IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Factor Analysis of Population Growth using Data Analytics

Anupama Girish[1*] , Aditya Dey[2], Ankit Sharma[3], Ketan Jain[4],Kumar Sanket[5], Amutha S.[6], Ramesh Babu D.R. [7]

[1,2,3,4,5,6,7]Dept. of CSE, Dayananda Sagar College of Engineering, VTU University, Bangalore, Karnataka, India

Corresponding Author: anupamayk@gmail.com

Available online at: www.ijcseonline.org

Accepted: 25/Sept/2018, Published: 30/Sept/2018

Abstract**—**According to the estimation billionth person was born in 1804 and the second billionth was born about 123 years later in 1927. Since then it has taken humans 60 years to reach the 5 billion mark and now we are closer to population of 8 billion. India contributes about 20% of this population making it as the second most populous country in the world. Originally, most of the important predictions were made using the Malthusian growth models. The science of data analytics has opened up new possibilities in the creation of prediction graphs. Prediction graphs give useful information about tackling the problem of increasing population. R programming language is used to identify factors that impact the rate of change of population**.** Important factors such as literacy rate, death rate, religion and so on, deeply impact the rate of population growth. From the Kaiser-Meyer-Olkin test and Factor Analysis found that out of all factors that were considered, religious differences and migration rate were the most important factors affecting the rate of population growth.

Keywords**—**Malthusian growth model,Data Analytics, Factors, R-Programmimg language, Kaiser-Meyer-Olkin

I. Introduction

Population standing at approximately 1.4 billion, India currently holds a fifth of the world's numbers. In the last three decades, the number has doubled to the value it is right now, making it safe to say that India's population has been increasing at an alarming rate. Successfully predicting the future values of Indian population can lead to an in-depth grasp of why this expansion has occurred. The factors behind it can be used to derive ways and means to perhaps control population growth.

Population growth models have been used in the past to predict future values of a country's population. Initial models assumed that the population of a country grew at a constant rate exponentially. This led to the development of formulae that could calculate future statistics using the value of the current population. These models were later revised to assume other factors besides the rate of growth of population, like restrictions on environmental resources and the maximum land carrying capacity of an area. Some more factors which may include birth rate, infant mortality rate, healthcare availability and so on can also be accounted for using such formulae. Once a model has been created, it can be used to predict the future population of India.

Factor analysis has been performed on the Indian population values to determine the factors or a list of principal components that have affected this growth. Factor analysis is the process in which the values of observed data are expressed as functions of multiple possible causes to find which are the most important. A possible list of factors can be assumed, namely birth and death rate, immigration rate, infant mortality, reproduction rates and likewise. The data for these factors would then have to collected along with preparation of data for growth analysis. The data obtained can then be used to perform factor analysis. This would result in confirmation that those factors are the main ones causing the growth following which some steps could be taken to implement laws or policies that would result in a decline in the growth of India's population.

The data obtained through the population growth analysis and the factors obtained through factor analysis can be combined to form an understanding of the state of India's population. India as a country might not be able to provide the sufficient amount of resources required for its people if the rate of growth stays as it is. It is imperative that some steps be taken to curb or control the rate of growth of people in our country. The implementation of laws around family control and incentivizing people to have lesser children has been considered.

A detailed understanding of the current situation of India's population requires a two-part model that includes population growth prediction to realize the future values and a factor analysis to conclude the principal components behind the alarming expansion.

Rest of the paper is organized as follows, Section I contains the introduction of population growth prediction and factor analysis, Section II contain the related work of factor analysis Section III contain the methodology and proposed system, Section IV contain the results and discussion, section V concludes research work with future directions.

II. RELATED WORK

Previous work into the analysis of India's population growth had been conducted by Dr. Samir Vohra [1] who stated that in terms of population India is the second largest country in the world. A high birth rate and a decline in the death rate has led to a rapid increase in population of India. A factor analysis had been performed by Suriani Hassan et al [2] who conducted a study on constructing the factors affecting students learning styles using factor analysis. To analyse the demographic differences on the new factors affecting students learning styles comparison means using the Kruskal - Wallis test was done. A survey questionnaire was prepared to collect data. The test results showed there was a significant difference between gender on students efforts outside class while there was no significant mean difference between genders on the other factors of students learning style and it was found after few years of study that students attitude before and after attending class influenced learning style.

III.METHODOLOGY

The Figure 1. shows the proposed system architecture. It includes data sources that are to be imported into R and then analysed as required. A KMO test is also performed on the datasets to confirm that there is minimum linearity required to perform satisfactory factor analysis.

In order to perform factor analysis on India's population data, an approximation must be made of India's current population. The census data obtained is used with predictive models and population growth models to obtain the required number. Factor analysis is then performed on state vs. time data obtained from the census. Each state's population data is arranged in the state vs. time format for each factor chosen to be analysed. The list of factors chosen include birth rate, death rate, literacy rate, migration rate, sex ratio and fertility rate. The datasets are initially cleaned and missing values are interpolated or extrapolated.

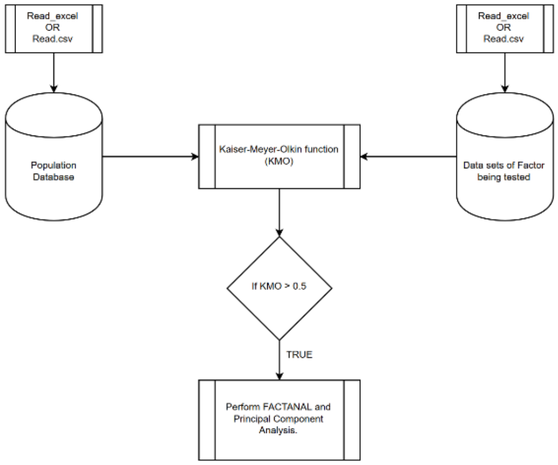

Read_excel
OR
Read.csv
Read_excel
OR
Read.csv
Population Database
Kaiser-Meyer-Olkin function (KMO)
Data sets of Factor being tested
If KMO > 0.5
TRUE
Perform FACTANAL and Principal Component Analysis.

Figure 1: Proposed system

Once the entire dataset is populated, it is put through a Kaiser-Meyer-Olkin(KMO) test to decide whether it is suitable for factor analysis or not and then factor analysis is performed to determine if these datasets had a significance on the overall population or not.

A. KMO Test

A KMO test is a statistical test that ensures that the dataset is suitable for factor analysis to be performed on it or not. If a dataset returns a KMO test value greater than 0.5, it is deemed suitable for factor analysis and if a value less than 0.5 is obtained, the dataset is rejected.

B. Factor Analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. The steps executed to perform factor analysis are as follows:

Step 1: Order the data according to the factor analysis function requirements with state data as rows and year-wise metrics as columns.
Step 2: Remove or extrapolate or interpolate any blank spaces or null values.
Step 3: Import data into R Studio using R.
Step 4: Create correlation matrix between the data sets.
Step 5: Use this matrix as input to the factor analysis function which provides us with a uniqueness of data variance in each column, a loading of each column onto the factors, and the correlation between the factors.

PCA (principal component analysis) allows us to divide the correlation matrix components which are later used to describe the impact of each column on the total variance of the data set. Once factor analysis has been performed, a significant value for each of the datasets is obtained which shows how significantly that dataset/factor has affected population growth. A higher significance value means that the dataset has affected overall population highly and must be considered as an important factor. On the other hand, lower significant values denote that while the factors do affect population numbers, their overall effect is much lesser than the other factors considered.

IV. Results and Discussion

The specifications of the computer used in the experiment are as follows:

a) CPU – Intel(R) Core(TM) i5-8250U CPU@ 1.8 GHz
b) Memory – 8 GB
c) Core – 8 core
d) HDD – 500 GB
e) OS – Windows 10

The software requirements are as follows:
a) R Programming Language Version 3.4
b) R Studio Version 1.1.423

The following four factors have been considered for testing:

1. Population Prediction: Population figures from each census over the last 10 decades have been used as input to predict population figures for the future using the forecast function in Excel. Prediction of the population has been done on the 2011 census data using MS Excel's built in FORECAST.ETS function which uses the AAA version of the Exponential Smoothing (ETS) algorithm. The simplest form of exponential smoothing is given by:

$$S_t = \alpha .x_t + (1-\alpha).S_{t-1} \qquad -(1)$$

In equation (1), α is the smoothing factor and S_t is a simple weighted average of the current observation x_t.

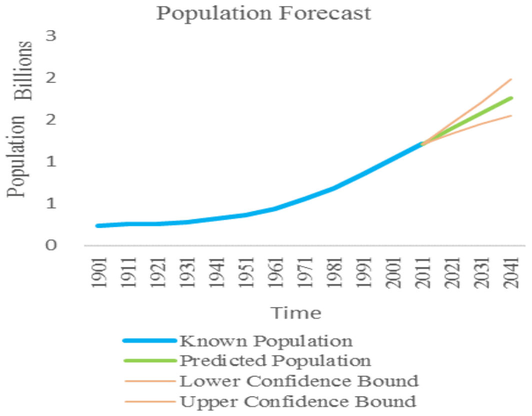
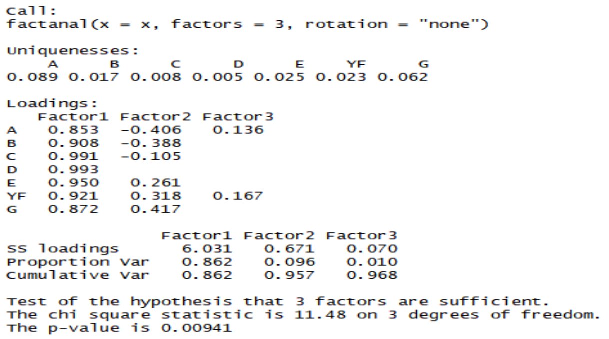


Figure 2: Population Prediction

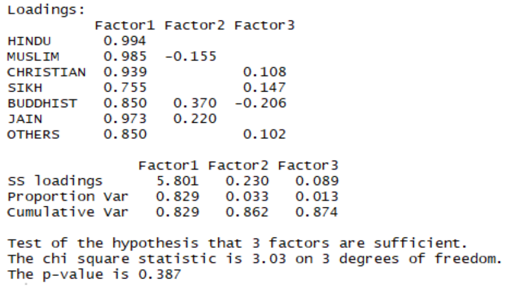
2. Literacy Rate: The increasing literacy rate has led to an increase in the population of India as can be seen from the significance value. It can be inferred that more people reproduce in order to provide more income for their family and an increase in literacy has led to an increase in population.



```
Call:
factanal(x = x, factors = 3, rotation = "none")

Uniquenesses:
    A     B     C     D     E    YF     G 
0.089 0.017 0.008 0.005 0.025 0.023 0.062 

Loadings:
   Factor1 Factor2 Factor3
A   0.853  -0.406   0.136 
B   0.908  -0.388         
C   0.991  -0.105         
D   0.993                 
E   0.950   0.261         
YF  0.921   0.318   0.167 
G   0.872   0.417         

               Factor1 Factor2 Factor3
SS loadings      6.031   0.671   0.070
Proportion Var   0.862   0.096   0.010
Cumulative Var   0.862   0.957   0.968

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 11.48 on 3 degrees of freedom.
The p-value is 0.00941
```

Figure 3: Literacy Rate

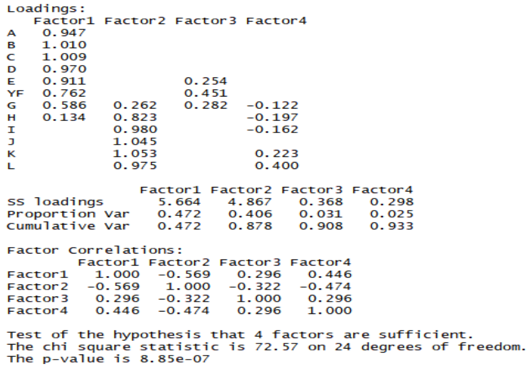
3. Religion: India has a variety of people from multiple religions residing within its borders. The difference of religion leads to an increase in population, seeing as people holding certain religious beliefs tend to reproduce more.



```
Loadings:
          Factor1 Factor2 Factor3
HINDU      0.994                 
MUSLIM     0.985  -0.155         
CHRISTIAN  0.939           0.108 
SIKH       0.755           0.147 
BUDDHIST   0.850   0.370  -0.206 
JAIN       0.973   0.220         
OTHERS     0.850           0.102 

               Factor1 Factor2 Factor3
SS loadings      5.801   0.230   0.089
Proportion Var   0.829   0.033   0.013
Cumulative Var   0.829   0.862   0.874

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 3.03 on 3 degrees of freedom.
The p-value is 0.387
```

Figure 4: Religion

4. Death Rate: A decline in the death rate due to better health care has resulted in a significant increase in the population numbers for India.

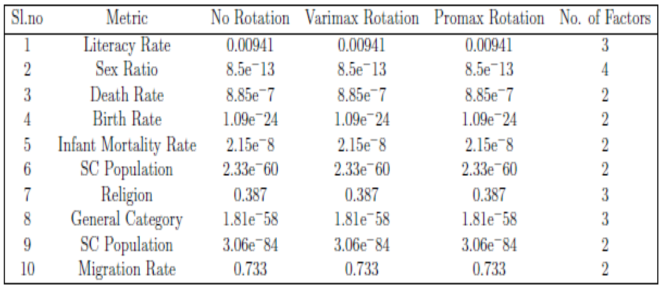


```
Loadings:
   Factor1 Factor2 Factor3 Factor4
A   0.947                         
B   1.010                         
C   1.009                         
D   0.970                         
E   0.911           0.254         
YF  0.762           0.451         
G   0.586   0.262   0.282  -0.122 
H   0.134   0.823          -0.197 
I           0.980          -0.162 
J           1.045                 
K           1.053           0.223 
L           0.975           0.400 

               Factor1 Factor2 Factor3 Factor4
SS loadings      5.664   4.867   0.368   0.298
Proportion Var   0.472   0.406   0.031   0.025
Cumulative Var   0.472   0.878   0.908   0.933

Factor Correlations:
        Factor1 Factor2 Factor3 Factor4
Factor1   1.000  -0.569   0.296   0.446
Factor2  -0.569   1.000  -0.322  -0.474
Factor3   0.296  -0.322   1.000   0.296
Factor4   0.446  -0.474   0.296   1.000

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 72.57 on 24 degrees of freedom.
The p-value is 8.85e-07
```

Figure 5: Death Rate

Once testing is carried out, the results obtained are collected and the factors with highest significance values are considered as the foremost factors affecting Indian population. The table below shows the results obtained.

Table 1: Experimental results with significant values



Sl.no	Metric	No Rotation	Varimax Rotation	Promax Rotation	No. of Factors
1	Literacy Rate	0.00941	0.00941	0.00941	3
2	Sex Ratio	$8.5e^{-}13$	$8.5e^{-}13$	$8.5e^{-}13$	4
3	Death Rate	$8.85e^{-}7$	$8.85e^{-}7$	$8.85e^{-}7$	2
4	Birth Rate	$1.09e^{-}24$	$1.09e^{-}24$	$1.09e^{-}24$	2
5	Infant Mortality Rate	$2.15e^{-}8$	$2.15e^{-}8$	$2.15e^{-}8$	2
6	SC Population	$2.33e^{-}60$	$2.33e^{-}60$	$2.33e^{-}60$	2
7	Religion	0.387	0.387	0.387	3
8	General Category	$1.81e^{-}58$	$1.81e^{-}58$	$1.81e^{-}58$	3
9	SC Population	$3.06e^{-}84$	$3.06e^{-}84$	$3.06e^{-}84$	2
10	Migration Rate	0.733	0.733	0.733	2

V. CONCLUSION and Future Scope

The rate of growth of population in India is a subject that needs to be discussed openly. There are many important factors such as literacy rate, death rate, religion and so on, which deeply impact the rate of population growth. From this study, it was learnt that out of all factors that were considered, religious differences and migration rate were the most important factors affecting the rate of population growth with significant values of 0.387 and 0.733 which tell us that these factors had a high impact on the population growth. These results can be improved with richer census data which would give more accurate insights. Future investigation in this field can be done by performing a deeper research on different population data sets and using more niche factors that allow a more defined look at what impacts rate of population.

REFERENCES

[1] Dr. Samir Vazidbhai Vohra, 'Population Growth-India's problem', PARIPEX – Indian Journal of Research, Vol. 4 , Issue : 11 , November 2015, ISSN - 2250-1991.

[2] Suriani Hassan, Norlita Ismail, Wan Yonsharlinawati Wan Jaafar, Khadizah Ghazali, Kamsia Budin, Darmesah Gabda and Asmar Shahira Abdul Samad , 'Using factor analysis on study of factors affecting students' learning style', International journal of Applied Mathematics and Informatics, Vol. 6, 2012.

[3] Lucas Gren, Alfredo Goldman, 'Useful statistical methods for human factors research in software engineering', ACM, p.p 121-124, 2016.

[4] Sanchita Patil,' Big data analytics using R ', IRJET, Vol. 03, Issue: 07, July-2016.

[5] Huabin Wei, Yanqing Jiang and Yuxing Zhang, 'A review of two population growth models', Asian Journal of Economic Modelling, 3(1): 8-20, 2015

[6] Shilpa S Kulkarni et al,' Analysis of population growth in India and estimation for future', IJIRSET, Vol. 3, Issue 9, Sep- 2014.

[7] Mohammed Yiha Dawed, Purnachandra Rao Koya, Ayele Taye Goshu,'Mathematical modelling of population growth: The case of logistic and Von Bertalanffy models', Open Journal of Modelling and Simulation, 2, 113-126, 2014.

[8] A. Alexander Beaujean, 'Factor analysis using R', Practical Assessment, Research & Evaluation, Vol. 18, Number 4, Feb-2013.

[9] Feng Zhiming et al, 'Natural environment suitability for human settlements in China based on GIS', Journal of Geographical Sciences, 19: 437-446 ,2009

[10] S,. Ramana, S. Sabitha, R. Senthil Kumar, T. Senthil Prakash "**Atmospheric Change on the Geographical Theme Finding Of Different Functions on Human Mobility",** International Journal of Scientific Research in Network Security and Communication, Vol.6, Issue.2 , pp.134-151, Apr-2018

[11] K. Sarmah, "Comparison Studies of Speaker Modeling Techniques in Speaker Verification System", International Journal of Scientific Research in Network Security and Communication , Vol.5 , Issue.5 , pp.75-82, Oct-2017

Authors Profile

Mrs. Anupama Girish completed Bachelor of Engineering from P.E.S. college of Engineering , Mandya in 2002, currently working as a Asssitant Professor at department of CSE, Dayanand Sagar College of Engineering affiliated to VTU university, Bangalore, Karnataka, India.

Mr.Adit Dey, Mr. Ankit Sharma, Mr.Ketan Jain and Mr. Sanket Kumar are pursuing Computer Science & Engineering from Dayanand Sagar College of Engineering affiliated to VTU university, Bangalore, Karnataka, India.

Dr. Amutha.S currently working as a Professor at department of Computer Sceine & Engineering, Dayanand Sagar College of Engineering affiliated to VTU university, Bangalore, Karnataka, India

Dr. Ramesh Babu. D.R currently working as a H.O.D. and Professor at department of Computer Sceine & Engineering , Dayanand Sagar College of Engineering affiliated to VTU university, Bangalore, Karnataka, India