FORECASTING OF LITERACY RATE USING STATISTICAL AND DATA MINING METHODS

[1]SWATI JAIN, [2]NITIN MISHRA

[1]M.Tech Scholar CSE dept RCET Bhilai India,
[2]Associate Professor CSE Dept RCET Bhilai India
E-mail: [1]js.jainswati@gmail.com, [2]drnitinmishra10@gmail.com

Abstract— Chhattisgarh state ranks 16th position in terms of population in India but it ranks 27th in terms of literacy rate. This backwardness is due to more tribal areas in this state. Since literacy is the basis for human development which in turn affects country's economic growth, it becomes important to improve the proportion of literates. Forecasting of literacy rate will help government & policy makers to make plans for improving the literacy rate. Number of literates in an area depends on various features like population of the area, density, type of area-rural or urban, age composition etc. In this research paper, total Population, male and female population is projected using a statistical method - logistic curve method. Using these projected populations, literacy rate is forecasted using a data mining method - multiple regressions. Using this combinational method, we forecast that the total population of Chhattisgarh in 2021 will be around 3 crores and the literacy rate will reach 89%. The accuracy of our combinational method is 92.69% which is much better than the results obtained from one of the proposed method - Global Age-specific Literacy Projections Model (GALP) for forecasting of literate and illiterate population.

Index Terms— Forecasting, Literacy Rate, Logistic Cuurve Method, Multiple Regression, Population Projection

I. INTRODUCTION

Forecasting means expected outcome in the near future. Literacy is the ability to read and write with the understanding of short and simple statements encountered in one's everyday life. Based on this definition of literacy, population can be categorized in two groups – Literate and Illiterate. Literate and illiterate population is characterized mainly by age (youth and adult), sex (male and female) and location (rural and urban). Since literacy plays a vital role for education and development, early formulation of literacy goals are required. Large number of policies is formed in almost every country which focuses on increasing literacy rate by encouraging school enrollment among children's. These policies may make progress but it will take decades to get significant results from such policies. Education for all(EFA) [14] had targeted to halve the adult illiteracy rate by 2015 but recent education system which is struggling to accommodate the growing population, marginalization of women and girls, maturing illiterate youth are hinders in achievement of the goals. Thus it becomes important to develop projection model that will help to understand the future scenario. Projection gives a picture of what future will look like, based on the past and assumptions about the future. Projection acts as the indicator of change over time, helps in evaluating impact of any policy or program by comparing the feature before and after the occurrence of an event. The demographic projections are based on the basis of projected population for particular area and for particular period. Population projection can be considered as an exercise consisting of calculation of population size in future. The past and present population record for any area required for projections can be obtained from census data. In India census are carried out after every 10 years. Census forms the basis for monitoring the past decade, understanding the current government policies and based on these plan the future. Census data play vital role for government planners and policymakers to predict the future requirement for food, water, energy, services. These predictions are useful to researchers, governments and various organizations for planning purpose, social and health research, for monitoring development goals and also for forecasting of various demographic characteristics. Thus population projections are not only useful for demographers but also for economist. Almost every country carries out census to collect different features of each geographic region and to have count on population in each region. Literacy statistics is collected using household surveys including Demographic and Health survey(DHS) and multiple indicator cluster survey(MICS).These surveys are not accurate way of collection and may lead to overestimation of literacy rate since the ability to read and write is self reported by respondents of survey.

A. Census of India

India's census data is the one of the most important resource of information related to demography, literacy, housing, economic activity, urbanization, fertility and mortality rate and many more features. Since 1871 census of India has been conducted every 10 years. Till 2011 decennial census has been conducted 15 times. Around 300BC India's population was around 100 – 140 million and then this statistics reached 225 million according to the 1881 census.[6] This was due to fall in mortality rate caused by factors

like immunization, better living conditions, improved nutrition, health care etc. According to 2011 census, this statistics has reached to 121 crores comprising of 62.31 crores males and 58.74 crores females. Literacy rate of India is 74% in 2011 census which has increased from previous census literacy rate of 64.83 percent. Literacy rate of male population being 82.14% is higher than that of female population which is 65.46%.

B. Census of Chhattisgarh

Chhattisgarh is situated at the centre of India and was separated out from Madhya Pradesh on 1st November; 2000.It is 10th in India in terms of area which is 135,190 sq-km [7]. According to 2011 census, Chhattisgarh population is 25,545,198 out of which 12,832,895 are male and 12,712,303 female which positions it as 16th populated state of India. Total population of Chhattisgarh in 2001 was 20,833,803 in which total males were 10,474,218 and total females were 10,359,585. Thus decadal population growth from 2001 to 2011 was recorded as 22%. In terms of literacy Chhattisgarh ranks 27th in India which is a bad situation .This backwardness is due to the more tribal areas which are less educated. Literacy rate increased from 64.66% in 2001 to 71.04% in 2011.

C. Forecasting Methods

Forecasting means to predict some events in future based on the study and analysis of available related data. Forecasting requires availability of accurate and timely data and one needs to understand the factors that affects the growth or decline of any feature in past. Plotting graph of variables against time helps to visualize changes over time. Extrapolation is the simplistic method for forecasting using historical data [9]. Extrapolated forecasting assumes constant increment, constant percentage change, average compounded growth or linear/nonlinear time trends. Simplistic models defined for population projection are arithmetic increase method, geometric progression method, incremental increase method and logistics method, graphical and comparative graphical method [10]. Any one of these methods can be used for population projection based on the growth pattern of the area in which projections are to be done.

Apart from these statistical methods, data mining methods are also available for predictions. Data mining offers possibilities for extracting hidden knowledge from even huge size of database. A variety of analyzing tools has been specified to discover patterns and relationship among data which helps in making predictions. Two commonly used models for data analysis are classification and prediction [13]. Classification model uses predefined set of classes for classifying the unknown objects. These classifications are based on the patterns observed in the training dataset. On other hand prediction models continuous valued function. The unknown or missing values are predicted using the prediction model being constructed. Regression is the major technique used for prediction. Regression technique is used to find relationship among variables. There are many different types of regression model available for prediction such as linear regression, non-linear regression, multivariate etc.

D. Objective of the study

The main purpose of this study is to forecast literacy rate of Chhattisgarh, one of the state in India. Literacy being the basis for both human and country's development requires plans to encourage children and youths for gaining education. Because of its much importance in country's progress it becomes crucial to develop projection model that forecast the literacy among people. The forecast of literacy rate directly depends on population projection. Thus, we will develop a model for forecasting of literacy rate which will be helping government to make educational plans. Population projection is done using logistic curve method. Based on this projection of population, literacy rate is forecasted using multiple regression method.

II. RELATED WORK

In this section, we discuss on a few important works that are closely related to our work as discussed. Dr. Wolfgang Lutz et.al [11] proposed GALP model for forecasting of literacy rate. He used cohort component method for population projection and uses these values to project literate and illiterate population. This model is data intensive as cohort component method itself requires data about fertility, mortality and migration. Literacy rate forecast using GALP model also requires age and sex specific proportion literates and the absolute numbers of literate men and women as input. It's difficult to get all these information for every area.

Dr R. Ravichandran [1] projects total population of India and one of its state Tamil Nadu using logistic curve method. He concluded that the population projection using logistic curve method gives reliable and accurate results.

Sitaram Asur et al. [2] used the chatterbox of a social site Twitter to forecast the box office revenues for different movies. A linear regression model was used for the prediction

Harshavardhan Achrekar et al. [8] used auto regression method on twitter data 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.

Adrian E. Raftery et al.[3] found that the most popular method, cohort component method, for population projection does not yield an estimation of uncertainty

about future population quantiles. To overcome this problem he proposed Bayesian probabilistic population projection model. In this model the input to cohort component method, total fertility rate and life expectancy, are predicted using Bayesian method.

F.Martinez Alvarez et al. [5] 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.

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.

III. METHODOLOGY

This section explains the work done. The work analyzes the census data collected online from website censusindia.gov.in. Methodology used in our study is discussed in detail as well as summarized in the fig. 1, as shown below.

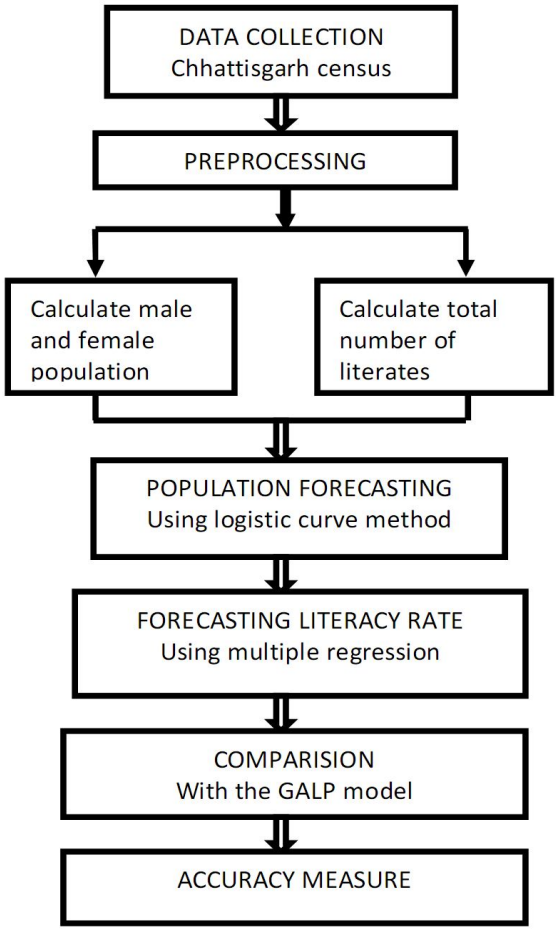

Fig 1: Methodology

E. Data Collection

Data was collected from India census website and also from indiabudget.co.in. This site contains census data for different states of India. We collected data for Chhattisgarh only. The data being collected contains the population of Chhattisgarh for last 12 census i.e. 1901-2011 and literacy rate is available for only 7 census i.e. 1951 - 2011. The censual data contains information about various demographic features like total population, population density, sex ratio, decadal growth rate, population by caste, working population, literacy rate etc.

F. Data Preprocessing

We are concerned with the forecasting of literacy rate which is mainly based on the number of persons of the area. Thus we select required attributes from the collected dataset for our research work. Selected attributes in our work are total population, sex ratio and literacy rate.

We have total population value from year 1951 to 2011 but we do not have separate male and female population for all these years. So male and female population are calculated using values of total population and sex ratio using (1) and (2). Sex ratio is the number of females per thousand male.

$$Malepop = \frac{Totalpop}{1 + \frac{Sexratio}{1000}} \quad (1)$$

$$Femalepop = Totalpop - Malepop \quad (2)$$

Similarly we also need to find number of literates during each census which will help us to forecast literacy rate. Literacy rate is calculated as the percentage of total number of literates divided by total population .Thus, total number of literates can be calculated using (3) where Litpop represents number of literates, Litrate is the literacy rate and Totalpop represents total population.

$$Litpop = \frac{Litrate * totalpop}{100} \quad (3)$$

G. Population forecasting using logistic curve method

Many mathematical or statistical methods are available for population projection but we choose logistic curve method because this method considers that the population grows under normal circumstances and is not affected by extraordinary changes like war, earthquakes. The plot of population of an area with respect to time using this method gives S-shaped curve. This shape shows that initially population grows at an geometric growth rate and once it reaches a saturation point it follows declining growth rate. The saturation point is due to the limited resources and economic opportunity.

Population at some time in future p_t is calculated using (4).

$$pt = \frac{psat}{1 + e\hat{}(a + b.dt)} \quad (4)$$

$$psat = \frac{2P0P1P2 - P1\hat{}2(P0 + P2)}{P0P2 - P1\hat{}2} \quad (5)$$

$$a = \ln \frac{Psat - P2}{P2} \quad (6)$$

$$b = \frac{1}{n} \ln \frac{P0(Psat - P1)}{P1(Psat - P0)} \quad (7)$$

Where,

Pt = Population to be forecasted
Psat = Population at saturation level
P0 = Base population
P1, P2 = Population at two time periods
n = time interval between successive census
dt = number of years after base years

H. Forecasting Literacy rate using multiple regression

Regression model estimates the independent variable as a function of one or more independent variables. It uses a mathematical equation of straight line to predict the value of independent variable based on the dependent variable. In our work, literacy rate is the independent variable that is to be forecasted and the dependent variables on which literacy rate depends are total male population and total female population.

$$Y = a + b1X1 + b2X2 \quad (8)$$

Where,

a:– Y-intercept, is the expected value of Y when X=0
b :– Regression coefficients for X

I. Comparison with the G ALP model

GALP model projects number of literate and illiterate population based on the national population projection using cohort component method. Based on the size of projected population GALP model uses logistic regression to forecast the literacy rate.

Actual literacy rate, our predicted literacy rate and literacy rate calculated using GALP model is shown in fig 2. It is evident from the graph shown in fig 2 that literacy rate predicted using our method is very close to actual method.

J. Accuracy Measure

Accuracy can be defined as the degree of correctness of variable or the closeness of the predicted value with its actual value. Accuracy measure is important as it helps to understand the appropriateness of the model.

The accuracy is measured using mean absolute percentage error (MAPE). MAPE expresses the error as percentage of actual values of the data. It is calculated using (9).

$$MAPE = \frac{\sum \left| \frac{At - Ft}{At} \right|}{n} * 100 \quad (9)$$

Where,

At is the actual value
Ft is the measured or predicted value
n represents total number of observations

MAPE calculate for values obtained from our combinational method is 7.3059, which indicates its accuracy is 92.8% whereas MAPE for Values using GALP model is 42.792 and thus the accuracy of GALP model is only 57.208%.

IV. RESULTS AND DISCUSSION

Using the actual population values for six censuses i.e. from 1901 to 1951 in the logistic curve method we calculate the population from census year 1961 to 2001. The actual and predicted populations for these years are shown below in Table I.

Similarly, male and female population is predicted using logistic curve method which takes male and female population for census 1901 to 1951 and forecast for census years 1961 to 2001. The actual and predicted male and female populations are shown below in Table II.

Table: I. Actual and forecasted values of population using logistic curve method

Year	Actual Population	Predicted Population
1961	9154498	9189477
1971	11637494	10912706
1981	14010337	13238590
1991	17614928	16541763
2001	20833803	21588429
2011	25545198	30228947

Table: II. Actual and predicted values of male and female population using logistic curve method

Year	Actual male population	Predicted male population	Actual female population	Predicted female population
1961	4559000	4557108	4595000	4634926
1971	5825000	5443775	5813000	5475571
1981	7019000	6659241	6991000	6595398
1991	8874000	8423603	8741000	8157098
2001	10475000	11209766	10359000	10480186
2011	12824000	16252664	12721000	14289985

These results show that the predicted values of total population, male population and female population are very close to their actual values. Using multiple regression equation and the values of male and female population predicted as the independent variables of regression equation literacy rate is forecasted from census 1961 to 2011.The actual and predicted literacy rate are shown below in Table III.

Table: III. Actual and predicted literacy rate using multiple regressions

Year	Actual Literacy Rate (%)	Predicted Literacy Rate (%)
1961	18.14	18.0709
1971	24.08	24.0578
1981	32.63	32.4001
1991	42.80	44.5802
2001	64.61	63.6760
2011	71.02	97.33

These results show that the predicted values of total population, male population and female population are very close to their actual values. Using multiple regression equation and the values of male and female population predicted as the independent variables of regression equation literacy rate is forecasted from census 1961 to 2011.The actual and predicted literacy rate are shown below in Table III.

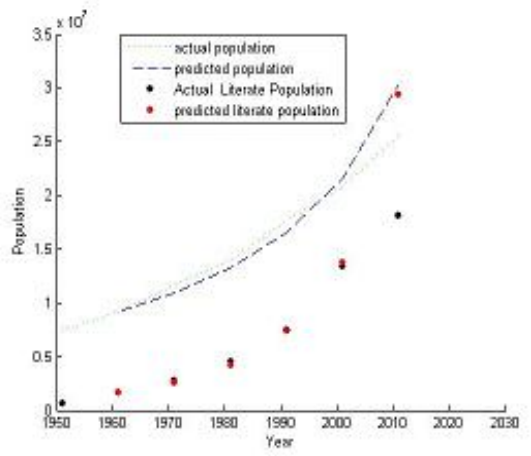

Fig.2: Actual and predicted total population and literate population

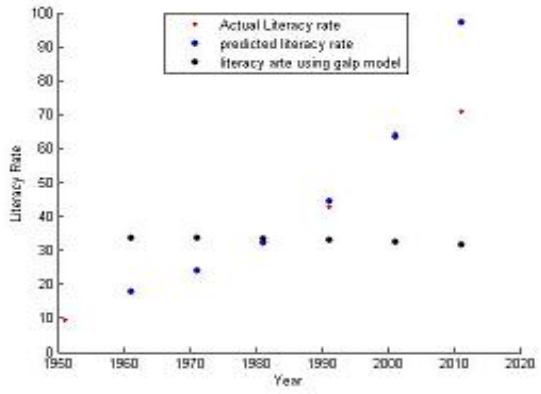

Fig.3: Actual, predicted and compared literacy rate

Figure 2 is the graph showing plot of actual values and predicted values for total population and literate population. The actual and predicted growth of literacy rate is presented graphically in figure 3. It also compares the literacy rate forecasted using GALP model.

From figure 2 and 3, we can see that predicted data points are very close to actual data points which indicate that error is very small. The graph shows the exponential growth of the attributes which means they are at the initial stage of logistic curve.

CONCLUSION

We conclude that the literacy rate mainly depends on the population size of the area. The population growth of Chhattisgarh well fits to the logistic curve method and is currently at its initial stage i.e. growing stage. Literacy rate calculated using multiple regressions also gives results very close to the actual literacy rate given by the census of India. Moreover, this combinational method for literacy rate forecasting would be useful to researchers who work with the demographic features and work related to it.

SCOPE FOR FUTURE WORK

Effective literacy rate can be calculated taking in to account the age composition. Effective literacy rate excludes population with age group 0-7 and hence it gives more accurate figures. Even it can be separated based on the rural and urban areas. Moreover, if more independent variables affecting literacy rate like income status, number of schools are included in multiple regression model it may be beneficial.

ACKNOWLEDGEMENT

I feel very pleased to thank all the supporting hands, who helped me in my ambitions work. I want to show my gratitude to my respected supervisor Dr Nitin Mishra for his guidance, help, support and encouragement. I am thankful to Prof. Toran verma, M.Tech Coordinator (CSE) for giving thoughtful suggestions during my work.

I owe the greatest debt and special respectful thanks to Mr. Santosh Rungta Sir, Chairman, Dr. Sourabh Rungta, Director(Tech.), Mr. Sonal Rungta, Director(Finance), Dr. S. M. Prasanna Kumar, Director, Rungta College of Engineering and Technology, Bhilai for their inspiration and constant encouragement that enabled me to present my work in this form.

REFERENCES

[1] Dr R. Ravichandrran," A Study on Population Projection using the Logistic Curve method in Time series analysis with reference to India", Indian Journal of Applied Research, Volume 3, May 2013

[2] Sitaram Asur and Bernardo A. Huberman," 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, 2010

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

[4] N.Rajasekhar and Dr. T. V. RajiniKanth. "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, July– 2014

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

[6] Wikipedia for Demographics of India, en.wikipedia.org/wiki/Demographics_of_**India**

[7] Chhattisgarh census data, http://www.census2011.co.in/

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

[9] Deborah Rosenberg, "Trend Analysis and Interpretation", Maternal and Child Health Information Resource Center, HRSA, PHS, DHHS , Dec 1997

[10] Population Forecasting – NPTEL IIT Kharagpur Web courses

[11] Dr. Wolfgang Lutz and Dr. Sergei Scherbov," Global Age-specific Literacy Projections Model (GALP): Rationale, Methodology and Software", Montréal (Québec), Canada, UNESCO Institute for Statistics (UIS),July 2006

[12] Augustus Wali, Epiphanie Kagoyire, Paci¯que Icyingeneye, "Mathematical Modeling of Uganda Population Growth", Applied Mathematical Sciences, Vol. 6, no. 84, 4155 - 4168, 2012

[13] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao," Data mining techniques and applications – A decade review from 2000 to 2011", Elsevier, Expert Systems with Applications 39, 11303–11311, 2012

[14] ADULT AND YOUTH LITERACY, National, regional and global trends, 1985-2015, UIS information paper, June 2013.

★★★