Prediction of death causes using regression models and moving averages

Irfan Abbas[1], and Umair Muneer[2]
[1, 2]Department of computer science and information technology, The University of Lahore, Gujrat, Pakistan
[1]abbairfan440@gmail.com, [2]umair.muneer@cs.uol.edu.pk

Corresponding author email: abbairfan440@gmail.com

***Abstract*— Savings lives of fellow citizens is the first priority of all governments in the world. Regression analysis is used to engine knowledge generation to predict the future threats and a massive support for policy makers and health organization for providing health care facilities. In this paper, our purpose is to address the causality which is causing death in France from 2001-2008[1]. The rapidly increasing data requiring the earlier change before the major human loss. Implementation is suggested through data filtering by regression analysis and moving averages for each cause. Also done the prediction model comparison by implemented linear and quadratic regression models. The storage and retrieval of huge data will enable to get meaningful conclusion out of that data to prevent death ratio in future.**

Keywords— Non-Communicable Diseases (NCDs), Moving Average (MA), Linear regression (LR), Quadratic regression (QR), Simple Moving Average (SMV), Exponential Moving Average (EMA).

1. Introduction

In recent years new diseases' outbreaks are currently being reported at different hospitals. Many of these diseases are proving fatal to human life. It urges health care department to deeply look at some of the diseases that are causing more deaths & plans for their prevention. Collected data [1] about causes of death depends upon age, gender, disease and value in a calendar year. Now a days technology plays a vital role in providing services in the area of health care. To make data useful it must be sorted and analyzed. Before linking the prediction analytics with healthcare, we demonstrate the role of Data visualization in healthcare.

The major focus of technology is to facilitate humans in every field. Analyzing death causes and tracking most prominent disease occurrences to improve public health surveillance and speed response for policy makers. Earlier reports, make more accurately targeted preparations for handling such as choosing the annual influenza patients; and turning large amounts of data into actionable information that can be used to identify needs, provide services, predict and prevent crises, especially for the benefit of city residents.

On world level in 2016, many people died with special causes, showing that plenty of deaths are caused by non-communicable diseases (NCDs) like Cancers, Diabetes and other long-term illness. About 70% of deaths are caused by NCDs. The data of past years till 2016 was collected by IHME and published on 2017[2].

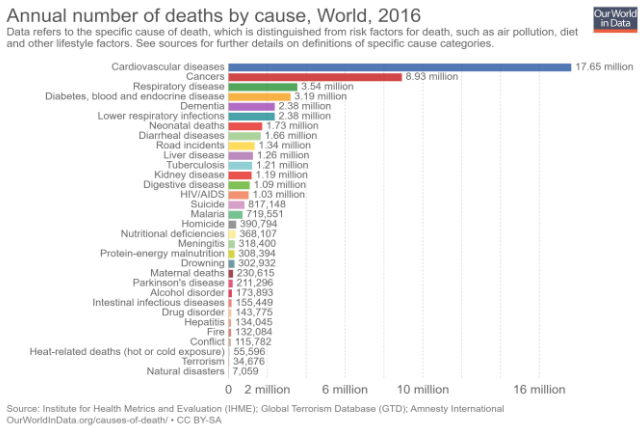


Figure 1. World death causes by NCD

1.1 Problem statement

The aim of this paper is to investigate the expansive datasets to extract the meaningful and cover up data to increase the efficiency of the whole system, which makes it possible to predict death causes in an effective manner. Thus, our aim is to reduce the number of deaths with regression models and measurable techniques. Paper attempts to overcome the death ratio with the utilizing the data that is associated with policy makers' evaluation for their planning's and instructions..

1.2 Regression

A statistical process that attempts to determine the strength of the relationship between the variable (dependent variable) and a series of other predictive variables (independent variables). The independent variable that is changed and generally symbolized by X and the dependent variable indicates Y. The regression estimates the relationship between variables to predict.

Regression analysis also measures optimal value, conditional expectation, measurement, probability distribution, and so forth. For the dependent variable in the independent variables.

The regression analysis is widely used for prediction, where its use is highly overlapping with the field of automatic learning. Regression analysis is also used to understand which independent variables are associated with the dependent variable and to explore the forms of these relationships. In restricted conditions, regression analysis can be used to infer causal relationships between independent and non-independent variables. However, this can lead to illusions or false relationships, so caution is advised [2].

Many techniques have been developed to perform regression analysis. Known methods, such as linear regression and normal quadratic regression, are parameters, because the regression function is defined in terms of a specified number of estimated anonymous parameters of data. Nonlinear regression refers to techniques that allow regression function in a specific set of functions, which can be unlimited in dimension.

The performance of regression analysis methods in practice depends on the form of the data creation process and how it relates to the regression approach used. Since the actual form of data generation is not generally known, regression analysis relies to a certain extent on assumptions about this process. These assumptions can sometimes be verified if sufficient data is available. Regression models are often useful to predict even when assumptions are violated moderately, although they may not work optimally. However, in many applications, especially with small effects or causal questions based on monitoring data, regression methods can produce misleading results. [3]

When the relationship between the dependent variables and the regression parameters are used to represent dependent variables in terms of linear independent variables is supposed to be called linear regression.

The case of the explanatory variable is called a simple linear regression. For more than one explanatory variable, the process is called a linear regression. This term differs from linear multivariate regression, where many dependent variables are predicted, rather than a single variable [10].

1.3 Data Visualization

Data visualization places data in a visual context to help people better understand the significance of data [3]. The objective of the visualization is to identify the useful data from data set and change it into an equitable structure for further use. The core challenge in performing visualization to filter the noisy data. Visualization generates the effectiveness in testing and predicting. The process of visualization Integrates large volumes of data to get at the soul of that data and convey the key insights. In this paper we visualize the data using LR , QR and MA (Moving Average) charts.

2. METHODOLOGY

2.1 How to predict the death causes

The study area of death causes observation is located in France. First, we find out a valid data source than we have to move that raw data in a proper information using excel sheet. Now the data has been organized into rows and columns. Now we can easily make the prediction. We have obtained data set to predict the disease's fatal effect on human lives. Range of the dataset is from 2001 to 2008[1]. To better understand the dynamics of deaths, we first extract the useful data from raw dataset. Now we organize the data in Alphabetical order to get the ability to differentiate between different diseases and other information. We filter the data by years to get the more accurate and measure able prediction.

While breaking up the data with more than two unique values it must be in an organized form. For better understanding, we can distinguish each fatal disease in corresponding years. Each disease categories in single chart. Good Visualization makes the understating of the situation easy. Data of each disease is being displayed using moving averages, QA and LR trend lines [3].

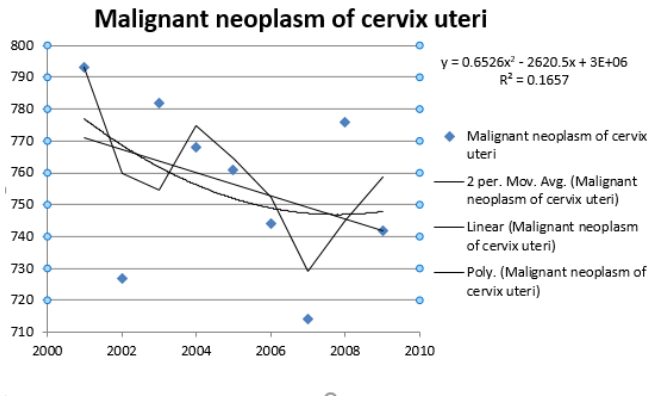


Figure 1 Malignant neoplasm of cervix uteri Disease

Total 65 diseases charts are represented to get help for prediction. Linear regression line, quadratic and moving average is also included on the charts, which help in prediction. Here is the chart of individual prediction of disease yearly. Same as above charts generated to get the prediction of all 65 diseases.

2.2 Linear regression

Linear regression relationships are designed using linear prediction functions in which parameters are estimated for an unknown model of data. These models are called linear models. [3] Mostly, the conditional average of the response given to the values of explanatory variables (or predictors) is assumed to be a function associated with those values; Occasionally, a conditional broker or some other quintiles are used [5]. Like all forms of regression analysis, linear regression focuses on the distribution of the conditional probabilities of the response given the values of the forecasters, rather than the common probability distribution of all these variables, the area of multivariate analysis.

Here is the Equation for simple linear regression:

$$yi=\beta 1+\beta 2xi+ei \quad (1)$$

VALUE	REPRESENT
YI	Dependent Variable
XI	Independent Variable
EI	Residual(A random error term)
$\beta 1$	Parameter for intercept(Regression coefficient)
$\beta 2$	Parameter for slope(Regression coefficient)

Table 1 Equation of linear regression

In principle, the residual should account for all the movements in Y that cannot be explained by X [7].

Linear regression was the first type of regression analysis that was carefully studied and widely used in practical applications [4]. If the objective is prediction or error reduction, linear regression can be used to match a predictive model to a set of observed data for response values and explanatory variables. After developing such a model, if additional values of explanatory variables are collected without an accompanying response value, the model can be used to predict the response values. If the objective is to clarify the difference in the response variable that can be attributed to the difference in the explanatory variables, linear regression analysis can be applied to measure the strength of the relationship between the response and the variables [5].

Here we know, how we could examine death from diseases an average death rate plus the deaths which are happen in each dataset year. So, the deaths forecast by using this methods:Some Common Mistakes

*Number of deaths = Last year death + increment value * increment*

Now, we can interpret the coefficient increment value as the increment in death for each year. And with this coefficient we can start making predictions by just knowing the rate of increment in death or decrement.

Year	Event of undetermined intent (No. of Deaths)
2001	553
2001	202
2002	594
2002	232
2003	506
2003	121
2004	399
2004	87
2005	376
2005	67
2006	414
2006	76
2007	425
2007	82
2008	300
2008	68
2009	154
2010	130
2011	112
2012	94

Table 2 undetermined intent deaths table

We calculate the rate of change by this coefficient in excel and demonstrate the increase and decrease in death rate by linear trend line. Calculate the forecasting year by sum up each determinant on the right-hand side of the equation. Highlighted value is shown in Table

2.3 Quadratic regression

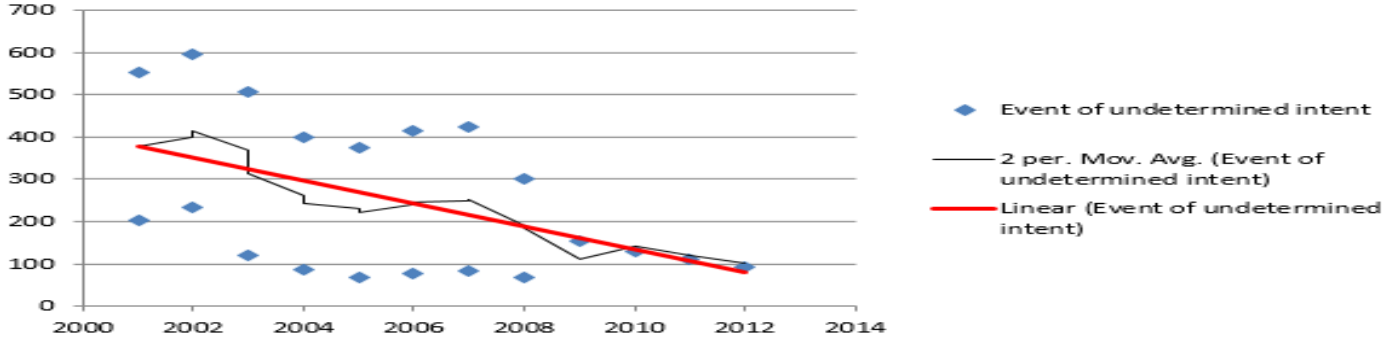
Quadratic regression model is used when the relationship between the variables is inherently polynomial in nature we use this model to get more accurate forecast value.

The quadratic equation not works like linear equation. If we put the values (slope) same like linear in linear equation its show an unexpected value. And to sum up the right side values to get forecast value. In this model we first calculate the R2 of the slope value. Than it adds into regression model [14].

$$y=ax^2+bx+c \text{ where } a\neq 0 \qquad (2)$$

Least square method is an ideal way to solve this quadratic equation. Before solving the equation we find the values of a,b, and c such that the squared vertical distance between each point (xi,yi) and the quadratic equation is minimal. While forecasting the death rate we calculate the xi and yi values. These values show the distance between the lines. In general, we calculate by just putting the value and their square. The following chart show the polynomial line for the endocrine disease and also show the equation and R2 value on the right side of chart which is calculated.

Figure 2 undetermined intent deaths with LR line



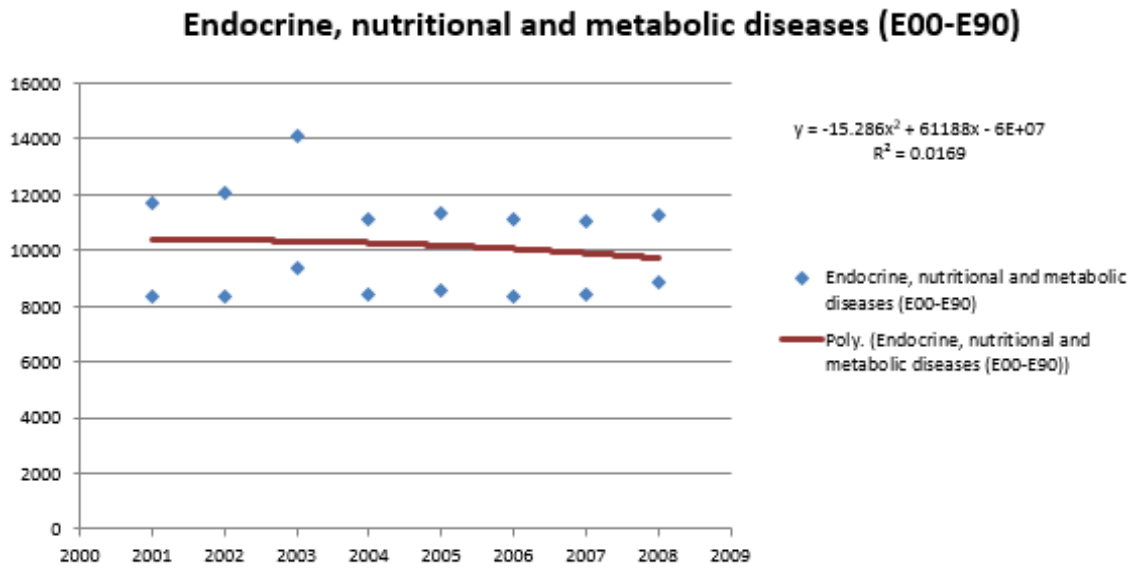


Figure 4 Endocrine disease deaths with Polynomial trend line.

When we calculate the quadratic equation two type of value we can forecast one is interpolate and second one is extrapolating. Interpolate is estimation of a value within two known values in a sequence of values. Extrapolating is an Assessment of a value based on increasing a known sequence of values or facts apart from the area that is certainly known. [15].

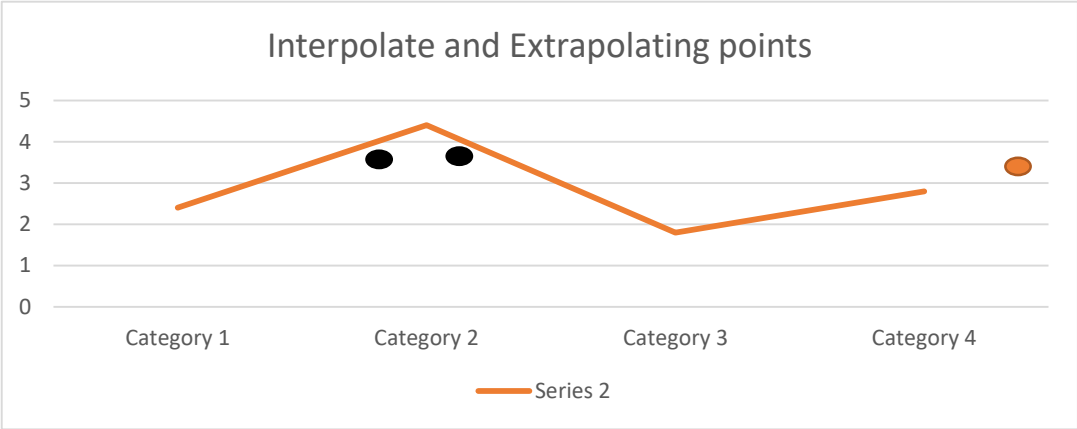


Figure 5 Interpolate and Extrapolating points in QR model

In this graph the black balls are show the interpolate value and the red balls are show the extrapolate value which is the forecasting value.

2.4 Moving averages

A moving average is used as indicator in technical analysis that helps smooth out deaths action by filtering out the noise (raw data) from random deaths fluctuations. It is a trend-following, or lagging, indicator because it is based on past deaths [11].

A simple moving average (SMA), method is calculated by taking the arithmetic mean of a given set of values. In this we take the number of deaths, are added together and then divided by the total number of deaths in last year by that disease. The main thing about moving averages came into mind is it only the mean? No, the new values become available, the oldest data points must be dropped from the set and new data points must come in to replace them. Thus, the data set is frequently "moving" and it leave space for new data. This method of calculation ensures that only the current information is being accounted for [12].

The prediction using moving average is done by EMA (exponential moving average) method. The EMA is a type of moving average that gives more weight to the recent death values in an attempt to make it more responsive to new information. Learning the somewhat complicated equation for calculating an EMA [13]. It is used because, it is more accurate than the SMA method due to smoothing factor calculation. Responsiveness is an also a factor for choosing this method for forecasting. Here is the equation EMA.

$$\mathbf{EMA = (D * \alpha) + (Pervious\ EMA * (1- \alpha))} \quad \mathbf{(3)}$$

D = Recently deaths
α = smoothing factor = 2/1+N
N = Number of years

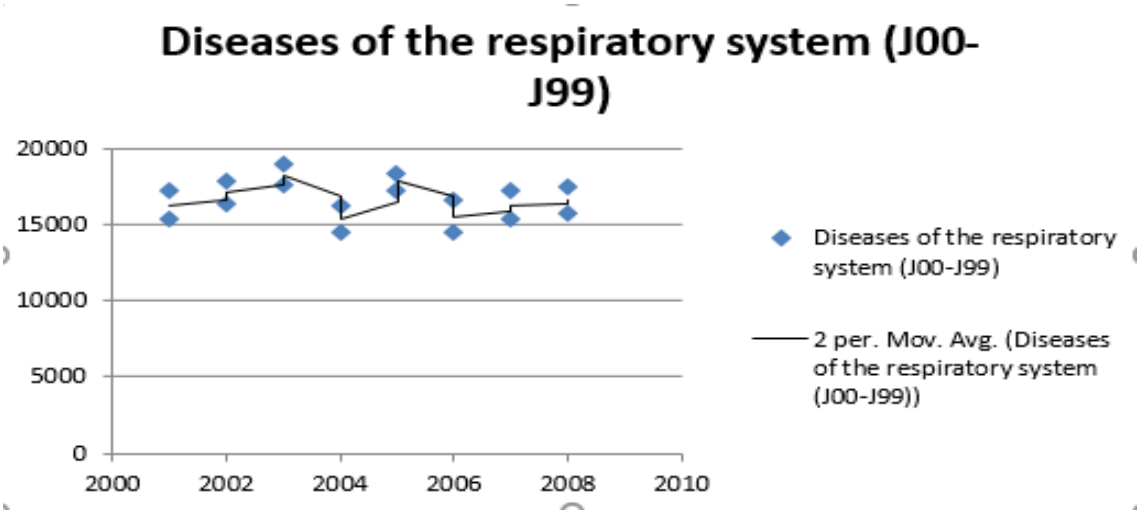


Figure 6 Diseases of respiratory system is shown by MA trend line

The MA chart represent the respiratory system disease deaths till year 2008. Moving average line indicate the values to help out in the forecasting trends.

3. ANALYSIS

In order to explore the optimistic approach for prediction of feature deaths we first make and comparison with the uses approaches. In this paper we successfully use the linear regression model to forecast the future death rate. In linear regression model suitability for condition is a specified and regular gap in the value. If the slope coefficient in this model change after some record it show only the line. That trends line choice is not an ideal prediction.

When we came into the non-linear regression model it provides the most flexible curve fitting functionality for forecasting. Quadratic model is best and updated model of regression than linear regression model. Linear models assume the functional form is linear, But not the relationship between your variables [9]

Nonlinear regression can be harder to perform. In QR the different pattern should be adopted to forecast the interpolated and extrapolate value. In calculation you also need to tell the starting values for the nonlinear algorithm. Some datasets can require substantial effort to find acceptable starting values. In staring it may be fail but we done QR model implementation after the linear regression model.

Nonlinear regression is a powerful alternative to linear regression but there are a few drawbacks. Fortunately, it's not difficult to try linear regression first. Quadratic model is essentially a linear model in two variables, one of which is the square of the other. We see that however good the linear model was, a quadratic model performs even better in additional variance.

3.1 Experimental reports

We take the real data for predicting the deaths by diseases. Real data gives true and good result rather than fuzzy data. In the prediction, the chart shows that each disease that causes death can more take deaths. Accident, Disease of the circulatory system, Disease of the respiratory system, ischemic heart disease, malignant neoplasms, Neoplasms, and other heart diseases will be the leading causes of deaths in 2020. The observed diseases which can be more effective the human lives are poisoning by exposure to noxious substance and ratio is 13% the main thing which causes this is foodstuff and energy drinks. In 2016 world health report by NCD shows that more than 1.19 million peoples die due to kidney diseases. In forecasted result about 10% chances to increase in these type of diseases.

In this modern generation, Drugs and Alcoholic drinks are used as a fashion icon. The predicted ratio of drugs is about 22% which is an alarming cause and a big challenge for policymakers to overcome this. Humans are born with all of the nerve cells that will serve them throughout life. When neurons die, so a human die [7]. Nervous system failure ratio will be increased by up to 19%.

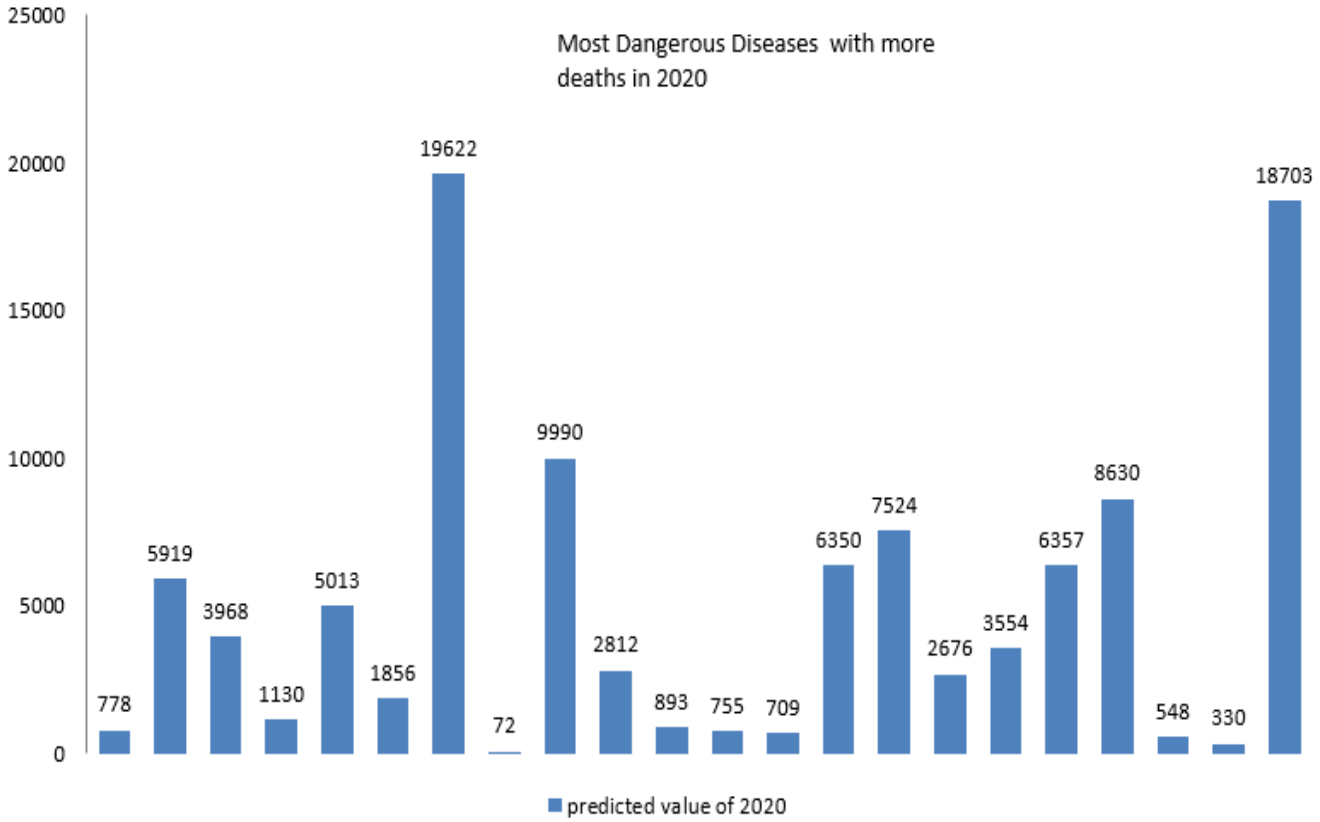


Figure 7 Alarming death causes

Extreme temperature can make breathing very difficult. Cold weather much harmful for lungs patient. Lungs disease ratio also increase and predicted increment ratio is about 24%.

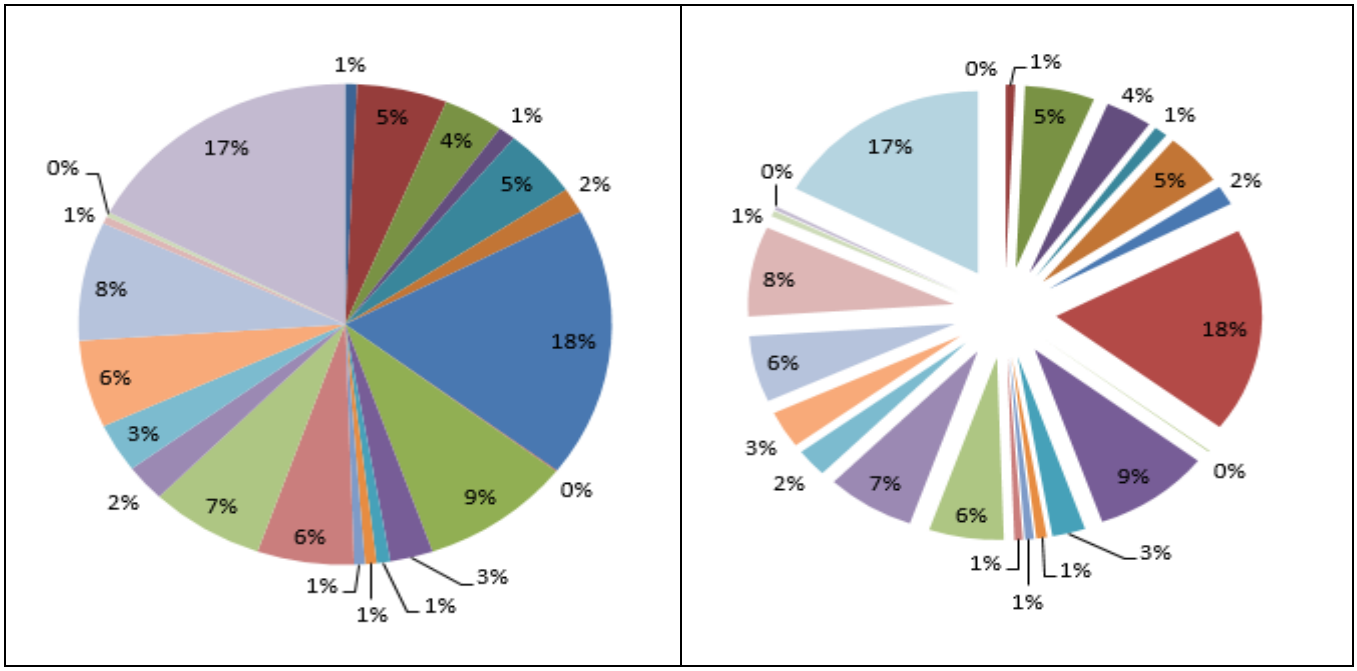


Figure 8 Controlled death ratio chart

Malignant neoplasm of breast and cervix uteri are also caused by death more than 33% but a number of deaths are less than 800 in 2020 which can be controlled easily in earlier stages.

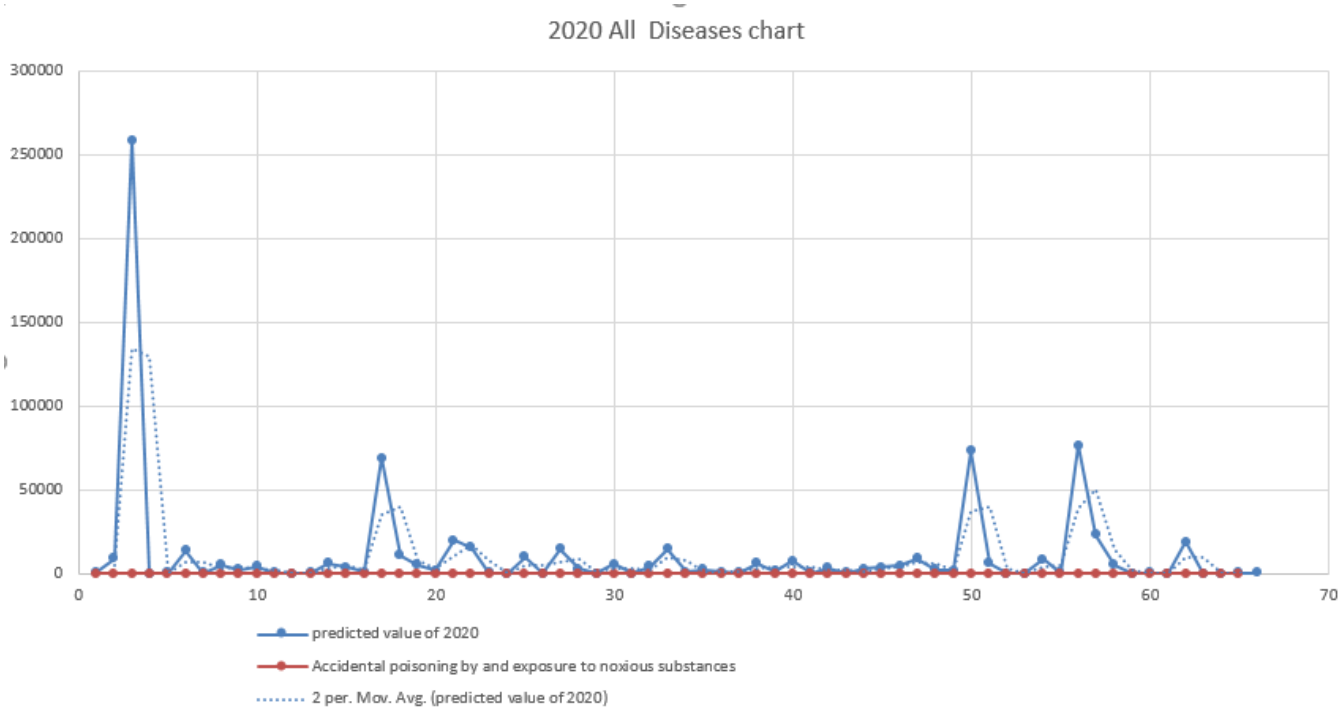


Figure 93 2020 prediction values of each disease

The Average scaling down death ratio is 3%. Which is low than United nations sustainable development goals 2030. The united nations seventeen goals contain health policy which makes it possible to overcome deaths ratio.

4. Conclusion

We have successfully predicted the death causes of year France in year 2009. The real data helps a lot to monitor the deaths rather than sample data. The work being done on the human lives has been demonstrated as a core factor. The results of yearly prediction are proven. Each disease of deaths predicts in separate form, which concludes that all death diseases are of a great help to forecast in a single chart.

We suggest by providing the reports to hospital's, that it will help to predict the future expected patients. It is also helpful for government health departments to take the right decision for policy making. In future it will also be helpful, in the work for curing of disease which are rapidly increasing and proving fatal for human health.

References

[1] https://github.com/EIS-Bonn/MA-INF4314-Lab-WS2014/blob/master/A-EvaluationLinDA-Fabrizio/dev/Evaluation%20reports/datasets/CausesOfDeath_France_2001-2008.csv

[2] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2016 (GBD 2016) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2017.

[3]. Brandeau, Margaret L., François Sainfort, and William P. Pierskalla, eds. Operations research and health care: a handbook of methods and applications. Vol. 70. Springer Science & Business Media, 2004.

[4] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, 2012.

[5] Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 329. John Wiley & Sons, 2012.

[6] Forster, Piers, et al. "Changes in atmospheric constituents and in radiative forcing. Chapter 2." Climate Change 2007. The Physical Science Basis. 2007.

[7] Hutchins, James B., and Steven W. Barger. "Why neurons die: cell death in the nervous system." The Anatomical Record: An Official Publication of the American Association of Anatomists253.3 (1998): 79-90.

[8] Nilsson, Måns, Dave Griggs, and Martin Visbeck. "Policy: map the interactions between Sustainable Development Goals." Nature News 534.7607 (2016): 320.

[9] Harrell, Frank E. "Ordinal logistic regression." Regression modeling strategies. Springer, Cham, 2015. 311-325.

[10] Cohen, Patricia, Stephen G. West, and Leona S. Aiken. Applied multiple regression/correlation analysis for the behavioral sciences. Psychology Press, 2014.

[11] Marple, S. Lawrence, and S. Lawrence Marple. Digital spectral analysis: with applications. Vol. 5. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[12] Johnston, F. R., et al. "Some properties of a simple moving average when applied to forecasting a time series."

Journal of the Operational Research Society 50.12 (1999): 1267-1271.

[13] Burgstahler, Lars, and Martin Neubauer. "New modifications of the exponential moving average algorithm for bandwidth estimation." Proc. of the 15th ITC Specialist Seminar. 2002

[14] Stinchcombe, John R., et al. "Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing?." Evolution: International Journal of Organic Evolution 62.9 (2008): 2435-2440.

[15] Verma, Surendra P., and Alfredo Quiroz-Ruiz. "Critical values for 33 discordancy test variants for outliers in normal samples of very large sizes from 1,000 to 30,000 and evaluation of different regression models for the interpolation and extrapolation of critical values." Revista Mexicana de Ciencias Geológicas 25.3 (2008).