International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review Paper on Forecasting of Demographic Features using Statistical and Data Mining Methods

Swati Jain[1], Dr. Nitin Mishra[2], Sakshi Bokade[3] , Saumya Chaturvedi[4]

[1] M.Tech Scholar, CSE, RCET, Bhilai, js.jainswati@gmail.com

[2]Associate Professor, CSE, RCET, Bhilai, *drnitinmishra10@gmail.com*

[3]Assistant Professor, CSE, RCE, Bhilai, *sakshisaxena12oct@gmail.com*

[4]Software Engineer, Aldoshik Delhi, *Saumyanmishra5@gmail.com*

Abstract: ***The word forecast means expected outcome in the future, also known as projection. Literacy is considered as an heart of human development. Because literacy plays a vital role in development, early formulations of literacy goals are required. Planning some improvements in education system to increase literacy rate require study, analysis and forecast of various features affecting it. Major features include population count, population density, economic conditioning and male to female ratio. Population projection is generally referred a challenging task specially in developing countries due to unavailability of reliable data .These predictions are useful to researchers, governments and various organizations for planning purpose, social and health research, for monitoring development goals and also as input for other areas of forecasting. In this paper, we provide a summary of various existing methods that have been used for population forecasting. We propose combination of these methods for forecasting of demographic features of Chhattisgarh. We have considered population forecast as an independent function however it must be recognized that economic factors play a vital role in population changes.***

Keywords: Population Projection, Statistics, Prediction, Correlation, Regression

1. Introduction

Data Mining is an analytic process which explores data to find consistent patterns and/or systematic relationships between variables, and then the detected patterns are applied to new subsets of data for validating the findings. Every kind of data has some information hidden in it. The data can be analyzed to find patterns and based on this patterns future events can be predicted. Data mining is also referred as knowledge discovery in database (KDD).Statistics being science of collecting, analyzing and presenting data, KDD is statistics and data mining is statistical analysis.

Population projection refers to the estimates of population in near future. Almost every country carries out census to collect different features of each geographic region and to have count on population in each region Collection, analysis and presentation of these data is called Statistics These censual data can be used to make future predictions which could help planning commission to make plans and focus on areas which require special attention. Government mostly relies on centralized and official sets of population forecasts based on which capital facilities are planned. If the value is underestimated, it will make system inadequate for the purpose intended; similarly if the value is overestimated, the system will become costly. Changes in the population of the city over the years occur, and the system should be designed taking into account these changes in the population at the end of the design period.

The terms 'projection' and 'forecast' have different meaning for demographers. A projection is defined as the numerical outcome of a set of techniques and assumptions related to future trends whereas a forecast is the specific projection which is most likely to provide an accurate prediction of future population change. But the authors use the terms interchangeably. I will do the same in present discussion.

Chhattisgarh is a growing state with total population of 2,55,40,196 (2.55 Cr) as of 2011 census. It ranks16th in India in terms of population with Males: 1, 28, 27,915 (1.28 Cr) and Females: 1, 27, 12,281 (1.27 Cr). Literacy rate has increased from 64.66% to 71.04%.Both male and female literacy rate has increased. Despite of increased literacy rate, literacy rate of Chhattisgarh is very low because tribal area is less educated.

The motto of research on demographic features of population data can be expressed as: Population problem is an important matter of concern in many countries especially in less developed and developing countries. The increased growth of population has resulted in decrease in productivity and income. The unplanned growth of population will result in scarcity of resources like food, water, houses etc. If population problem are not handled in a right way and at right time its effects can be dangerous

Forecasting future gained importance as it aids individual and organizations in decision making. Effective forecasting of demographic features has lot of advantage like providing early information about increase or decrease in various features of population census data. This information will be useful for the government or planning commissions to make adequate plans and to plan capitals required beforehand. Predicting characteristics of local populations improves local governance in planning, politics and policy analysis.

2. Literature Survey

Literacy is essential for development and for effective development planning proportion of literate men and women need to be known. Literacy has no standard definition but can be referred as "context-bound continuum of reading, writing and numeracy skills, acquired and developed through processes of learning and application, in schools and in other settings appropriate to youth and adults"(UNESCO,2005). Most researchers attempting to forecast literacy focused on age pattern of literacy and its change over time.

Most widely used method for forecasting population is cohort-component method. This method is based on future levels of fertility, mortality, sex composition, migration, and other parameters.[3]This method separately projects each of the factors that causes population change.

Basic equation for this model is:
P (t+n) = P(t)+Births−Deaths+Immigrants−Emigrants

Where, *t* is the starting point of time; *n* is the projection interval; *P(t)* is the population-size at time *t*; and *P(t+n)* is the population size at time *t+n*..

Although Cohort-Component method is one of the most widely used method for population forecasting, but it is data intensive. (George et al. 2004, Smith Tayman and Swanson 2001).Adrian E. Raftery et.al[8] found in his study that widely used cohort component method for population projection does not yield an estimation of uncertainty about future population quantiles. To overcome this he modeled a Bayesian probabilistic population projection method. Total fertility rate and life expectancy predicted using this model was then used as input to cohort component method.

Population forecast was also done through regression based approach using Hamilton-Perry method [26]. Data of two recent census was used to predict population from time t to time t+k using cohort changing ratio (CCR) [1]. It also satisfies the fundamental demographic equation as cohort-component model but has far less intensive data requirements.

Regression analysis has also been used in financial institution, with time series data. It was used to periodically forecast stock market prices and been proven complement to other numeric forecasting method.[21] Global age-specific Literacy projection(GALP) model[15] was used to forecast the literacy rate based on age disaggregated literacy data and demographic data collected from UN population division.

Several authors have used time series methods mostly autoregressive integrated moving average (ARIMA) methods to forecast total birth [25]. Although statistical methods like smoothing or ARIMA are mainly used for time series forecasting [19].

Some computational intelligence (CI) technique has recently been proposed. Examples for these include: Artificial neural network (ANN), Support vector machines (SVM), Fuzzy technique or combination of them. Sven F.crone found in his study that ensemble of CI technique outperforms compared to combination of statistical method [2]. Valanzula et al. proposed hybridization of intelligent techniques which takes advantage of easy to use ARIMA models and computational power of ANN.

N.Rajasekhar et. al[4] propose hybrid support vector machine technique for weather prediction. .K-means clustering algorithm was applied initially to form clusters and then SVM technique was applied on previously clustered dataset. Clustering is done using Euclidean distance on monthly mean of each year average temperature.

Ryan W.Kirk[5] uses quantile regression method for evaluation of landscape trends of new building construction using parcel building dataset of Macon country. Quantile regression (QR) is distinguished from linear regression (LR) in that, LR uses mean or average of distribution while QR uses conditional probability distribution function like median. Also future development pattern are forecasted using combination of methods –population growth projection, spatial logistic regression model and extrapolation of recent development density trends.

Social media content has also been used for prediction of real world data [6].Sitaram Asur et.al used chatter box of Twitter, a social site, to forecast the revenues of box office from different movies. Linear regression model was used for this purpose and the result found was far more accurate as Hollywood stock exchange. Similarly auto regression method has also been used on twitter data [7] to track and predict the influence and spread of influenza epidemic in population. The result of this model was tested with Centers for Disease Control and Prevention (CDC) data and it was found to be accurate in predicting influenza-like illness (ILI) cases.

A prediction model based on statistical approach with its application to discrete time data was proposed by Neda sadhegi et.al[9].This model uses non-linear parametric temporal function and non-liner mixed effects modeling (NLME) for analysis of early infant brain maturation.

Jagdish Prasad et. al[10] forecasted the need of contraceptives based on the trends in different districts of Rajasthan. Least square method is applied to obtain the regression equation for each of the districts and based on these equations values of different family planning methods are forecasted.

A crime analysis tool was developed by Malathi. A et. al[11] using different data mining techniques namely data cleaning, clustering , classification and outlier detection. The model thus proposed enabled investigators to characterize and analyze crime data economically to find out trends and patterns over years using clustering and to forecast the crime using classification technique.

Data mining techniques combined with temporal abstraction were used to analyze patients' biochemical data [12].Patients' hospitalization probability can be mined using temporal hospitalization pattern and this can be used by doctors to suggest some immediate solution to patient avoiding hospitalization.

F.Martinez Alvarez et. al[13] proposed a model for prediction of earthquakes. The techniques used for prediction were Quantitative association rule (QAR) and regression. Predictions are made based on the b-value, which reflects the tectonics and geospatial properties of rock and also the fluid pressure variation on the surface, as given by Gutenberg-Richter law.

Human population growth was used to forecast the threatened mammals and bird species by nation by Jeffrey K. McKeea[14].Stepwise regression method was used to correlate number of threatened species per unit area and it was found that number of threatened species have strong correlation with human population density and species richness.

Warren C.Sanderson studied various forecasting methods [16] to answer a question "can Knowledge improve forecast" and found that method with less accuracy of forecasting when combined with more accurate forecast method results in more accurate forecasting compared to methods being used alone.

The accuracy of forecast was measured using number of methods such as root mean squared error, mean absolute error, and mean absolute proportional error. Felix Salfner et al[23] proposed method for online failure prediction. Uncertainty in measuring forecast was measured by David A. Swanson et.al[17] using two approaches namely projections based on alternative scenario and statistical forecast intervals(model based and empirical based).Hamilton-Pery method, which is a regression based approach, was used on four states(US) and nine test points to determine the factors affecting forecast accuracy.

Samir K.C et.al proposed a hybrid model for population projection based on age, sex and education [24].Firstly, Cohort-component method was used to project population by sex and age group. Then, this two dimensional projection was converted to three dimensional projection by adding level of education.

David A. Swanson used rescaled version of Mean Absolute Percent Error(MAPE) for evaluating cross-sectional ,sub national forecasts .He suggested that rescaled version should be used rather than MAPE as it preserves the useful information about error structure when substantial outliers exists.

3. Statistical Techniques for forecasting

Once the present and past data of population of an area is collected from the census, various methods can be used for predicting the population at end of the designed period based on the growth pattern of that area. the methods are as follows:

3.1 Arithmetical Increase Method

It is most suitable for large area with considerable developments. But if used for small, average or newer area, the estimated value will be less than actual value. From the past census data the average increase in population per decade is calculated. This value is then added to present population to find next decade population. This method takes in to consideration that population changes at a constant rate. Population after nth decade will be -

$Pn = P + n.C$ (1)

Where,

Pn is the population after 'n' decades

'P' is present population.

3.2 Geometrical Progression method

This method assumes that percentage increase in population remains constant from decade to decade. Since this method gives higher value, it is suitable for new area which is developing. Population at the end of nth decade (P_n) –

$P_n = P (1+ IG/100) n$

Where,

IG = geometric mean (%)

P = Present population,

n = no. of decades

3.3 Incremental Increase Method

This method is modification of arithmetical increase method and is suited for average sized area where population growth is in increasing order. The future population is calculated using increase in increment determined from past population and then the average value as well as average rate of increase is added to the present population. Population after nth decade is

$Pn = P+ n.X + \{n (n+1)/2\}.Y$

Where,

Pn = Population after nth decade,

X = Average increase,

Y = Incremental increase

3.4 Graphical Method

In this method, a graph is plotted on suitable scale for the population of last few decades. The curve thus obtained is smoothly extended to predict future population. This method requires proper experience and judgment. The best way to extend the curve is to compare the population curve with the growth curve of similar cities.

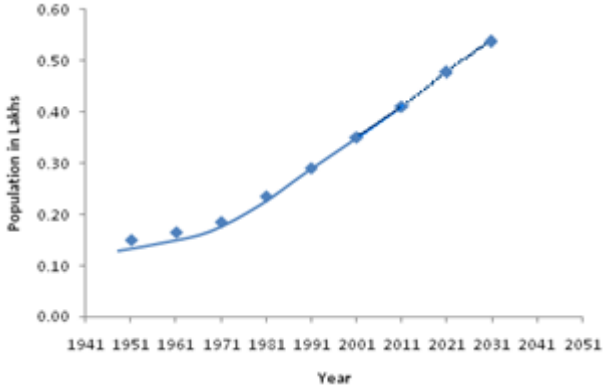

Figure 1: Population growth over years using graphical method [18]

3.5 Comparative Graphical Method

In this method, graphs are plotted for population of area already developed under similar condition and on the same graph curve for the past population of area under consideration is plotted. This curve is further extended by comparing it with the curve of other similar areas with similar growth patterns.

3.6 Logistic Curve Method

This method is useful when population growth occurs under normal situation and is not affected by extraordinary changes like epidemic, war or natural disaster. When population of an area under normal condition is plotted with respect to time, it forms S shape and thus called logistic curve.

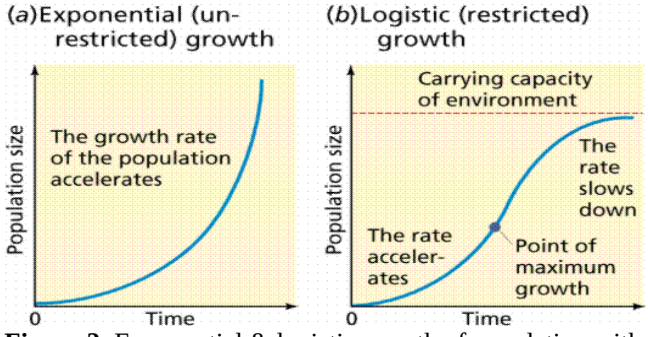

Figure 2: Exponential & logistic growth of population with time [18]

4. Proposed Methodology

4.1 Pre-Processing

Data on which we are going to work need to be analyzed before making predictions. Preprocessing of data involves:

4.1.1. Sampling
We need to determine the number of time periods (year) for which data will be evaluated.

4.1.2 Outlier Detection
Another consideration while making prediction is to find the presence of extreme observation or outlier if any.

4.1.3 Interpolation or Extrapolation
When inters censual data are available, values for intermediate years can be found using linear interpolation or extrapolation. This requires the calculation of average annual percentage change.

4.2 Correlation

This is used to find relationship between the literacy rate and demographic features (like population growth, male-female ratio) that affects literacy rate.

4.3 Prediction

The different methods for prediction are as following:

4.3.1 Multiple Regressions
Through this method, literacy rate is predicted as a function of one or more independent variables. Independent variables are estimated independent of the regression equation. The value of log(Y) the dependent variable (Y) is predicted using mathematical equation for a straight line on the basis of the independent variable (X).

$Y = a + b1X1 + b2X2 + biXi + e$
Where,
a – Y-intercept, is the expected value of Y when X=0
b – Regression coefficient for X i.e amount by which Y changes for unit change in X
e – Error term, difference between actual value and predicted value

4.3.2 Statistical based method
Based on the growth pattern of features affecting literacy rate based on the censual data available, one of the statistical method mentioned above will be used to predict the demographic feature. Study of dataset reveals that population is growing in increasing order, so we will use Incremental increase method to predict population growth in coming years.

4.4 Accuracy Measure

Root mean square error is used to measure the difference between value predicted by model and the actual value. If we represent actual data by A_t and forecasted value as F_t and n representing number of forecast made then root mean square error(RMSE) is given by :
$RMSE = \text{sqrt}\ (1/n.\ \sum(A_t - Ft)^2)$

5. Conclusion

Based on the study of various methods for prediction and forecasting and their application on various areas, we reached to a conclusion that we can achieve improvement in the forecasting accuracy by using combination of methods rather than using any single technique. So to predict the literacy rate we will find correlation of literacy rate with other demographic attributes and then these individual attribute prediction need to be done.

References

[1] David A. Swanson. 2013. Measuring uncertainty in population forecasts: A new approach. UNECE
[2] Sven F. Crone. 2011. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction, Elsevier
[3] Md. Mahsin and Syed Shahadat Hossain, .2012 , Population Forecasts for Bangladesh, Using a Bayesian Methodology, Journal of health ,population and nutrition(JHPN) , 30(4): 456–463.
[4] N.Rajasekhar and Dr. T. V. Rajini Kanth. July– 2014. Hybrid SVM Data mining Techniques for Weather Data Analysis of Krishna District of Andhra Region. International Journal of Research in Computer and Communication Technology, Vol 3, Issue 7
[5] Ryan W. Kirk, Paul V. Bolstad, Steven M. Manson. 2012. Spatio-temporal Trend analysis of long-term

development patterns (1900–2030) in a Southern Appalachian County- Landscape and Urban Planning. Elsevier-104 47– 58

[6] Sitaram Asur and Bernardo A. Huberman . 2010. Predicting the Future with Social Media. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 Washington, DC, USA Pages 492-499

[7] Harshavardhan Achrekar , Avinash Gandhe ,Ross Lazarus , Ssu-Hsin Yu , Benyuan Liu . 2011. Predicting Flu trends using Twitter data. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai

[8] Adrian E. Raftery, Leontine Alkema and Patrick Gerland . 2014. Bayesian Population Projections for the United Nations. Statistical Science 2014, Vol. 29, No. 1, 58–68 DOI: 10.1214/13-STS419 Institute of Mathematical Statistics.

[9] Neda Sadeghi, P. Thomas Fletcher, Marcel Prastawa, John H. Gilmore and Guido Gerig . 2014. Subject-specific prediction using nonlinear population modeling-Application to early brain maturation from DTI. Springer International Publishing Switzerland

[10] Jagdish Prasad and Nand Kishore Rawat . 2007. Trends and Forecasting the Estimation of Different Contraceptive Needs in the Districts of Rajasthan. Health and Population-Perspectives and Issues, 30(2):107-123

[11] Malathi. A, Dr. Santhosh Baboo. May 2011. An Enhanced Algorithm to Predict a Future Crime using Data Mining. International Journal of Computer Applications (0975 – 8887) Volume 21– No.1,

[12] Jinn-Yi Yeh, Tai-Hsi Wu, Chuan-Wei Tsao. 2011. Using data mining techniques to predict hospitalization of hemodialysis patients. Elsevier Decision Support Systems, 50 439–448

[13] F.Martínez-Álvarez, A. Troncoso1, A. Morales-Esteban, and J.C. Riquelme. 2011. Computational Intelligence Techniques for Predicting Earthquakes. HAIS, Part II, LNAI 6679, Springer-Verlag Berlin Heidelberg. 287–294

[14] Jeffrey K. McKee, Paul W. Sciulli, C. David Fooce, Thomas A. Waite. 2013. Forecasting global biodiversity threats associated with human population growth. Elsevier Biological Conservation, 115 161–164

[15] Friedrich Huebler, Weixin Lu. 2013. ADULT AND YOUTH LITERACY- National, regional and global trends, 1985-2015. UIS Information Paper

[16] Warren C.Sanderson. 2008. Knowledge can improve Forecasts-A review of selected socioeconomic population projection model. Population and development review, vol24, supplement, Frontiers of population Forecasting

[17] David A. Swanson, Jeff Tayman. 2013. Measuring uncertainty in population forecasts: A new approach. UNECE Work Session on Demographic Projections organized in cooperation with Istat (Rome, Italy)

[18] Population Forecasting – NPTEL IIT Kharagpur Web courses

[19] Jaoa Luiz Maurity Saobia. 1977. Autoregressive Integrated Moving Average (ARIMA) models for birth forecasting. Journal of American statistical association, pp. 264-270

[20] Alho J, Spencer B. 2005. Statistical demography and forecasting. New York, NY: Springer Series in Statistics

[21] S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S,Jimoh, R. G. 2011. Stock Trend Prediction Using Regression Analysis – A Data Mining Approach. ARPN Journal of Systems and Software

[22] Samir K.C., Wolfgang Lutz, Warren Sanderson, Sergei Scherbov , Erich Striessnig. October 2013, Rome, Italy. Result of the New Wittgenstein Centre Population Projections by age, sex and level of education for 171 countries. Joint Eurostat/UNECE Work Session on Demographic Projections organized in cooperation with Istat

[23] Felix Salfner, Maren Lenk, and Miroslaw Malek. 2013. A Survey of Online Failure Prediction Methods. ACM Journal Name, Pages 1–68.

[24] Stanley K. Smith. 2013. Further thoughts on simplicity and complexity in population projection models. Elsevier International Journal of Forecasting

[25] Saigal S. and Mehrotra D. 2012. Performance Comparison of time series data using predictive data mining techniques. Advances in Information Mining Volume 4, Issue 1, pp.-57-66.

[26] David A. Swanson, Alan Schlottmann, Bob Schmidt. 2010. Forecasting the population of census tracts by age and sex .Springer Population Research and Policy Review, Volume 29, Issue 1, pp 47-63,