificant weather variable became significantly and positively associated for both male and overall suicide rate. For people over 65, the life expectancy variable was significantly and negatively associated for male and female suicide rates whereas all the other variables were no longer significant for any of the groups.

Machado et al. (2015) investigated the effects of socio-economic factors on the suicide rate of both genders in 5,507 municipalities of Brazil in a 12 year period between

2000 and 2011. Primary focus among these factors was on the income inequality, which was quantified by the Gini index of each municipality. Other factors were education levels of individuals, income levels, urbanization rate, divorce rate and religious affiliations which were grouped into three different categories. Negative binomial models with fixed effect specifications were used for the analysis which were decided by conducting a Hausman test (Glen, 2020). From the results of the models, it is found that Gini index, percentage of under educated people and two of the religious affiliations have positive and significant relationship with the overall suicide rate. On the other hand; income, urbanization rate, divorce rate, mean amount of resident and one of the religious affiliations has negative relationship with the overall suicide rate. When we look at the effects of these factors on the suicide rate of each gender, we saw that most of the results didn't change. Most notable difference was on the divorce rate where it was positively associated for men and negatively associated for women.

Ferreira et al. (2019) explored the effects of economic factors on suicide rate of the EU countries over a 24 year period between 1990 and 2013. There were multiple models constructed using the suicide rate of two genders and two age groups as the dependent variables. Independent variables in these models were public expenditures on health(PEH), public expenditures on social welfare(PESW), unemployment, GDP per capita, fertility rate, divorce rate and consumption of alcohol per capita. From the first three models that modeled the male, female and overall suicide rate of all age groups, following results are obtained: PESW, divorce rate and unemployment were significant in all three models but PESW had a negative relation with the suicide rate while other two variables had positive relation. GDP per capita was significant for only the male suicide rate model and had negatively association. In the next four models male and female suicide rates of people over and under 65 years old were used as dependent variables. Results were overall similar with few differences: PESW and unemployment rate were significant in all four models and had negative and positive associations respectively. GDP per capita was significant in the young male and older female suicide rate models with negative and positive relation respectively. Divorce rate was positively related in all four models but was only significant for three of the four models which excluded the older male suicide rate model. Previously insignif-

icant in all three models, fertility rate was significant and had negative relation with older female suicide rate.

Emamgholipour et al. (2021) examined social and economic factors related to the suicide numbers of 30 provinces of Iran between 2001 and 2016. Factors that were used were: Divorce, literacy, unemployment and urbanization rates, per capita income and industrialization index. As a result of Hausman test, fixed effect model was chosen when constructing the model. In the model, industrialization index was significant only for male suicide number whereas all the other covariates were significant for both genders. Urbanization rate was positively associated with female suicide numbers and negatively associated with male suicide numbers. Associations of all the other variable were in the same direction for both genders where; unemployment rate and divorce rate were positively related whereas household income and literacy rate were negatively related.

After examining different studies which focused on span of different countries and provinces' suicide rate, this thesis aims to add to them by modelling all 81 provinces of Turkey. Methodology of this thesis expands the standard longitudinal data modelling methods and adds hybrid machine learning model which differs from the methods used in literature. In the next section, methodology of this thesis will be explained in detail.

CHAPTER 3

METHODOLOGY

In this section, techniques that were used to model the suicide rate of provinces of Turkey as well as the comparison metrics used to compare the models are explained in detail. This section will be divided into three sections which will explain the three standard longitudinal data models and three hybrid data models as well as the comparison metrics for the models.

3.1 Longitudinal Data Models

Longitudinal data also known as the panel data, refers to the dataset consisting of multiple subjects that are tracked across multiple time points. It encompasses the attributes of two types of data in one, which are cross sectional data where multiple subjects are shown in single time point and time series data where a single subject is tracked across multiple time points. There are several different areas where the longitudinal data is commonly used including; econometrics,epidemiology as well as health and social sciences.There are three well known longitudinal data models that will be used in this thesis which include the random effects models,fixed effects models and transition models.

3.1.1 Random Effects Models

Random effects model allows model parameters to vary from one subject to another which results in heterogeneity among individuals. The main difference between fixed effects model and random effects model is that in random effects model, the omit-

ted variables are uncorrelated with the independent variables whereas the omitted variables in fixed effects model are correlated with the independent variables. Representing equation for the random effects models is the following:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} ... + \beta_n X_{kit} + b_{0i} + b_{1i} Z_{1it} ... + b_{li} Z_{lit} + \varepsilon_{it}$$

where

Y_{it} represents the dependent variable where i=1,2,...,n indicates the subjects and t=1,2...n_i indicates the time,

X_{it} represents the independent variables and β_{it} represents the coefficients of the independent variables,

Z_{it} is the subset of X_{it} while b_{i0} and b_{im} represent the intercept and random coefficients where $\underline{b_i} = (b_{1it}, b_{2it} ... b_{lit}) \sim N(0, D)$,

ε_{it} represents the error term where $\varepsilon_{it} \sim N(0, R_i), R_i = \sigma^2 I_n$.

3.1.2 Fixed Effects Models

Fixed effects model is the most widely used model in the longitudinal data analysis. The main premise behind these type of models is modelling the dependent variable as a function of independent variables, while taking the within-subject correlation into account. When choosing this type of model over the others, the goal is to compare groups or subgroups rather than individuals. Fixed effects model are represented by the following equation:

$$Y_{it} = \beta_0 + \beta_1 X_{it,1} + \beta_2 X_{it,2} + ... + \beta_{p-1} X_{it,p-1} + \varepsilon_{it}$$

where

Y_{it} represents the dependent variable where i indicates the subjects and t indicates the time,

X represents the independent variables and β represents the coefficients of the independent variables,

ε represents the error term with the covariance matrix of R_i which can have several different covariance structures including; autoregressive(AR), exchangeable, toeplitz and unstructured structures.

3.1.3 Transition Models

Main premise behind the transition models is that the response at time t is modelled depending on the responses at the previous time points and covariates at the current time points. The aim of this approach is to learn from the previous observations. We can also consider this model as a special case of fixed effect models in which it includes the lag of the response, as well as the covariates, to the right hand side of the model. If the model for the conditional mean is correctly specified, we can treat repeated transitions for the subject as independent events and use standard statistical methods It is important to remember however that unlike the previous models, data needs to be balanced. Representing the formula for the model can be found below:

$$Y_{it} = \beta_0 + \beta_1 X_{it,1} + \beta_2 X_{it,2} ... + \beta_{p-1} X_{it,p-1} + \alpha_1 Y_{i,t-1} ... + \alpha_k Y_{i,t-k} + \varepsilon_{it}$$

where

Y_{it} represents the dependent variable where i indicates the subjects and t indicates the time,

X_{it} represents the independent variables and β_i represents the coefficients of the independent variables,

$Y_{i,t-k}$ represents dependent variables with k lag and α_k represents the coefficients of the lagged dependent variables,

ε_{it} represents the error term with the variance structure of an identity matrix.

3.2 Hybrid Machine Learning Models

In this thesis we made use of three hybrid machine learning models that were introduced as a way to integrate tree regression methods into longitudinal data. One of the key advantages of this integration is to bypass the assumptions that were required by the standard longitudinal data models for the datasets. While most of the methods that are utilized have similarities, details and differences of these methods will be explained in this section.

3.2.1 Mixed Effect Regression Trees (MERT)

Hajjem et al. (2011) introduced the first hybrid method that we use in this thesis.The main idea behind this approach is to calculate the fixed effect component using standard regression trees and calculate random effect component using linear mixed modelling(LMM) for each node of the tree. A Classification and Regression Tree (CART) algorithm was used when constructing the tree. Following equation can be used to represent this method:

$$y_{it} = f(X_i) + b_{1i}Z_{1it}... + b_{li}Z_{lit} + \varepsilon_{it}$$

where;

$f(X_{it})$ is the regression tree model that will be used estimate the fixed effects of the model,

y_{it} represents the dependent variable where i indicates the subjects and t indicates the time,

X_{it} represents the independent variables,

Z_{it} is the subset of X_{it} while b_{li} represent the random coefficients where $b_{li} \sim N(0, D)$.

ε_{it} represents the error term with the covariance matrix of $R_i = \sigma^2 I_n$.

The steps of the algorithm for the regression tree $f(X_i)$ where r=1,2,... represents the iteration number, can be described as following:

1. Let r=1, $\hat{b_i} = 0$, $\hat{D}_0 = I_l$ and $\hat{\sigma}_0^2 = 1$

2. Let r=r+1, update $y^*_{i(r)}$, $\hat{f}(X_i)_{(r)}$ and $\hat{b}_{i(r)} = 0$

   - $y^*_{i(r)} = y_i - Z_i\hat{b}_{i(r-1)}$ i=1,2,...,n
   - Let $\hat{f}(X_i)_{(r)}$ an estimate of $f(X_i)$ obtained from a standard tree algorithm with $y^*_{i(r)}$ as responses and X_i where i=1,2,...,n as covariates. Note that the tree is built as usual using all n individuals as inputs along with their covariate vectors.
   - $\hat{b}_{i(r)} = \hat{D}_{(r-1)}Z_i^T\hat{V}^{-1}_{i(r-1)}(y_i - \hat{f}(X_i)_{(r)})$, i=1,2,...,n where $\hat{V}^{-1}_{i(r-1)} = Z_i\hat{D}_{(r-1)}Z_i^T + \hat{\sigma}^2_{(r-1)}I_{n_i}, i = 1, 2, ..., n$

3. Update $\hat{\sigma}^2_{(r)}$ and $\hat{D}_{(r)}$ using:

$$\hat{\sigma}^2_{(r)} = N^{-1}\sum_{i=1}^{n}\{\hat{\epsilon}^T_{i(r)}\hat{\epsilon}_{i(r)} + \hat{\sigma}_{(r-1)}[n_i - \hat{\sigma}_{(r-1)}trace(\hat{V}^{-1}_{i(r-1)})]\}$$

$$\hat{D}_{(r)} = n^{-1}\sum_{i=1}^{n}\{\hat{b}^T_{i(r)}\hat{b}_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)}Z_i^T\hat{V}^{-1}_{i(r-1)}Z_i\hat{D}_{(r-1)}]\}$$

   where $N = \sum_{i=1}^{n} n_i$.

4. Repeat steps 2 and 3 until convergence.

To describe the steps of the algorithm, it starts at step 1 with default values for random effect values($\hat{b}_i$), random effect variance ($\hat{\sigma}^2$), and random effect covariance structure ($\hat{D}$). At step 2, it first calculates the fixed step of the response variable, y_i, which is the response variable from which current available value of the random step is removed. Second, the fixed component $\hat{f}(X_i)$ is estimated using a standard tree algorithm with y_i as responses and X_i as covariates. Third, $\hat{b}_i$ is updated. At step 3, the variance components ($\hat{\sigma}^2$) and ($\hat{D}$) are updated based on the residuals after the estimated fixed component $\hat{f}(X_i)$ is removed from the raw data y_i . Iterations continue by repeating steps 2 and 3 until convergence which is monitored by computing, at each iteration, the following generalized log-likelihood (GLL) criterion:

$$GLL(f, b_i|y) = \sum_i^n\{[y_i - f(X_i) - Z_ib_i]^TR_i^{-1}[y_i - f(X_i) - Z_ib_i] + b_i^TD^{-1}b_i + log|D| + log|R_i|\} \quad (3.1)$$

3.2.2 Mixed Effect Random Forest (MERF)

After introducing MERT, Hajjem et al. (2012) later expanded that method by replacing regression trees in each iteration with random forests. This new approach has been named mixed effect random forest(MERF).

$$y_{it} = f(X_i) + b_{1i}Z_{1it}... + b_{li}Z_{lit} + \varepsilon_{it}$$

where;

$f(X_i)$ is the random forest model that will be used estimate the fixed effects of the model,

y_{it} represents the dependent variable where i indicates the subjects and t indicates the time,

X_i represents the independent variables,

Z_{it} is the subset of X_{it} while b_{li} represent random coefficients where $b_{li} \sim N(0, D)$.

ε_{it} represents the error term with the covariance matrix of $R_i = \sigma^2 I_n$.

The steps of the algorithm for the regression tree $f(X_i)$ where r=1,2,.. represents the iteration number, can be described as following:

1. Let r=1, $\hat{b_i} = 0$, $\hat{D}_0 = I_l$ and $\hat{\sigma}_0^2 = 1$

2. Let r=r+1, update $y^*_{i(r)}$, $\hat{f}(X_i)_n$ and $\hat{b}_{i(r)}$

   - $y^*_{i(r)} = y_i - Z_i\hat{b}_{i(r-1)}$ i=1,2....,n
   - Build a random forest using a standard RF algorithm with $y^*_{i(r)}$ as the training set responses and x_i as the corresponding training set of covariates $i = 1, 2...., n$ and $t = 1, 2, ..., n_i$. The bootstrap sampling method is used to build the random forests.
   - Estimate of $\hat{f}(x_i)_{(r)}$ is obtained using random forests.

3. Update $\hat{\sigma}^2_{(r)}$ and $\hat{D}_{(r)}$ using:

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \{\hat{\epsilon}^T_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}_{(r-1)}[n_i - \hat{\sigma}_{(r-1)} trace(\hat{V}^{-1}_{j(r-1)})]\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \{\hat{b}^T_{i(r)} \hat{b}_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}]\}$$

4. Repeat steps 2 and 3 until convergence.

To describe the steps of the algorithm, it starts with step 1 with default values for random effect values($\hat{b}_i$), random effect variance ($\hat{\sigma}^2$), and random effect covariance structure ($\hat{D}$). At step 2, it first calculates the fixed step of the response variable, y_i, which is the response variable from which current available value of the random step is removed. Second, the algorithm uses bootstrap sampling method to build a random forest. At step 3, variance components $\hat{\sigma}^2$ and $\hat{D}$ are updated based on the updated estimates of the residuals. The algorithm keeps iterating by repeating steps 2 and 3 until convergence which is monitored by computing, at each iteration, the following generalized log-likelihood (GLL) criterion:

$$GLL(f, b_i|y) = \sum_i^n \{[y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] + b_i^T D^{-1} b_i + log|D| + log|R_i|\} \quad (3.2)$$

3.2.3 Regression Expectation Maximization Trees (RE-EM Trees)

Sela and Simonoff (2011) proposed another tree method that can be used to analyze clustered and longitudinal data that is named random effects expectation maximization trees (RE-EM Trees). Although this method was proposed independently from the previous two methods, there are similarities between these methods. The representing equation is the same as the equation which is same as the equation used for MERT model can be described as following:

$$y_{it} = f(X_i) + b_{0i} + b_{1i} Z_{1it} ... + b_{li} Z_{lit} + \varepsilon_{it}$$

$f(X_i)$ is the unknown regression tree model that will be used estimate the fixed effects of the model,

Y_{it} represents the dependent variable where j indicates the subjects and t indicates the time,

Z_{it} is the subset of X_{it} while b_{0i} and b_{li} represent the random coefficients where $b_{li} \sim N(0, D)$,

ε_{it} represents the error term with the covariance matrix of R_i.

Similar to MERT, fixed effect and random effect equations are separated and calculated using regression tree and LMM methods respectively. The algorithm can be summarised as following:

1. Random effects are to be initialised as $\hat{b}_i = 0$.
2. Steps 2a and 2b are to be iterated until the estimated random effects, $\hat{b}_i$, converge.
   (a) Regression tree approximating f is to be estimated , based on the target variable, $y_{it} - Z_{it}\hat{b}_i$ and attributes, $x_{it} = (x_{it1}, x_{it2}...x_{it(p-1)})$ for i=1,2,..,n and t=$1, 2, ..., n_i$. This regression tree is to be used to create a set of indicator variables, $N(x_{it} \in g_p)$, where g_p ranges over all of the terminal nodes in the tree.
   (a) Linear mixed effects model, $y_{it} = Z_{it}b_i + N(x_{it} \in g_p)\mu_p + \epsilon_{it}$ is to be fitted and $\hat{b}_i$ is to be extracted from the estimated model.
3. Predicted response at each terminal node of the tree is to be replaced with the estimated population level predicted response $\hat{\mu}_p$ from the linear mixed effects model fit in 2b.

3.3 Comparison Metrics

After constructing all of the models, we are required to compare them in order to determine the most accurate method for modelling the suicide rate. One of the two metrics that we used to compare the final models of each method is the Root Mean Square Error(RMSE) which is calculated by equation 3.3.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{3.3}$$

Here y_i and $\hat{y}_i$ represent the observed and predicted values respectively where $i = 1, 2, ..., n$. This metric quantifies the accuracy by looking at the square root of the average of the squared difference between the observed and predicted values.

Second metric that was used is the mean absolute error(MAE) which is calculated by equation 3.4.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{3.4}$$

Here y_i and $\hat{y}_i$ represent the observed and predicted values respectively where $i = 1, 2, ..., n$. This metric quantifies the accuracy using the average of the absolute value of differences between the observed and predicted values. Since RMSE takes the square of the differences before square rooting, larger differences between the observed and calculated values have much bigger influence on the calculated error than the MAE (Willmott & Matsuura, 2005).

CHAPTER 4

RESULTS

In this chapter, results of exploratory analysis as well as the results of all of the constructed models are explained in detail.

Data from the all 81 provinces of Turkey between 2012 and 2019. All of the data is compiled from Turkish Statistical Institute (TURKSTAT) . The dependent variable we used was the annual suicide rate of each province in the given timeline which was calculated by dividing the suicide number of the given year by the population of the country in the middle of the year. The predictor variables used in the model include the following: Fertility rate, elderly dependency ratio, categorical variable to indicate whether the given province is considered metropolitan or not, portion of hospitals to population of province, portion of people who successfully earned a masters degree or PhD to population of province, portion of people who are illiterate to population of province, portion of unemployed people to population of province, portion of divorces to population of province and GDP per capita of each province. These variables are chosen because they were used in the similar models from the literature. While all the variables are used in the initial models, additional models are also constructed using only the significant variables obtained from the previous models. R programming language was used for all of the models and all of the plots. R codes that were used for this thesis can be found in the Appendix C.

4.1 Exploratory Analysis

For this analysis, we have obtained data from the all 81 provinces of Turkey between 2012 and 2019. The dependent variable we used was the annual suicide rate of each

province in the given timeline which is calculated by dividing the suicide numbers of the province at the end of the year to population at the middle of the year and multiplying the result by 100,000. The predictor variables used in the model include the following:

- Fertility Rate: Calculated by dividing the number of births in the province at the end of the year to population of fertile(women aged between 15-44) in the middle of the year and multiplying the result by 1000.
- Elderly Dependency Ratio: Calculated by dividing the elderly population(people aged over 65+) at the end of the year in the province to working age(people aged between 15-64) population at the end of the year in the province and multiplying the result by 100.
- Metropolitan Variable: Categorical variable indicating whether the given province is Metropolitan or not. Currently there are 30 metropolitan provinces which were indicated by "1" by Metropolitan variable.
- Portion of Hospitals: Number of hospitals in the province divided by the population of the province at the end of the year and multiplying the result by 100,000.
- Portion of Higher Ed.: Calculated by dividing number of people who successfully earned a masters degree or PhD in the province to population of province then multiplying the result by 1000.
- Portion of Dropouts: Calculated by dividing number of people who are illiterate in the province to population of province then multiplying the result by 100.
- Portion of Unemployed: Calculated by dividing the population of unemployed people to the total population and multiplying it by 100
- Portion of Divorces: Calculated by dividing the population of unemployed people to the total population and multiplying it by 100.
- GDP per Capita: Gross Domestic Product of the province divided by its population.

These variables are chosen because they were used in the similar models from the literature. While all the variables are used in the initial models, additional models are also constructed using only the significant variables and variables that have higher variable importance obtained from the initial models.

Table 4.1: Descriptive Statistics

Variable	N	Mean	Median	Std. Dev	Min	1st Quar.	3rd Quar.	Max
Suicide Rate	648	4.366	4.204	1.6	0	3.298	5.249	11.631
Fertility Rate	648	2.122	1.860	0.651	1.324	1.694	2.333	4.574
Elderly Dependency Ratio	648	14.348	14.648	4.794	4.209	10.923	17.374	29.163
Metropolitan:No	409							
Metropolitan:Yes	239							
Portion of Hospitals	648	2.403	2.2047	0.991	0.765	1.747	2.958	7.3
Portion of Higher Ed.	648	8.017	7.218	4.3	1.474	5.006	10.176	32.927
Portion of Dropouts	648	3.982	3.6421	1.861	0.865	2.382	5.378	9.025
Portion of Unemp.	648	3.991	3.670	1.54	1.693	3.038	4.509	15.735
Portion of Divorces	648	1.373	1.4269	0.629	0.112	0.98	1.83	2.954
GDP per Capita	648	8.37	7.866	3.045	2.946	6.26	9.827	20.883

From Table 4.1, we can see that minimum suicide rate is 0, which is obtained from province of Bayburt. The reason for the 0 suicide rate is because this province had 0 suicides in both 2017 and 2019. Highest suicide rate of 11.631 is observed in the province of Tunceli. While this number may seem high, it is important to remember that Tunceli is one of the least populated provinces in the country, and therefore even a small number of suicides creates high suicide rates compared to the more populated provinces. Since the mean and median are close for the most of the variables, it can be interpreted that the majority of the variables shows symmetric distribution. Since there are 30 metropolitan cities in Turkey, we would expect the number of metropolitan indicator variables to be 240 in an 8 year period. However, since the province of Ordu only became metropolitan in 2013, and since our data contains the period between 2012 and 2019, the metropolitan variable is split as 409 to 239.

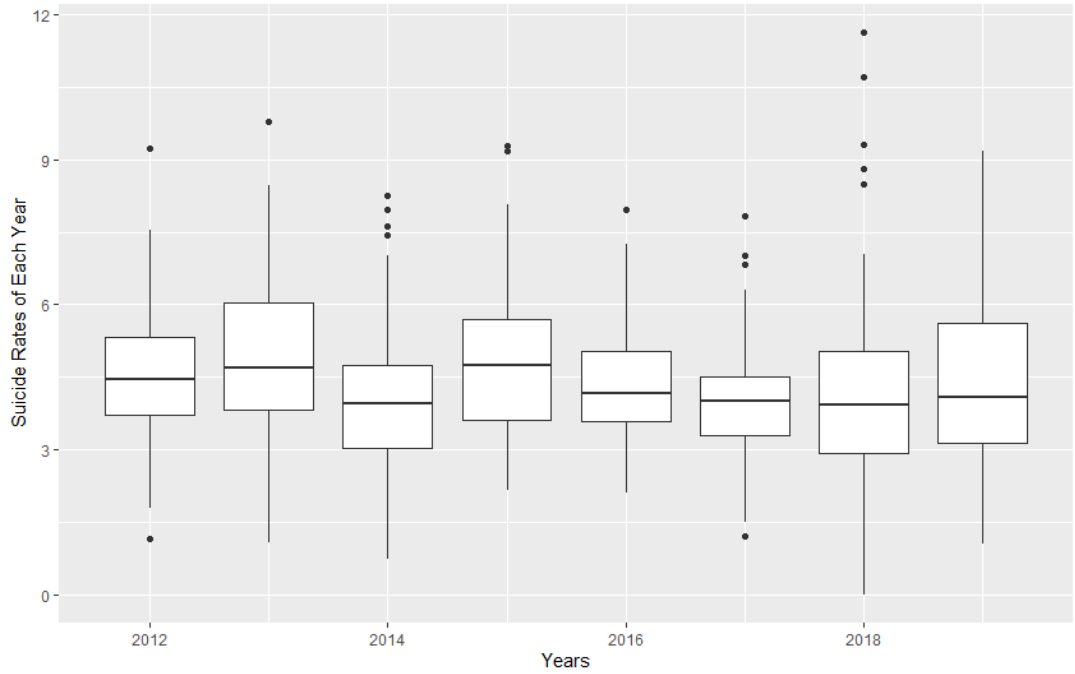

Figure 4.1: Boxplots of suicide rates of each year

From Figure 4.1, we can see the boxplots of suicide rates of all provinces in each year. From the shapes of each boxplots we can infer that the majority of them don't have normal distribution. In order to confirm this, Shapiro-Wilk test has been performed on the each year's suicide rate.

Table 4.2: Shapiro Wilk Normality Test Results of each year

Year	W Statistic	p-value
2012	0.971	0.067
2013	0.973	0.082
2014	0.942	0.001
2015	0.983	0.369
2016	0.939	0.001
2017	0.956	0.007
2018	0.958	0.009
2019	0.985	0.478

Normality assumption is required for the panel data models that we will apply, so it is important to check each year of the dependent variable for normality. From the

histogram plots in Figure 4.1 we can see that the both 2015 and 2019 have normal distribution which is confirmed with normality tests. However majority of the years don't satisfy the normality condition, therefore it is required to apply transformations to the dependent variable.

For transformation, number of different methods are tried, but unfortunately none of them satisfy the normality conditions for all of the years. The satisfactory method that worked for the majority of the years can be found in Equation 4.1.

$$y(\lambda) = \sqrt{\lambda + 1} \tag{4.1}$$

After the transformation Shapiro Wilks normality test performed again, and results are given in Table 4.3.

Table 4.3: Shapiro Wilk Normality Test Results after transformation

Year	W Statistic	p-value
2012	0.988	0.668
2013	0.984	0.421
2014	0.974	0.0952
2015	0.988	0.696
2016	0.977	0.167
2017	0.957	0.008
2018	0.982	0.313
2019	0.979	0.223

From Table 4.3 it can be seen that all years except 2017 satisfy the normality conditions. As mentioned while this method doesn't transform all of the years it can still be considered satisfactory.

After transformation, the pairwise correlations between variables which are shown in Table A.1 in Appendix A, are observed to make an initial assessment about the relations between the variables. From these observations, we can see that divorce variable(Portion of Divorce) has the highest correlation with the suicide rate followed by portion of hospitals and elderly dependency ratio. It is important to note however,

these correlations are rather small with the highest being around 0.21. Other high correlations that can be observed are the illiterate population that has high negative correlation with divorce rate and high positive correlation with fertility rate. This indicates that the under educated population tend to have more children while also having fewer divorces.

Table 4.4: Temporal Correlations of Suicide Rate

	2012	2013	2014	2015	2016	2017	2018	2019
2012	1	0.277	0.304	0.334	0.349	0.286	0.319	0.361
2013	0.277	1	0.361	0.604	0.623	0.581	0.455	0.569
2014	0.303	0.361	1	0.542	0.525	0.415	0.344	0.381
2015	0.334	0.604	0.542	1	0.541	0.469	0.443	0.435
2016	0.349	0.623	0.525	0.541	1	0.615	0.532	0.646
2017	0.286	0.581	0.415	0.469	0.615	1	0.562	0.583
2018	0.319	0.455	0.344	0.442	0.532	0.562	1	0.498
2019	0.361	0.569	0.381	0.435	0.646	0.583	0.498	1

Table 4.4 shows the correlation structure of the suicide rate across all of the given years. From Table it can be observed that, there is no specific correlation structure, which will be important when constructing the models.

4.2 Standard Longitudinal Data Models

In the first part of the modelling process, we used three of the most commonly used longitudinal data models which included the fixed effect(marginal) model, random effect model and transition model. For each method we started by constructing a model using all of the variables then refine the model until only significant variables remained.

4.2.1 Random Effect Models

We started the modelling with random effect model. Initial model was constructed using all of the covariates. After constructing the model, variance inflation factors(VIF) of the variables are observed in order to detect any possible multicollinearity in the model.

Table 4.5: VIF values of the each variable in the random effect model

Variables	VIF
Year	7.74
GDP per Capita	3.02
Portion of Hospitals	1.72
Portion of Divorces	2.32
Elderly Dependency Ratio	3.09
Metropolitan	1.51
Portion of Unemployment	1.45
Portion of Highly Educated	4.43
Portion of Illiterate	4.65
Fertility Rate	3.86

From Table 4.5 we can see that all of the VIF values are quite small and therefore we can confirm that there is no multicollinearity problem in the model. After constructing the model, it has been confirmed that only the divorce,hospital and GDP variables are significant at the $\%95$ confidence level. Then the next model was constructed using the significant variables.

Table 4.6: Results of the Random Effect Model with significant variables

Coefficients	Estimate	Std. Err.	P value
Intercept	1.941	0.076	$2x10^{-16}$
Portion of divorces	0.139	0.039	0.001
Portion of hospitals	0.087	0.024	0.001
GDP per capita	-0.006	0.007	0.378

Second model indicates that only hospital and divorce variables are significant and have increasing effect on suicide rate, whereas GDP per capita is not significant and therefore excluded in the next model.

Table 4.7: Results of the Final Random Effect Model

Coefficients	Estimate	Std. Err.	P value
Intercept	1.908	0.084	$2\text{x}10^{-16}$
Portion of divorces	0.127	0.037	0.001
Portion of hospitals	0.085	0.024	0.001

From Table 4.7 it can be seen that both variables are highly significant and therefore this model can be considered to be the final.

$$\hat{TSR}_{it} = 1.908 + 0.127PoD_{it} + 0.085PoH_{it} + \hat{b}_i \tag{4.2}$$

where TSR refers to the transformed suicide rate which is calculated using Equation 4.2. PoD and PoH represent suicide rate,portion of divorces and portion of hospitals respectively. From Equation we can infer that both divorce and hospital proportions have increasing effect on the suicide rate. In the event where both covariates are zero, we can see that suicide rate would be positive with the value around 1.91. This indicates that at any time and in every province without any hospitals and divorces, suicide rate would still be positive and significant.

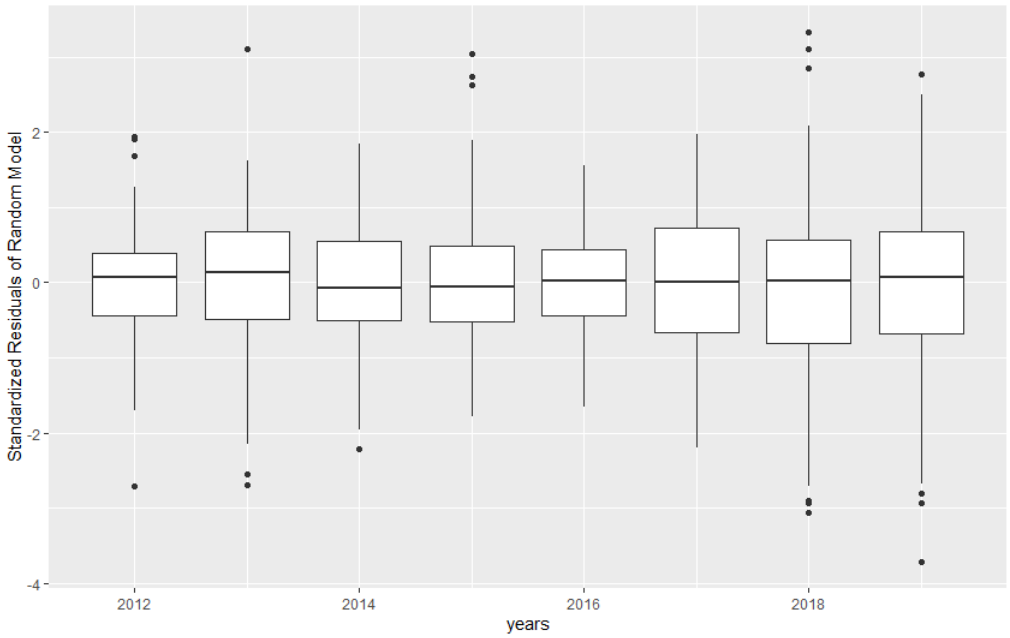

Figure 4.2: Box-plots of Standardized Residuals of Random Effect Model in each year

From the Figure 4.2 it can be seen that while there are few outliers in the several years, there is no particular pattern that would indicate a bad fit for our model. From Table B.1 in the Appendix B we can see that residuals from all but the year 2016 are normally distributed.

Figure 4.3: Random effect model fitted values vs true values

From the figure 4.3 it can be seen that the scatterplot of the response shows a decent fit with small discrepancies and shows increasing trend. Root mean squared error(RMSE) is calculated as 1.097.

4.2.2 Fixed Effect Model

Second model that we used from the standard longitudinal data models was the fixed effect model. From the exploratory analysis we determined that the correlation structure of the dependent variable doesn't resemble a predefined structure, therefore when constructing the model unstructured correlation structure was used as a first choice. Few other correlation structures such as AR(1) and exchangeable correlation structures were also used for comparison purposes. Similar to the random effect model, all of the covariates were included in the initial model. From the results of the model we found that only; GDP per capita,portion of hospitals and portion of divorces variables were significant. Using only these variables we reconstructed the model with different correlation structure. Results for these models are given in Tables 4.8-4.10 .

Table 4.8: Results of the Fixed Effect Model for unstructured correlation structure

Coefficients	Estimate	Std. Err.	P value
Intercept	1.969	0.079	$2x10^{-16}$
Portion of Divorce	0.084	0.027	0.001
Portion of hospitals	0.138	0.039	0.001
GDP per Capita	-0.007	0.005	0.177

Table 4.9: Results of the Fixed Effect Model for AR(1) Correlation Structure

Coefficients	Estimate	Std. Err.	P value
Intercept	1.965	0.095	$2x10^{-16}$
Portion of Divorce	0.093	0.033	0.004
Portion of hospitals	0.143	0.045	0.001
GDP per Capita	-0.011	0.007	0.157

Table 4.10: Results of the Fixed Effect Model for Exchangeable Correlation Structure

Coefficients	Estimate	Std. Err.	P value
Intercept	1.944	0.091	$2x10^{-16}$
Portion of Divorce	0.0861	0.031	0.006
Portion of hospitals	0.139	0.039	0.001
GDP per Capita	-0.006	0.006	0.278

Comparing all of the results with different correlation structures, we can see that there isn't a significant difference between standard errors or the coefficient estimates. We also used quasi information criterion (QIC) to compare the three models which gave 665, 670 and 671 for unstructured, AR(1) and exchangeable correlation structures respectively. From these results it can be seen that the unstructured correlation structures performs slightly better than the others and thus it is chosen for the fixed effect model.

Table 4.11: Results of the Final Fixed Effect Model

Coefficients	Estimate	Std. Err.	P value
Intercept	1.928	0.076	$2x10^{-16}$
Portion of divorces	0.126	0.036	0.001
Portion of hospitals	0.081	0.028	0.003

Similar to the random effect model, GDP variable while significant in the model with all of the covariates, is no longer significant in the newer models. Therefore, in the final constructed model only the portion of divorce and portion of hospitals are included. From Table 4.11 we can confirm that both of the remaining covariates remain highly significant in the final model.

$$\hat{TSR}_{it} = 1.928 + 0.126PoD_{it} + 0.081PoH_{it} \tag{4.3}$$

From Equation 4.3 we can see that where both the divorce and hospital variable have an increasing effect on the suicide rate. From the positive intercept value, we can infer

that in an event where both covariates are zero, we can see that suicide rate would be positive with the value around 1.93.

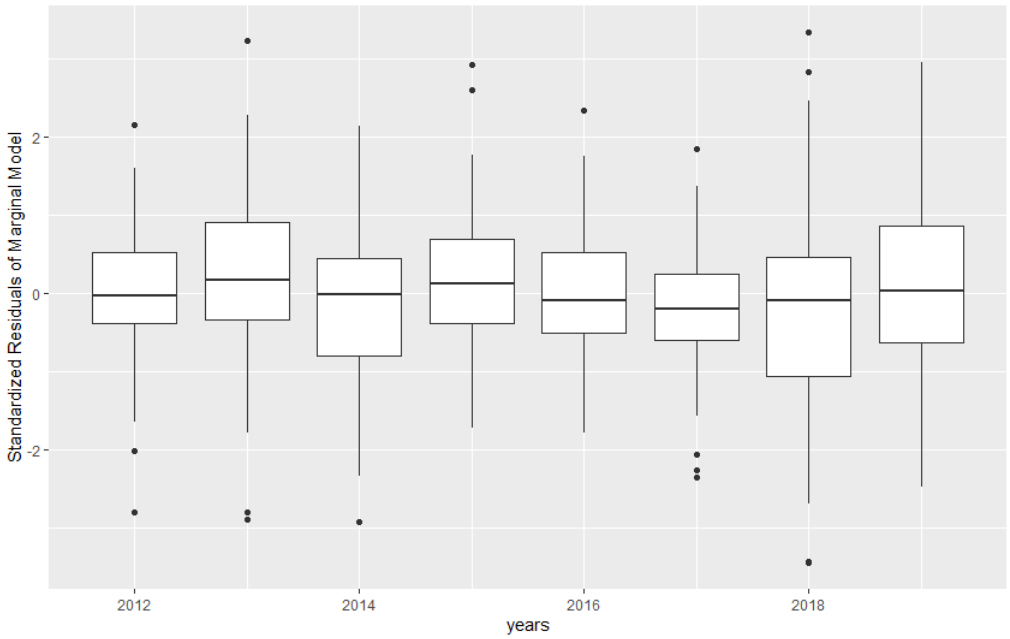

Figure 4.4: Box-plots of Standardized Residuals of Fixed Effects Model in each year

After parameter estimation, the residuals of the model are examined over time for possible patterns and outliers. From Figure 4.4 it can be seen that the residuals do not fluctuate much over time and have only few outliers in several years. Tale B.2 confirms that the residuals have normal distribution each year.

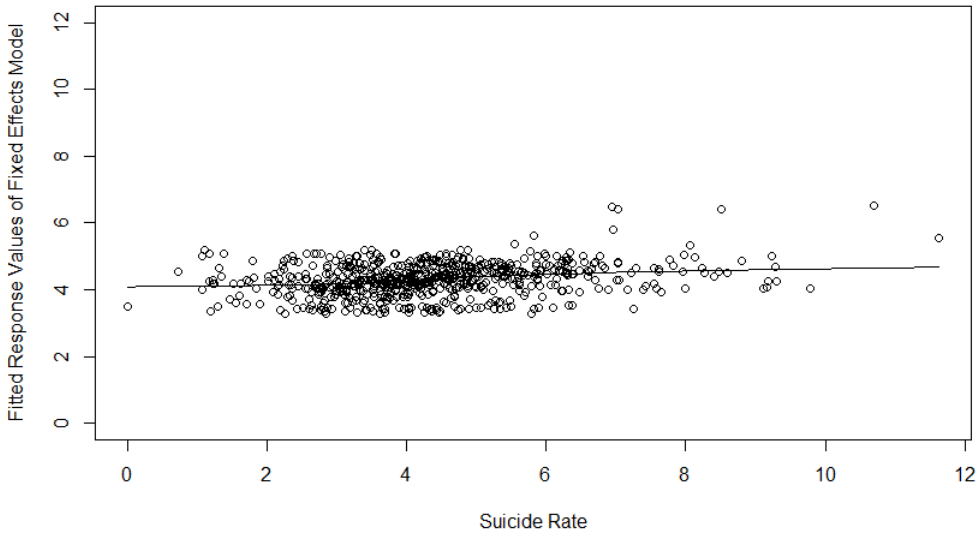

Figure 4.5: Fixed effect model fitted values vs true values

Figure 4.5 shows the scatterplot of the response values and the fitted values of the final fixed effect model. Plot shows some discrepancy between observed and fitted, in the sense that model predicts most of the suicide rates between 3 and 6.

4.2.3 Transition Models

Last of the three standard longitudinal data models that was used was the transition models. Different than the previous models, before constructing the model, the response variable was separated into two variables which included itself and its one year lagged version. Afterwards the model constructed similar to the fixed effect model but included the lag 1 of the dependent variable as a covariate. Unlike the previous models only the divorce variable was significant in this model alongside with the lagged suicide rate variable. Therefore the next model was constructed using only these variables.

Table 4.12: Results of the Transition Model

Coefficients	Estimate	Std. Err.	P value
Intercept	1.128	0.134	$2\text{x}10^{-16}$
Suicide Rate Lagged 1	0.461	0.063	$2.3\text{x}10^{-13}$
Portion of divorces	0.075	0.027	0.005

From Table 4.13 it can be seen that all the variables are still significant and therefore we can consider this model to be the final model for the transition model.

$$\hat{TSR}_{it} = 1.128 + 0.461 TSR_{i(t-1)} + 0.075 PoD_{it} \tag{4.4}$$

From Equation 4.4, we can see that the lagged Suicide rate variable has increasing effect on the suicide rate which is expected. Similar to the previous models divorce rate also has an increasing effect with the coefficient of 0.075 on the suicide rate. In the event where both covariates are zero, we can see that suicide rate would be positive with the value around 1.128. This indicates that at any time and in every province without any divorces, suicide rate would still be positive and significant.

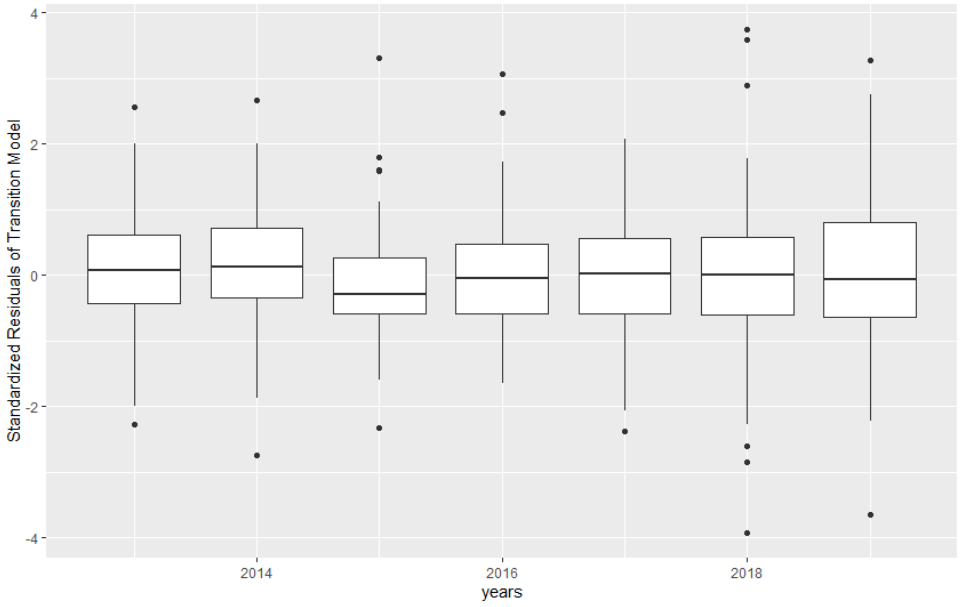

Figure 4.6: Box-plots of Standardized Residuals of Transition Model in each year

From the Figure 4.6 it can be seen that the residuals fluctuate slightly more than the previous two models. Also from the plots we can observe that some of the years don't have a normal distribution which we can also confirm from Table B.3. While these differences may affect the overall fit of the model, since there aren't many outliers and drastic fluctuations, the effects won't be significant.

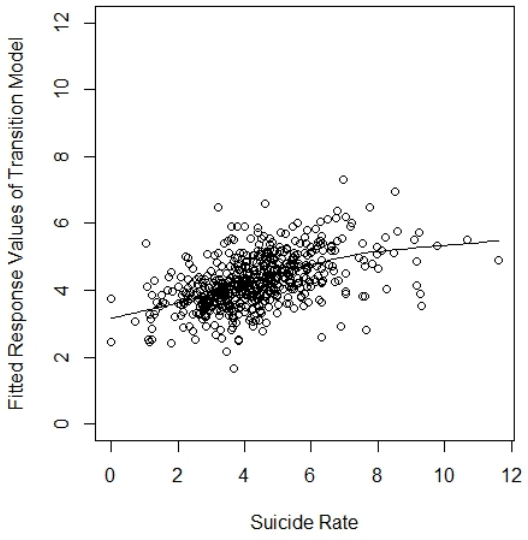

Figure 4.7: Transition model fitted values vs true values

From the Figure 4.7 it can be seen that the scatterplot of the response shows a decent fit which was quantified by calculating root mean squared error(RMSE) as 1.40.

4.3 Hybrid Methods

For the second part of the modelling, we proceeded to form the models which combined the standard longitudinal data models with machine learning methods. The methods that we used in this section included Mixed Effect Regression Trees(MERT),Mixed Effect Random Forest(MERF) and Random Effect -Expectation Maximization Trees(RE-EM Trees) methods. Similar to the standard models, the modelling process for each of the model was done by constructing an initial model which included all of the variables and then refining that model by choosing variables by looking at the variable importance values.

4.3.1 MERT Model

After constructing the three standard longitudinal data models, we then proceeded to the hybrid data models that combine the longitudinal data methods with the machine learning methods. First method that was used was the Mixed Effect Regression Trees(MERT) method. Similar to the previous models we used all of the variables in the first model. After building this model variable importance of each variable were obtained from the regression tree in the model. Based on the results from Table 4.13, two more models were constructed.

Table 4.13: Variable Importance of the Regression Tree in MERT Model

Variables	Variable Importance
Portion of Divorces	8.628
Portion of Hospitals	8.251
Elderly Dependency Ratio	7.061
Portion of Dropout	5.679
Fertility Rate	5.022
GDP per Capita	3.445
Portion of Highly Educated	3.098
Metropolitan	2.469
Portion of Unemployment	0.409

Next model was constructed using the variables with the highest importance values which include elderly dependency ratio,hospital and divorce variables. Third model was constructed using all but the unemployment variable which has a significantly less variable importance compared to the other variables.

Table 4.14: Comparison of the three MERT Models

	MERT.1	MERT.2	MERT.3
RMSE	0.999	1.009	1.019
RESD	0.146	0.159	0.154

From Table 4.14 we can see the comparison of the three MERT models where MERT.1 represents the model with all of the variables, MERT.2 represents the model with divorce,hospital and elderly dependency ratio variables and MERT.3 represents the model with all variables except the unemployment variable. Since the first model has both the smallest root mean squared error(RMSE) and smallest random effect standard deviation(RESD) it was selected.

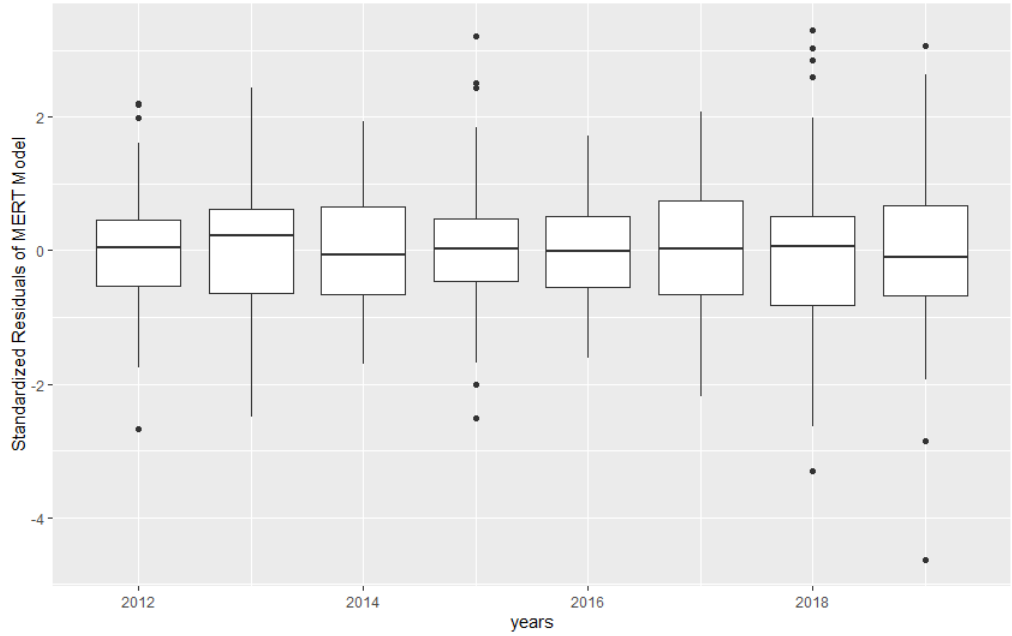

Figure 4.8: Box-plots of Standardized Residuals of MERT Model in each year

Boxed plots of each year for the MERT model can be seen in the Figure 4.8. Like the rest of the residuals models we don't detect any pattern or significant outliers from the plots. From the plots we can see that residuals at each year has normal distribution which we can confirm from Table B.4.

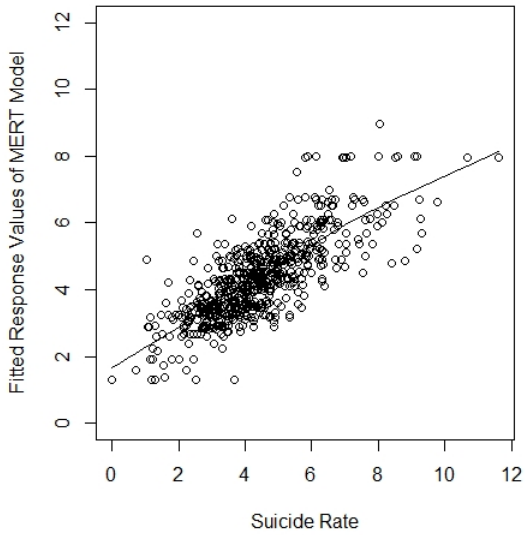

Figure 4.9: MERT model fitted values vs true values

In the Figure 4.9 the fit of the MERT model can be seen. Several of the suicide rate values between 6 and 10 are all predicted as 8 by the model which shows a line like pattern. This may be due to majority of the observations being clustered between 2 and 6 which caused the model to predict larger values inaccurately. RMSE has been calculated approximately as 0.99 which can also be seen in Table 4.14.

4.3.2 MERF Model

Second hybrid method that was used was to continuation of the MERT model which is mixed effect random forest(MERF) model. Similar to the MERT model, an initial model that included all of the variables are constructed and importance of variables are calculated from the random forest part of the model. Variable importance of covariates are calculated by mean decrease accuracy (%IncMSE) which shows how much model accuracy decreases if given variable is left out. Other parameter for calculating variable importance is the Mean Decrease Gini (IncNodePurity) which is defined as the decrease in node impurities from splitting on the variable, averaged over all trees.(Han et al., 2016)

Table 4.15: Importance of variables in the Random Forest part of the MERF Model

Variables	%IncMSE	IncNodePurity
Portion of Hospitals	20.351	7.492
Portion of Divorces	18.866	6.159
Fertility Rate	18.246	5.374
Portion of Dropout	17.959	5.403
Elderly Dependency Ratio	14.768	4.751
GDP per capita	12.477	4.577
Portion of Highly Educated	11.927	4.525
Metropolitan	5.959	0.407
Portion of Unemployment	5.373	4.217

It is important to note that since MERF method uses random forests, values on Table 4.15 as well as their order can change. After running this model for multiple iterations, we observed that the importance of Metropolitan and unemployment variables stay lower than rest of the variables. In contrast, divorce, hospital,illiterate variables as well as fertility rate had consistently high importance values. Therefore for the next two models we constructed we took these observations into account. The second model was constructed using only the variables with the highest importance values which included divorce, hospital,illiterate variables and fertility rate. The third model was constructed by removing the variables with the lowest importance values which were metropolitan and unemployment variables.

Table 4.16: Comparison of the three MERF Models

	MERF.1	MERF.2	MERF.3
RMSE	0.519	0.559	0.516
RESD	0.152	0.155	0.146

In Table 4.16 we can see the three models where MERF.1 represent the model with all variables included, MERF.2 represents the model with the divorce, hospital,illiterate variables and fertility rate variables and MERF.3 represents the model with all vari-

ables except metropolitan and unemployment variables as predictors. From Table 4.16 it can be seen that the third model has both the least RMSE and the least RESD values therefore it is chosen.

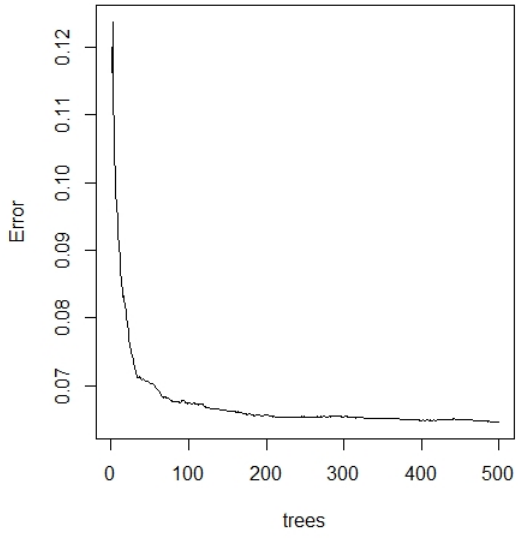

Figure 4.10: Mean square error (MSE) vs number of trees in the random forest part of MERF model

Plot of the MSE with respect to the number of trees in the random forest is given in the Figure 4.10. For our model we used the default tree number of 500 from the function *MERF* which is included in the *LongituRF* package in R. We can see from the figure 4.10 however, that the error values converge after 200 trees so more trees than that amount wouldn't improve our model significantly.

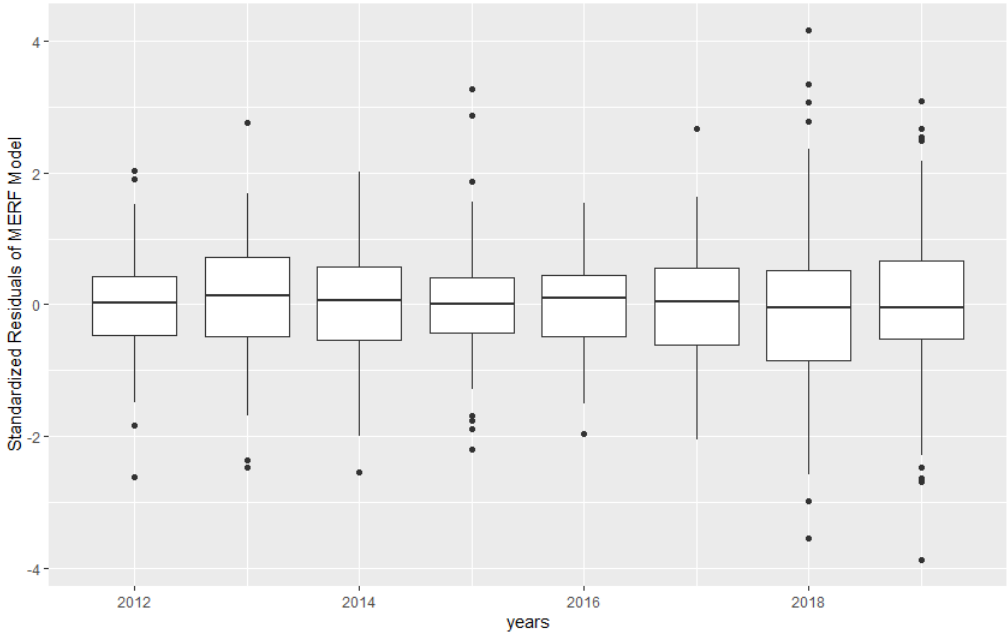

Figure 4.11: Box-plots of Standardized Residuals of MERF Model in each year

Similar to the previous methods, residual boxplots of the MERF model is checked and while there are more outliers than the previous models, their effects to the overall fit is not significant. From the plots we can see that residuals at each year are not all normally distributed which can also be confirmed by Table B.5.

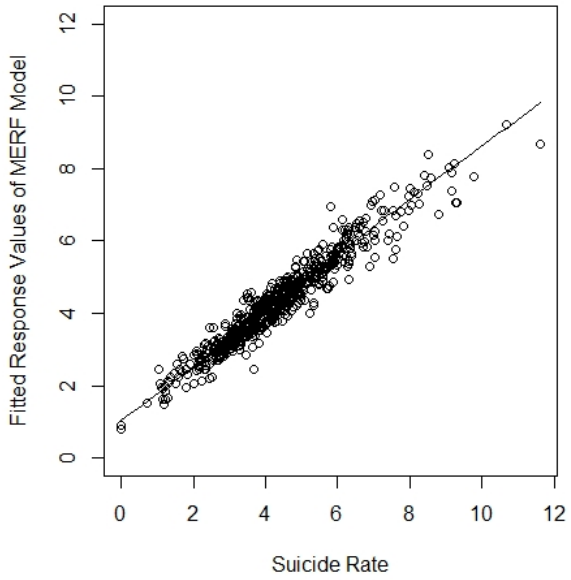

Figure 4.12: MERF model fitted values vs true values

Figure 4.12 shows the plot of the fit for the MERF model. From both the plot and the RMSE value from Table 4.15 we can consider this plot to be a good fit.

4.3.3 RE-EM Trees Model

The last model that was used in this thesis was the Random effects-Expectation maximization(RE-EM) tree model. Similar to the MERT method we first created a model with all of the variables then observed the variable importance values of the regression tree in the RE-EM Tree method.

Table 4.17: Variable Importance of the Regression Tree in RE-EM Tree Model

Variables	Variable Importance
Portion of Hospitals	5.227
Portion of Dropout	4.512
Portion of Divorces	4.429
Fertility Rate	3.214
Elderly Dependency Ratio	2.952
Metropolitan	2.558
GDP per Capita	1.944
Portion of Highly Educated	1.822
Portion of Unemployment	0.031

After constructing the first model and analyzed the variable importance in the model, we then proceeded to construct the next models. From Table 4.17 it can be seen that the variable importance of the hospital,divorce and illiterate variables are higher than the rest of the variables. Therefore the second model has been constructed using only these variables. Also from Table 4.17 it can be seen that unemployment variable has significantly less variable importance compared to the other variables. Thus in the third model we used all of the variables but excluded the unemployment variable.

Table 4.18: Comparison of the three REEM-Tree Models

	REEM.1	REEM.2	REEM.3
RMSE	1.026	1.034	1.022
RESD	0.175	0.189	0.174

Table 4.18 shows the three RE-EM Tree models we constructed where REEM.1 represent the model with all of the variables, REEM.2 represents the model with hospital,divorce and illiterate variables and REEM.3 represents the model with all but the unemployment variables respectively. Since it results in the lowest RMSE and lowest RESD values, REEM.3 model was used as the RE-EM Tree model. Decision tree from the selected model can be found in the Appendix C.

After deciding the model, we then constructed the resulting regression tree to get a better understanding of how the variables are used how the fitted values are calculated.

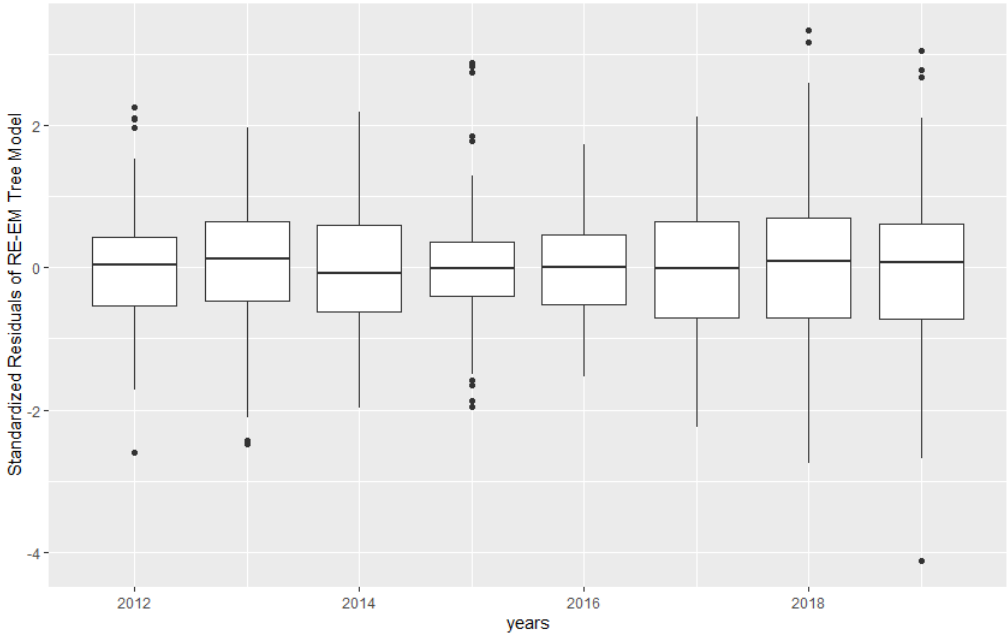

Figure 4.13: Box-plots of Standardized residuals of RE-EM Tree model over time

The residual boxplots of the RE-EM Tree model for each year is given in Figure 4.13. The plots show no discernible pattern or concerning outliers. The distributions all seem to be normally distributed which can be confirmed by the results from Table

B.6.

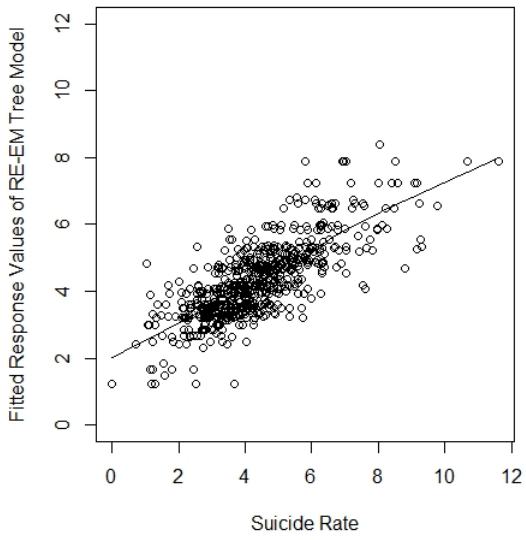

Figure 4.14: RE-EM Tree model fitted values vs true values

Fitted values of the RE-EM Tree model can be seen in the Figure 4.14. Resulting RMSE from the model is around 1.03 thus both this value and the plot indicates a good fit.

4.4 Model Comparisons and Discussion

After constructing all of the models and obtaining the results, we proceeded to compare these results by considering few different categories. First and foremost the accuracy of each model is compared using the root mean squared error(RMSE) and mean absolute error(MAE) metric. We can also compare the models in terms of complexity and computational demand if their accuracies aren't significantly different.

Table 4.19: Comparison Metrics for all models

	Fixed Effects	Random Effects	Transition	MERF	MERT	REEM-Tree
RMSE	1.535	1.097	1.396	0.512	0.999	1.034
MAE	1.157	0.811	1.030	0.366	0.755	0.778

From Table 4.19 we can see the RMSE and MAE values for each of the final models that were constructed. While random effects model has the lowest values among the standard longitudinal data models, MERF model has the lowest values overall. RE-EM tree model and MERT models have similar error rates with RE-EM tree model having slightly higher values for both metrics.

Table 4.20: Correlations between the predicted suicide rate and actual suicide rate values for each model

Methods	Correlation
Random Effect	0.734
Fixed Effect	0.284
Transition	0.494
MERT	0.781
MERF	0.964
RE-EM Tree	0.766

Another comparison method for the models is to look at the correlation values between the predicted and actual values of the dependent variable for the each model. In Table 4.20 we can see those values for all of the models. MERF model has the highest correlation value which indicates the best fit among all of the models whereas the fixed effect model gives the smallest correlation value which indicates the worst fit among all of the models.

After concluding MERF to be the best model, we observed the results for some of the provinces that is given by this model. Tables 4.21 - 4.24 shows the actual and predicted yearly results for four different provinces. Overall the difference between actual and predicted values decrease when the actual values are between 2 and 5.

Table 4.21: Comparison of Actual Values with Predicted Values for Bayburt Province

Year	Actual	Predicted	Difference
2012	5.245	4.0218	1.223
2013	2.645	2.818	0.177
2014	1.280	1.759	0.479
2015	2.514	2.355	0.159
2016	1.187	1.516	0.331
2017	0.000	0.924	0.924
2018	3.690	2.467	1.223
2019	0.000	0.759	0.759

Table 4.22: Comparison of Actual Values with Predicted Values for Trabzon Province

Year	Actual	Predicted	Difference
2012	2.776	3.014	0.242
2013	4.485	4.183	0.302
2014	4.983	4.421	0.562
2015	3.648	3.875	0.227
2016	4.006	3.953	0.053
2017	4.725	4.126	0.599
2018	3.136	3.199	0.063
2019	1.608	2.476	0.868

Table 4.23: Comparison of Actual Values with Predicted Values for Konya Province

Year	Actual	Predicted	Difference
2012	3.911	3.951	0.040
2013	3.921	4.158	0.237
2014	3.677	3.919	0.241
2015	4.859	4.575	0.283
2016	4.287	4.232	0.055
2017	4.007	4.137	0.129
2018	4.423	4.341	0.082
2019	4.641	4.433	0.208

Table 4.24: Comparison of Actual Values with Predicted Values for Istanbul Province

Year	Actual	Predicted	Difference
2012	3.944	3.689	0.255
2013	3.526	3.470	0.055
2014	3.125	3.230	0.105
2015	2.975	3.157	0.181
2016	3.136	3.142	0.006
2017	3.378	3.327	0.051
2018	3.674	3.553	0.121
2019	2.739	3.255	0.514

CHAPTER 5

CONCLUSION

Suicide remains to be an important public health and social concern in the numerous countries around the world. To address this issue it is important to determine the factors that may impact on the suicide rate. Research from the various fields throughout history show that the socio-economic factors of the countries,states and provinces have significant effects on the suicide rate. Therefore determining the possible relations between these factors and the suicide rate is important in order to address these issue of suicide.

To analyze all 81 provinces in a 8 year period between 2012 and 2019, longitudinal data methods were utilized in this thesis. Those models can be separated in two categories; standard longitudinal data models which include fixed effect,random effect and transition models. Second category is the hybrid machine learning models which include: Mixed effect regression tree(MERT) model, mixed effect random forest model(MERF), random effect-expectation maximization tree(RE-EM) tree method.

The predictor variables that were selected among different socio-economic factors include: Fertility rate, elderly dependency ratio, categorical variable to indicate whether the given province is considered metropolitan or not,portion of hospitals to population of province,portion of people who successfully earned a masters degree or PhD to population of province,portion of people who are illiterate to population of province, portion of unemployed people to population of province,portion of divorces to population of province and GDP per capita of each province.

We started analysis by using the standard longitudinal data models. For the fixed effect and random effect models our final model, after removing non significant pre-

dictors, included only two variables which were portion of divorces and portion of hospitals. In the transition model we found out that only the divorce variable was significant aside from the lagged variable. Among these models, random effect model performed significantly better than the rest which was verified by RMSE and MAE metrics as well as the fitted vs predicted values plots.

In the next part we constructed the hybrid models for the analysis. For the MERT model, we used the model that uses all of the variables. For the MERF model, we used the model which had the metropolitan and unemployment variables removed due to those variables having significantly lower variable importance values than the rest of the variables. Similar choice was made for RE-EM tree model where the selected model had all but the unemployment variable due to it having the lowest variable importance. Among these models MERT and RE-EM Tree methods performed similarly which is expected due to the similar methods used in constructing the models. MERF model on the other hand, performed significantly better than both of the hybrid methods as well as all of the other methods.

For the future studies, several other variables can be added to the longitudinal models that couldn't be obtained in this thesis. Those variables include the alcohol consumption and substance use, meteorological data such as average temperature and humidity, and the usage of the prescription drugs specifically the ones diagnosed for mental health problems. Hopefully this study can help in identifying possible factors affecting the suicide rate in order for the government to address and help solve these problems.

REFERENCES

Barth, A., Sögner, L., Gnambs, T., Kundi, M., Reiner, A., & Winker, R. (2011). Socioeconomic factors and suicide. *Journal of Occupational & Environmental Medicine*, *53*(3), 313–317.

Breuer, C. (2014). Unemployment and suicide mortality: Evidence from regional panel data in Europe. *Health Economics*, *24*(8), 936–950.

Durkheim, É. (1951). *Suicide: A study in sociology*. Free Press. New York.

Emamgholipour, S., Arab, M., & Shirani, R. (2021). Socioeconomic determinants of suicide in Iran: Panel Data Study. *Iranian Journal of Public Health*, *50*(11), 2309–2316.

Ferreira, E. R., Monteiro, J. D., & Pires Manso, J. R. (2019). "Death by Economic Crisis": Suicide and self-inflicted injury in the European Union (EU15) during the worst of times. *Society and Economy*, *41*(1), 145–164.

Glen, S. (2020). Hausman test for endogeneity (Hausman Specification Test). https://www.statisticshowto.com/hausman-test/

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451–459.

Hajjem, A., Bellavance, F., & Larocque, D. (2012). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Han, H., Guo, X., & Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*.

Kelly, C., & Dale, E. (2011). Ethical perspectives on suicide and suicide prevention. *Advances in Psychiatric Treatment*, *17*(3), 214–219.

Kõlves, K., Milner, A., & Värnik, P. (2013). Suicide rates and socioeconomic factors in Eastern European countries after the collapse of the Soviet Union: Trends between 1990 and 2008. *Sociology of Health &; Illness*, *35*(6), 956–970.

Machado, D. B., Rasella, D., & dos Santos, D. N. (2015). Impact of income inequality and other social determinants on suicide rate in Brazil. *PLOS ONE*, *10*(4).

Milner, A., McClure, R., & De Leo, D. (2010). Socio-economic determinants of suicide: An Ecological Analysis of 35 countries. *Social Psychiatry and Psychiatric Epidemiology*, *47*(1), 19–27.

Minoiu, C., & Andrés, A. R. (2008). The effect of public spending on suicide: Evidence from u.s. state data. *The Journal of Socio-Economics*, *37*(1), 237–261.

Ross, J. M., Yakovlev, P. A., & Carson, F. (2012). Does state spending on mental health lower suicide rates? *The Journal of Socio-Economics*, *41*(4), 408–417.

Sela, R. J., & Simonoff, J. S. (2011). Re-em trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169–207.

Steller, H. (1995). Mechanisms and genes of cellular suicide. *Science*, *267*(5203), 1445–1449.

Turkstat. (n.d.). https://www.tuik.gov.tr/

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, *30*, 79–82.

Yamamura, E. (2010). The different impacts of socio-economic factors on suicide between males and females. *Applied Economics Letters*, *17*(10), 1009–1012.

Appendix A

CORRELATIONS BETWEEN NUMERIC VARIABLES

Table A.1: Correlations between numeric variables

	Suicide Rate	Fertility Rate	Elderly Dependency Ratio	Portion of Hospitals	Portion of Higher Ed.	Portion of Illiterate	Portion of Unemployment	Portion of Divorces	GDP per capita
Suicide Rate	1	-0.1200019	0.1325913	0.14632854	0.053093573	-0.06825326	-0.010308803	0.21327797	0.07310548
Fertility Rate	-0.1200019	1	-0.7519023	-0.37417021	-0.540294057	0.65583691	0.105704649	-0.65545983	-0.51938498
Elderly Dependency Ratio	0.1325913	-0.7519023	1	0.59422072	0.264572171	-0.31431380	-0.108249465	0.34736502	0.13201029
Portion of Hospitals	0.14632854	-0.37417021	0.59422072	1	0.03115720	0.10420927	-0.05516149	-0.05565176	0.04614689
Portion of Higher Ed.	0.053093573	-0.540294057	0.264572171	0.03115720	1	-0.598990743	-0.008609472	0.513269142	0.431380778
Portion of Illiterate	-0.06825326	0.65583691	-0.31431380	0.10420927	-0.598990743	1	0.13077419	-0.79168853	-0.60134531
Portion of Unemployment	-0.010308803	0.105704649	-0.108249465	-0.05516149	-0.008609472	0.13077419	1	-0.239548197	-0.310927396
Portion of Divorces	0.21327797	-0.65545983	0.34736502	-0.05565176	0.513269142	-0.79168853	-0.239548197	1	0.63045893
GDP per capita	0.07310548	-0.51938498	0.13201029	0.04614689	0.431380778	-0.60134531	-0.310927396	0.63045893	1

Appendix B

NORMALITY TEST TABLES FOR RESIDUALS

Table B.1: Shapiro Wilk Normality Test Results for Random Effect Model Residuals

Year	W Statistic	p-value
2012	0.983	0.375
2013	0.989	0.772
2014	0.971	0.066
2015	0.985	0.468
2016	0.959	0.033
2017	0.978	0.608
2018	0.945	0.230
2019	0.852	0.061

Table B.2: Shapiro Wilk Normality Test Results for Fixed Effect Model Residuals

Year	W Statistic	p-value
2012	0.974	0.104
2013	0.995	0.993
2014	0.989	0.741
2015	0.979	0.207
2016	0.978	0.294
2017	0.961	0.177
2018	0.935	0.143
2019	0.941	0.563

Table B.3: Shapiro Wilk Normality Test Results for Transition Model Residuals

Year	W Statistic	p-value
2013	0.989	0.749
2014	0.982	0.301
2015	0.972	0.075
2016	0.964	0.024
2017	0.948	0.034
2018	0.960	0.373
2019	0.958	0.741

Table B.4: Shapiro Wilk Normality Test Results for MERT Model Residuals

Year	W Statistic	p-value
2012	0.985	0.438
2013	0.996	0.996
2014	0.986	0.519
2015	0.983	0.359
2016	0.990	0.895
2017	0.958	0.149
2018	0.961	0.481
2019	0.886	0.155

Table B.5: Shapiro Wilk Normality Test Results for MERF Model Residuals

Year	W Statistic	p-value
2012	0.982	0.294
2013	0.989	0.702
2014	0.947	0.002
2015	0.944	0.001
2016	0.955	0.020
2017	0.979	0.688
2018	0.931	0.112
2019	0.826	0.029

Table B.6: Shapiro Wilk Normality Test Results for RE-EM Tree Model Residuals

Year	W Statistic	p-value
2012	0.982	0.314
2013	0.987	0.599
2014	0.983	0.369
2015	0.988	0.633
2016	0.974	0.205
2017	0.970	0.373
2018	0.932	0.123
2019	0.923	0.386

Appendix C

R CODES

```
library(vtable)
library(gplots)
library(car)
library(forecast)
summs=colnames(Paneldata3)[ (1:2)]
summarytable=st(Paneldata3,out="return",vars = summs)
summarytable

 Histograms

par(mfrow=c(2,4))
shapirosp=vector()
shapirosw=vector()
suispeeds=vector()
preBoxed=list()




for(i in 0:7)
  n=2012
  year=as.character(n+i)
  xlabel=paste("Suicide Rate in",year)
  hist(Paneldata2[Paneldata2Year==n+i,]SuicideRate,
  xlab=xlabel,main="",label=T)
  preBoxed=
  lappend(preBoxed,Paneldata3[Paneldata3Year==2012+i,]
```

```
    SuicideRate)
  shap=shapiro.test(Paneldata2[Paneldata2Year==n+i,]
    SuicideRate)
  shapirosp=append(shapirosp,shap[[2]])
  shapirosw=append(shapirosw,shap[[1]])
  suispeeds=cbind
  (suispeeds,Paneldata2[Paneldata2Year==n+i,]SuicideRate)
```

```
 Normality Tests

Year=c(2012:2019)
normaltable=data.frame(cbind(Year,shapirosw,shapirosp))
rownames(normaltable)=NULL
colnames(normaltable)=c("Year","W Statistic","p value")
normaltable

 Transformation for normality

Boxcox = function (data)
  return(sqrt(data+1))


revboxcox= function (data)
  return((data 2) 1)



boxed=lapply(preBoxed,Boxcox)
boxshap=lapply(boxed,shapiro.test)
shapirosbp=vector()
shapirosbw=vector()

par(mfrow=c(2,4))
```

```
for(i in 0:7)
  n=2012
  year=as.character(n+i)
  xlabel=paste("Suicide Rate in",year)
  hist(boxed[[i+1]],xlab=xlabel,main="")
  shapirosbp=append(shapirosbp,boxshap[[i+1]][[2]][[1]])
  shapirosbw=append(shapirosbw,boxshap[[i+1]][[1]][[1]])
  suispeeds=cbind(suispeeds,boxed[[i+1]])


 Normality Tests after transformation

Year=c(2012:2019)
normaltable=data.frame(cbind(Year,shapirosbw,shapirosbp))
rownames(normaltable)=NULL
colnames(normaltable)=c("Year","W Statistic","p value")
normaltable



 Correlation

boxeddf=data.frame(boxed)
colnames(boxeddf)=Year
yearcorrs=cor(boxeddf)
boxeddf2=data.frame
(newcol = c(t(boxeddf)), stringsAsFactors=FALSE)
Paneldata4=Paneldata3
Paneldata4$SuicideRate=boxeddf2$newcol
yearcorrs=cor(boxeddf)
covcorrs=cor(Paneldata4[, c(1,2,6)])
covcovs=cov(Paneldata4[, c(1,2,6)])
covcorrs
covcovs
yearcorrs

 Scatterplot
```

```
par(mfrow=c(1,1))
scatter.smooth(Paneldata3$Year,
Paneldata3$SuicideRate,
main="Scatterplot of Crude Suicide Rate versus Year",
               xlab="Year",ylab="Suicide Rate",xaxt= n )
axis(side = 1, at=2012:2019)

 Heterogenity Plots

plotmeans(SuicideRate   Year,
main="Heterogeineity across years",
data=Paneldata3)

library(gee)
library(plm)
library(lme4)
library(REEMtree)
library(MuMIn)
library(panelr)
library(Metrics)
library(geepack)
library(MixRF)
library(LongituRF)
library(lmerTest)
library(car)
library(rpart.plot)


 Marginal Models

 Unstructured correlation


marunst=geeglm(SuicideRate Year+
GDPPerCapitaTh+FertilityRate+
OldAgeDependencyRatio+Metropolitan+
```

```
PortionofDivorces+PortionofHospitals+
PortionofHigherEd+PortionofIgn+PortionofUnemp,
data=Paneldata4,id=Province,
family=gaussian,corstr = "unstructured")
summary(marunst)



marunstsiglm=geeglm(SuicideRate PortionofHospitals+
PortionofDivorces+
GDPPerCapitaTh,
data=Paneldata4,id=Province,
family=gaussian,corstr = "unstructured")
summary(marunstsiglm)



 AR Correlation



marglmarm=geeglm(SuicideRate PortionofHospitals+
PortionofDivorces+
GDPPerCapitaTh,
data=Paneldata4,id=Province,
family=gaussian,corstr = "ar1")
summary(marglmarm)

 Exchangable correlation



marglmexc=geeglm(SuicideRate PortionofHospitals+
PortionofDivorces+GDPPerCapitaTh,
data=Paneldata4,id=Province,
family=gaussian,corstr = "exchangeable")
summary(marglmexc)

 Final Model
```

```
marglmfin=geeglm(SuicideRate PortionofHospitals+
PortionofDivorces,
data=Paneldata4,id=Province,
family=gaussian,corstr = "unstructured")
summary(marglmfin)
```

Different Marginal

```
MuMIn::QIC(marunstsiglm)
MuMIn::QIC(marglmarm)
MuMIn::QIC(marglmexc)
```

Residuals

```
resmar=marunstsiglmresiduals
stdresmar=scale(resmar)
```

getting rmse mae:

```
marfitted1=marunstsiglmfitted.values
marfitted2=revboxcox(marfitted1)
rmsemargp3=rmse(Paneldata3SuicideRate,marfitted2)
rmsemargp4=rmse(Paneldata4SuicideRate,marfitted1)
maemargp3=mae(Paneldata3SuicideRate,marfitted2)
maemargp4=mae(Paneldata4SuicideRate,marfitted1)
 margmae=rmse(Paneldata3SuicideRate,marfitted)
scatter.smooth(Paneldata4Year,stdresmar,
              xlab="Year",ylab="Standardized Residuals of Marginal
                  Model",xaxt= n )
axis(side = 1, at=2012:2019)
```

standardized residuals vs. covariates (marginal model):

```
scatter.smooth(Paneldata4PortionofHospitals,stdresmar,
              main="Scatterplot of Standardized Residuals of
                 Marginal Model versus Portion of Hospitals",
```

```
              xlab="Portion of Hospitals",ylab="Standardized
                 Residuals of Marginal Model")

scatter.smooth(Paneldata4PortionofDivorces,stdresmar,
              main="Scatterplot of Standardized Residuals of
                 Marginal Model versus Portion of Divorces",
              xlab="Portion of Divorces",ylab="Standardized
                 Residuals of Marginal Model")


 fitted response values vs. real response values (marginal model):
scatter.smooth(Paneldata3SuicideRate,marfitted2,
              xlab="Suicide Rate",ylab="Fitted Response Values of
                 Fixed Effects Model",ylim=c(0,12))



 Random Models

rand1=lmer(SuicideRate Year+GDPPerCapitaTh+FertilityRate+
   OldAgeDependencyRatio+Metropolitan+
         PortionofHospitals+PortionofHigherEd+PortionofIgn+
            PortionofUnemp+PortionofDivorces+(1 Province),
        data=Paneldata4)
summary(rand1)

rand2=lmer(SuicideRate Year+GDPPerCapitaTh+FertilityRate+
   OldAgeDependencyRatio+Metropolitan+
           PortionofHospitals+PortionofHigherEd+PortionofIgn
              +PortionofUnemp+PortionofDivorces+(Year Province
              ),
          data=Paneldata4)
summary(rand2)

randsig=lmer(SuicideRate GDPPerCapitaTh+PortionofHospitals+
   PortionofDivorces+(1 Province),
```

```
              data=Paneldata4)
summary(randsig)

randsig2=lmer(SuicideRate PortionofHospitals+
   PortionofDivorces+(Year Province),
              data=Paneldata4)
summary(randsig2)

randsig=lmer(SuicideRate PortionofHospitals+PortionofDivorces
   +(1 Province),
               data=Paneldata4)
summary(randsig)

 Residuals

resrand=resid(randsig)
stdresrand=scale(resrand)
 getting rmse:

randfitted=fitted.values(randsig)
randfitted2=revboxcox(randfitted)
rmserandp3=rmse(Paneldata3SuicideRate,randfitted2)
rmserandp4=rmse(Paneldata4SuicideRate,randfitted)
maerandp3=mae(Paneldata3SuicideRate,randfitted2)
maerandp4=mae(Paneldata4SuicideRate,randfitted)
scatter.smooth(Paneldata4Year,stdresrand,
              xlab="Year",ylab="Standardized Residuals of Random
                 Effects Model",xaxt= n )
axis(side = 1, at=2012:2019)

 standardized residuals vs. covariates (random effects model):

scatter.smooth(Paneldata4PortionofHospitals,stdresrand,
              main="Scatterplot of Standardized Residuals of
                 Random Effects Model versus Portion of Hospitals
                 ",
              xlab="Portion of Hospitals",ylab="Standardized
```

```
                    Residuals of Random Effects Model")

scatter.smooth(Paneldata4PortionofDivorces,stdresrand,
                main="Scatterplot of Standardized Residuals of
                    Random Effects Model versus Portion of Divorces",
                xlab="Portion of Divorces",ylab="Standardized
                    Residuals of Random Effects Model")



 fitted response values vs. real response values (random effects
    model):
par(mfrow=c(1,2))

scatter.smooth(Paneldata3SuicideRate,randfitted2,
                xlab="Suicide Rate",ylab="Fitted Response Values of
                    Random Effects Model",ylim=c(0,12))


par(mfrow=c(1,1))
randdiff=randfitted2 randfitted
hist(randdiff)

vif(rand1)
resrand=resid(rand1)
stdresrand=scale(resrand)



 Transition Model
behind=subset(Paneldata4,Year 2019,select=SuicideRate)
transdata=subset(Paneldata4,Year 2012)
transdata2=subset(Paneldata3,Year 2012)
transdata["Behind"]=behind

trans=geeglm(SuicideRate Behind+Year+GDPPerCapitaTh+
    FertilityRate+OldAgeDependencyRatio+Metropolitan+
```

```
          PortionofHospitals+PortionofHigherEd+PortionofIgn+
             PortionofUnemp+PortionofDivorces,
         data=transdata,id=Province,family=gaussian,corstr="
            independence")
summary(trans)

transsig=geeglm(SuicideRate Behind+PortionofDivorces,
               data=transdata,id=Province,family=gaussian,corstr="
                  independence")
summary(transsig)

 Residuals

restrans=transsigresiduals
stdrestrans=scale(restrans)
 getting rmse:

library(Metrics)
y3=transdataSuicideRate
transfitted1=transsigfitted.values
transfitted2=revboxcox(transfitted1)
rmsetransp3=rmse(transdata2SuicideRate,transfitted2)
rmsetransp4=rmse(transdataSuicideRate,transfitted1)
maetransp3=mae(transdata2SuicideRate,transfitted2)
maetransp4=mae(transdataSuicideRate,transfitted1)
 transmae=mae(backed,transfitted)
scatter.smooth(transdataYear,stdrestrans,
               xlab="Year",ylab="Standardized Residuals of
                  Transition Model",xaxt= n )
axis(side = 1, at=2013:2019)

 standardized residuals vs. covariates (transition model):

scatter.smooth(transdataBehind,stdrestrans,
               main="Scatterplot of Standardized Residuals of
                  Transition Model versus Lagged Suicide Rate",
               xlab="Lagged Suicide Rate",ylab="Standardized
```

```
                Residuals of Random Effects Model")

scatter.smooth(transdataPortionofDivorces,stdrestrans,
              main="Scatterplot of Standardized Residuals of
                  Transition Model versus Portion of Divorces",
              xlab="Crude Birth Rate (Per Thousand)",ylab="
                  Standardized Residuals of Random Effects Model")


 fitted response values vs. real response values (Transition model)
    :
scatter.smooth(transdata2SuicideRate,transfitted2,
              xlab="Suicide Rate",ylab="Fitted Response Values of
                  Transition Model",ylim=c(0,12))


 Hybrid Models

z=matrix(rep(1,648),nrow=648,ncol=1)  Random Effect Predictors

 Mixed Effects Random Forest


set.seed(111)
merf=MERF(Y = Paneldata4SuicideRate, X = Paneldata4[, c(1,2,3,6
    ,10,11)], Z=z , id=Paneldata4Province ,time=Paneldata4Year ,
    sto="none" )
predmerf=predict(merf, X = Paneldata4[, c(1,2,3,6,10,11)], Z=z ,
    id=Paneldata4Province ,time=Paneldata4Year)
revpredmerf=revboxcox(predmerf)
resmerf=Paneldata4SuicideRate   predmerf
stdresmerf=scale(resmerf)

plot(merf[["forest"]],main="")

scatter.smooth(Paneldata4Year,stdresmerf,
              xlab="Year",ylab="Standardized Residuals of MERF
```

```
                    Model",xaxt = n )
axis(side = 1, at=2012:2019)

 fitted response values vs. real response values (Mixed Effects
    Random Forest  model):
scatter.smooth(Paneldata3$SuicideRate,revpredmerf,
                xlab="Suicide Rate",ylab="Fitted Response Values of
                    MERF Model",ylim=c(0,12))

 Rmse
rmsemerfp3=rmse(Paneldata3$SuicideRate,revpredmerf)
rmsemerfp4=rmse(Paneldata4$SuicideRate,predmerf)
maemerfp3=mae(Paneldata3$SuicideRate,revpredmerf)
maemerfp4=mae(Paneldata4$SuicideRate,predmerf)

 Mixed Effects Regression Tree

mert=MERT(Y = Paneldata4$SuicideRate, X = Paneldata4[, c(1,2,3,1
    1)], Z=z , id=Paneldata4$Province ,time=Paneldata4$Year ,sto="
    none" )
predmert=predict(mert, X = Paneldata4[, c(1,2,3,11)], Z=z , id=
    Paneldata4$Province ,time=Paneldata4$Year)
revpredmert=revboxcox(predmert)
resmert=Paneldata4$SuicideRate   predmert
stdresmert=scale(resmert)

scatter.smooth(Paneldata4$Year,stdresmert,
                xlab="Year",ylab="Standardized Residuals of MERT
                    Model",xaxt = n ,ylim=c( 4,4))
axis(side = 1, at=2012:2019)

 fitted response values vs. real response values (Mixed Effects
    Random Tree  model):
scatter.smooth(Paneldata3$SuicideRate,revpredmert,
                xlab="Suicide Rate",ylab="Fitted Response Values of
                    MERT Model",ylim=c(0,12))
```

```
 Rmse
rmsemertp3=rmse(Paneldata3SuicideRate,revpredmert)
rmsemertp4=rmse(Paneldata4SuicideRate,predmert)
maemertp3=mae(Paneldata3SuicideRate,revpredmert)
maemertp4=mae(Paneldata4SuicideRate,predmert)

 Random Effects Expectation Maximization Trees

reem=LongituRF::REEMtree(Y = Paneldata4SuicideRate, X =
   Paneldata4[,c(7,9,12)], Z=z , id=Paneldata4Province ,time=
   Paneldata4Year ,sto="none")
predreem=predict(reem, X = Paneldata4[,c(7,9,12)], Z=z , id=
   Paneldata4Province ,time=Paneldata4Year)
revpredreem=revboxcox(predreem)
resreem=Paneldata4SuicideRate   predreem
stdresreem=scale(resreem)

rmsereemp4=rmse(Paneldata4SuicideRate,predreem)
rmsereemp3=rmse(Paneldata3SuicideRate,revpredreem)
maereemp4=mae(Paneldata4SuicideRate,predreem)
maereemp3=mae(Paneldata3SuicideRate,revpredreem)

rpart.plot(reem[["forest"]])


scatter.smooth(Paneldata4Year,stdresreem,
              xlab="Year",ylab="Standardized Residuals of RE EM
                 Tree Model",xaxt= n ,ylim=c( 4,4))
axis(side = 1, at=2012:2019)

 fitted response values vs. real response values (Mixed Effects
   Random Tree  model):
scatter.smooth(Paneldata3SuicideRate,revpredreem,
              xlab="Suicide Rate",ylab="Fitted Response Values of
                 RE EM Tree Model",ylim=c(0,12))
```

```
 Comparison

rmsesp3=c(rmsemargp3,rmserandp3 ,rmsetransp3,rmsemerfp3,
   rmsemertp3,rmsereemp3)
rmsesp4=c(rmsemargp4,rmserandp4 ,rmsetransp4,rmsemerfp4,
   rmsemertp4,rmsereemp4)
maesp3=c(maemargp3,maerandp3 ,maetransp3,maemerfp3,
   maemertp3,maereemp3)
maesp4=c(maemargp4,maerandp4 ,maetransp4,maemerfp4,
   maemertp4,maereemp4)
comps=rbind(rmsesp3,rmsesp4,maesp3,maesp4)
rownames(comps)=c("RMSEp3","RMSEp4","maesp3","maesp4")
colnames(comps)=c("Marginal","Random","Transition","MERF","MERT","
   REEM")

 Result Correlations
marcorr=cor(Paneldata3SuicideRate,marfitted2)
randcorr=cor(Paneldata3SuicideRate,randfitted2)
transcorr=cor(transdata2SuicideRate,transfitted2)
mertcorr=cor(Paneldata3SuicideRate,revpredmert)
merfcorr=cor(Paneldata3SuicideRate,revpredmerf)
reemcorr=cor(Paneldata3SuicideRate,revpredreem)
Correlation=c(randcorr,marcorr,transcorr,mertcorr,merfcorr,
   reemcorr)
Methods=c("Random Effect","Fixed Effect","Transition","MERT","MERF
   ","RE EM Tree")
rescorrs=data.frame(Methods,Correlation)


 Results for some of the provinces

konya1=subset(Paneldata3,Province=="Konya")
konya2=revpredmerf[as.numeric(row.names(konya1))]
data.frame(Years=c(2012:2019) ,Actual=konya1SuicideRate,Predicted
   =konya2,Difference=abs(konya1SuicideRate konya2))

istanbul1=subset(Paneldata3,Province==" stanbul ")
```

```
istanbul2=revpredmerf[as.numeric(row.names(istanbul1))]
data.frame(Years=c(2012:2019) ,Actual=istanbul1$SuicideRate,
    Predicted=istanbul2,Difference=abs(istanbul1$SuicideRate
    istanbul2))

trabzon1=subset(Paneldata3,Province=="Trabzon")
trabzon2=revpredmerf[as.numeric(row.names(trabzon1))]
data.frame(Years=c(2012:2019) ,Actual=trabzon1$SuicideRate,
    Predicted=trabzon2,Difference=abs(trabzon1$SuicideRate trabzon2
    ))

bayburt1=subset(Paneldata3,Province=="Bayburt")
bayburt2=revpredmerf[as.numeric(row.names(bayburt1))]
data.frame(Years=c(2012:2019) ,Actual=bayburt1$SuicideRate,
    Predicted=bayburt2,Difference=abs(bayburt1$SuicideRate bayburt2
    ))

 Additions after thesis defence

library(ggplot2)
reshapirosp=vector()
reshapirosw=vector()
for(i in 0:7)
  shap=shapiro.test(stdresmar[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])

Year=c(2012:2019)
margres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(margres)=NULL
colnames(margres)=c("Year","W Statistic","p value")
margres


marginerrors=data.frame(years=rep(2012:2019,each=81),stdresmar)
ggplot(marginerrors, aes(x = years, y = stdresmar,group=years)) +
     labs(y="Standardized Residuals of Marginal Model") +
```

```
  geomboxplot()

 Random Effect

reshapirosp=vector()
reshapirosw=vector()
for(i in 0:7)
  shap=shapiro.test(stdresrand[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])


Year=c(2012:2019)
randres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(randres)=NULL
colnames(randres)=c("Year","W Statistic","p value")
randres

randerrors=data.frame(years=rep(2012:2019,each=81),stdresrand)

ggplot(randerrors, aes(x = years, y = stdresrand,group=years)) +
labs(y="Standardized Residuals of Random Model") +
  geomboxplot()

 Transition

reshapirosp=vector()
reshapirosw=vector()
for(i in 0:6)
  shap=shapiro.test(stdrestrans[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])


Year=c(2013:2019)
transres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(transres)=NULL
```

```
colnames(transres)=c("Year","W Statistic","p value")
transres


transerrors=data.frame(years=rep(2013:2019,each=81),
stdrestrans)

ggplot(transerrors, aes(x = years, y = stdrestrans,group=years))
    +  labs(y="Standardized Residuals of Transition Model") +
  geomboxplot()

 MERF
reshapirosp=vector()
reshapirosw=vector()
for(i in 0:7)
  shap=shapiro.test(stdresmerf[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])


Year=c(2012:2019)
merfres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(merfres)=NULL
colnames(merfres)=c("Year","W Statistic","p value")
merfres


merferrors=data.frame(years=rep(2012:2019,each=81),stdresmerf)

ggplot(merferrors, aes(x = years, y = stdresmerf,group=years)) +
     labs(y="Standardized Residuals of MERF Model") +
  geomboxplot()

 MERT

reshapirosp=vector()
reshapirosw=vector()
```

```
for(i in 0:7)
  shap=shapiro.test(stdresmert[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])


Year=c(2012:2019)
mertres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(mertres)=NULL
colnames(mertres)=c("Year","W Statistic","p value")
mertres

merterrors=data.frame(years=rep(2012:2019,each=81),stdresmert)

ggplot(merterrors, aes(x = years, y = stdresmert,group=years)) +
    labs(y="Standardized Residuals of MERT Model") +
  geomboxplot()

 RE EM Tree

reshapirosp=vector()
reshapirosw=vector()
for(i in 0:7)
  shap=shapiro.test(stdresreem[81 i+1:81 (i+1)])
  reshapirosp=append(reshapirosp,shap[[2]])
  reshapirosw=append(reshapirosw,shap[[1]])


Year=c(2012:2019)
reemres=data.frame(cbind(Year,reshapirosw,reshapirosp))
rownames(reemres)=NULL
colnames(reemres)=c("Year","W Statistic","p value")
reemres

reemerrors=data.frame(years=rep(2012:2019,each=81),stdresreem)

ggplot(reemerrors, aes(x = years, y = stdresreem,group=years)) +
```

```
   labs(y="Standardized Residuals of RE EM Tree Model") +
  geomboxplot()

 Suicide Rate Boxplots

suirates=data.frame(Years=rep(2012:2019,each=81),sui=Paneldata3
    SuicideRate)

ggplot(suirates, aes(x = Years, y = sui,group=Years)) + labs(y="
    Suicide Rates of Each Year") +
  geomboxplot()
```