Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2016/15738

Amalgamating Data Analytics and Machine Learning for Predicting Sex Ratio and Infant Mortality Rate to Improve Gender Composition

Amita Jain[1]*, Amishapriya Singh[2], Devendra Kumar Tayal[2] and Sonakshi Vij[2]

[1]Department of CSE, AIACTR, New Delhi – 110031, Delhi, India; amita_jain_17@yahoo.com

[2]Department of CSE, IGDTUW, New Delhi - 110006, Delhi, India; amisha.sh22@gmail.com, dev_tayal2001@yahoo.com, sonakshi.vij92@gmail.com

Abstract

Different parameters are used to obtain a detailed account of population compositionwithin a country or its states in terms of sex. These parameters include sex ratio, infant mortality rate, foeticide rate etc. In this paper, we analyse the trends in sex ratio of India and some of its states, the number of children ever born to a married women given that all are females and the number of surviving females is zero, infant mortality rate, child birth ratio, percent of boys more than girls. Through these parameters, we developed a prediction system for infant mortality rate and sex ratio of India, provided the current trends continue without substantial changes. We have collected the data from reliable government sources and used platforms of R programming for prediction purposes and RStudio for visualisation in order to present a visually appealing user interface. Although the data available to us was highly limited, our system was able to make predictions with an accuracy of 93%. Using the results, we have developed several conclusions and observation regarding the skewed gender composition in India and some of its states. Leveraging our observations, we have not only identified the problem areas but also tried to direct attention towards the policies/laws that have proved to be inefficient in alleviating this skewed sex ratio (in favour of males) in India.

Keywords: Infant Mortality Rate, Regression, R Programming, RStudio, Sex Ratio

Manuscript Accepted: 02-July-2016; **Originality Check:** 13-July-2016; **Peer Reviewers Comment:** 23-July-2016; **Double Blind Reviewers Comment:** 02-Aug-2016; **Author Revert:** 15-Aug-2016; **Camera-Ready-Copy:** 10-Sep-2016)

Introduction

Data Analysis has a dynamic field of application. It has been deployed in the fields of health services research[1], policy research[2], e-learning[3], landscape ecology[4], meteorology and oceanography[5], finance etc. Data analysis involves and covers the process of cleaning, examining, transforming, visualising and modelling of the available data in order to uncover interesting patterns or trends, make deductions, extract information and support decision-making. Here the objective is to understand the bigger picture to identify similarities and dissimilarities. One of the many techniques for performing data analysis is data mining[9]. It covers modelling and knowledge discovery for the purpose of making predictions instead of information. In this paper, we will be utilising the concepts of data mining[8] to a large extent.

This paper tries to identify the problem area for the continual skewed sex ratio in India and the failure of government policies and initiatives to alleviate the same. Sex Ratio is defined by the number of females per 1000 males in the population. We have also analysed population of girls and boys state-wise, Infant Mortality Rate, defined as the number of infant deaths before the age of one per 1000 live births, and ever married women by the number of children born. We have taken the case of all female children and zero survivors. Additionally, we have made predictions for sex ratio and infant mortality rate to provide and insight into the future if current trends continue. This will allow the government and the policy makers to directly view and compare the improvements and impacts that their policies bring.

For the purpose of implementations, we have used Rstudio and R programming. RStudio was used for all the analysis to present a better looking user interface and R was used for the purpose of predicitions. All the data have been collected from reliable government sources.

The remainder of the paper is organised with section 2 introducing the concepts that have been leveraged in this paper, section 3 outlining the implementation and results, section 4 covering the following discussions and finally section 5 concluding our work.

*Author for correspondence

2. Data Analysis Concepts

In this paper, we have used several data mining concepts[10]. The various steps followed for knowledge discovery from data[11] include:

i. Data cleaning: It involves removal of invalid entries and noise from data. Sometimes, this step may also involve bringing regularity in data definitions to introduce uniformity in the data set.

ii. Data integration: Data collected from various sources are combined and they are consolidated to remove ambiguity and introduce uniformity.

iii. Pattern mining: Useful and informative patterns are extracted from the data set with the help of exploratory tools such as modelling techniques.

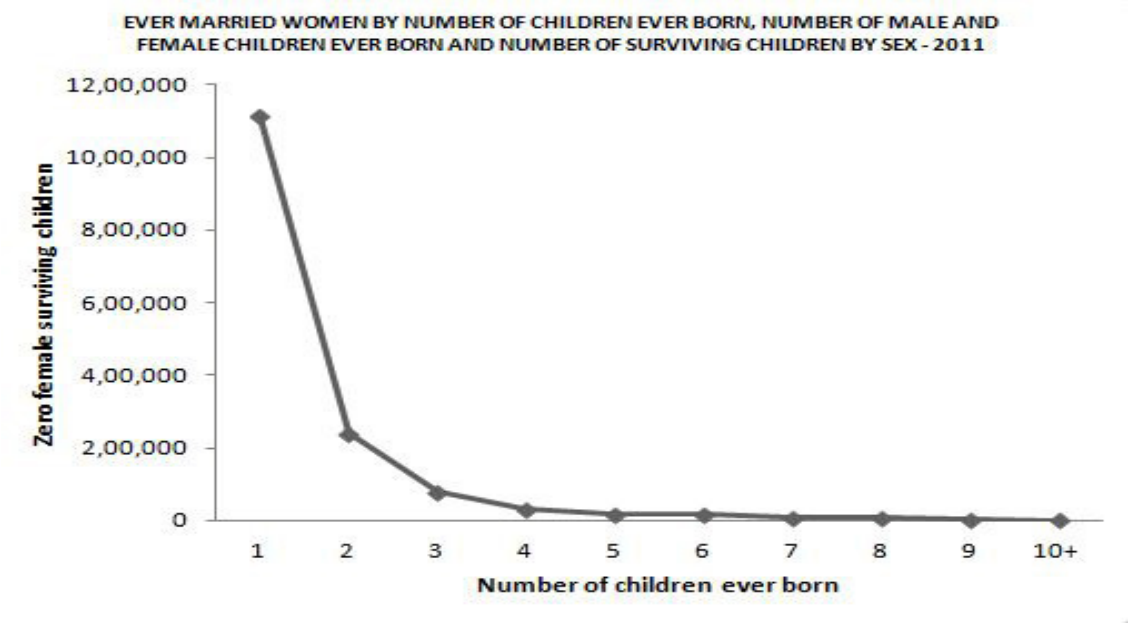

Figure 1. Ever married women by the number of children born assuming all are girls and the number of cases of zero survivors in India in both rural and urban settlements.

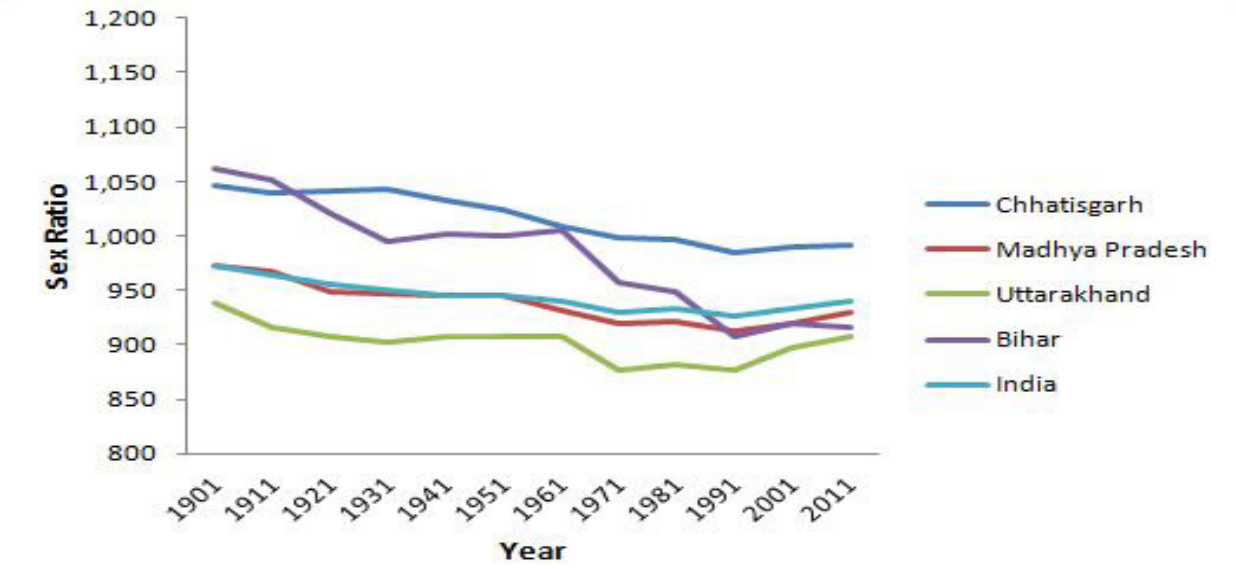

Figure 2. Sex Ratio in India and four of its sates-Chhatisgarh,

iv. Data Visualisation: It involves presentation of knowledge collected from the data set after applying modelling.

This whole procedure is also known as knowledge discovery from data (KDD)[13]. All the above steps were followed to prepare the data and extract useful patterns from the same. The following section highlights our implementation results.

For the purpose of prediction/modelling, we have used regression. Regression12 can be classified as:

a. Supervised[15]: As the name suggests, supervised regression uses supervision for regression i.e. labelled data to train and test the prediction model. It is ones of the easiest methods for implementing a prediction module. Inputs and outputs are used to find a function that generalises this behaviour.

Unsupervised7: Here, only inputs are provided and no outputs. The problem and individual knowledge help in prediction technique. It is similar to unsupervised classification.

Semi-supervised6: In semi-supervised technique, the data used is a mix of labelled and unlabelled. As labelled data is more expensive to generate, this method finds a trade off between supervised and unsupervised technique to achieve highest efficiency without compromising the accuracy or the ease of implementation.

In this paper, we have used supervised regression called linear regression[14]. Linear Regression is a statistical approach for modelling the relationship between a scalar dependent variable and one or more independent/explanatory variables. When there is only one independent variable, the technique is called simple linear regression. The data set available to us was such that it facilitated the use of supervised learning and we found it to be the most feasible and efficient option in our case.

3. Implementation and Results

The data collected from different sources were cleansed, integrated and then mined to identify useful patterns and trends. We have presented the different visualisation techniques that we used in our implementation for different types of data sets. The visualisation shows interesting patterns and allows us to draw important conclusions which are summarised in the following sections.

Figure 1 shows ever married women by the number of children born assuming all are girls and the number of cases of zero survivors in India in both rural and urban settlements. From the plot, it is clear that the number of only child being a girl and not surviving is around 11,50,000. It decreases as the number of children increases.

Figure 2 shows trends seen in sex ratio in India as compared to the sex ratio of four different states. In 2011, the sex ratio of India was 940.

Madhya Pradesh, Bihar, Uttarakhand from 1901 to 2011.

Figure 3 and 4 show the sate-wise percentage of boys more than girls and number of girls born per 1000 boys respectively. This visualisation allows us to identify the areas that have low girl child birth rates and the areas that have higher percentage of boys than girl.

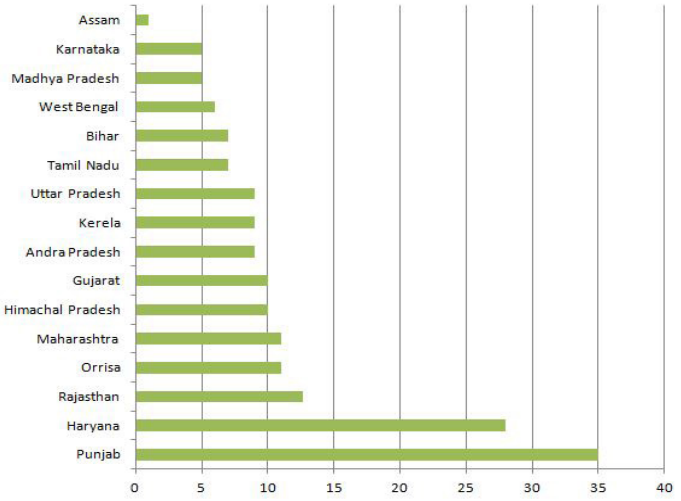

Figure 3. Percentage of boys more than girls for different states.

Figure 5 shows the Infant Mortality Rate in India. As the figure suggests, IMR in India has shown continual decrease in its value which is an evidence of the fact that the death rate of newborns has decreased over the years which in turn implies that the high (and in some cases increasing) death rate of girl child cannot be attributed solely to healthcare problems.

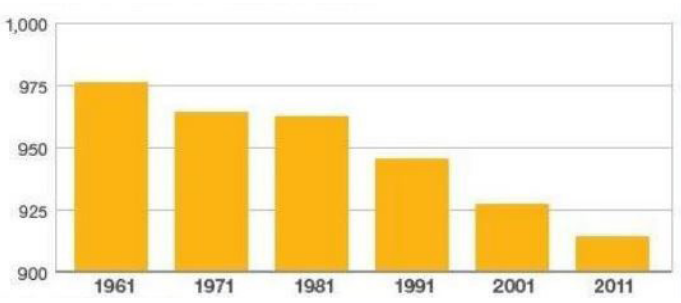

Figure 4. Number of girls born per 1000 boys.

We have also made predictions on IMR and sex ratio of India using Linear Regression. The results are summarised in Table 1 and Table 2 respectively. Although the data available to us was

highly limited, our system was able to make predictions with an accuracy of 93%.

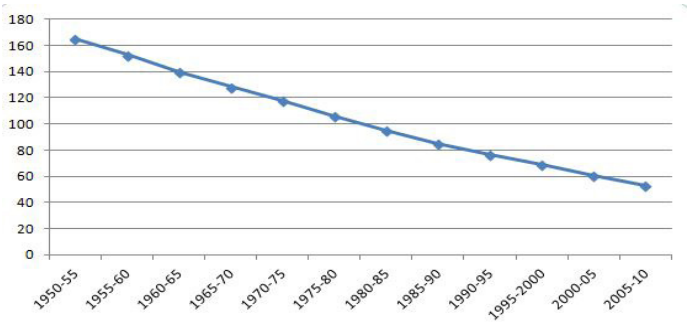

Figure 5. Infant Mortality Rate (IMR) for 1950 to 2010.

Table 1. IMR in India Prediction Data Comparison Chart

Year	Observed Value	Predicted Value
1997	77	77
2003	65	62
2007	52	49

Table 2. Sex Ratio in India Prediction Data Comparison Chart

Year	Observed Value	Predicted Value
1991	927	928
2001	933	930
2011	940	936

4. Discussions

Our implementation results show that the sex ratio of states like Madhya Pradesh and Uttarakhand is among the lowest in India. Moreover, the rate of increase in the ratio value is either too low or in some cases even negative. One direct implication of the result is identifying such regions and increasing vigilance in the local hospitals for illegal abortions, pre-natal sex determination of the child and female foeticide. There is a need for revision of healthcare policies and abortion laws in such states/regions.

Similarly, the analysis done on female children ever born and zero survivors shows an honest problem that exists in our country that the government is continually failing to tackle.

Our prediction results have shown that the IMR is decreasing continuous which is an evidence of the success of the governmental health policies in this area. However, the sex ratio in India shows unpredictable increases and decreases in its value. Although the value has never exceeded that in 1901, our prediction shows that if the current trends follow, the sex ratio in India will be greater than its 1901 value of 972 by 2031.

5. Conclusion

Data Analysis has a varied field of application. In this paper, we have applied it to indentify and tackle the gender composition problems existing in India. We have identified the informative patterns that exist in the data and made predictions based on those to provide a thorough understanding of future implications of these patterns. We have highlighted the areas that have witnessed improvements and those that are in dire need on it.

In the future, we aim to extend the implementation to include more varied data sets and also using semi-supervised and/or unsupervised regression for faster and more reliable predictions. Moreover, a more detailed data set can be employed for prediction and the accuracy of the prediction system can be future improved by using a higher degree polynomial for curve fitting etc.

6. References

1. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Services Research. 2007 Aug 1; 42(4):1758–72. https://doi.org/10.1111/j.1475-6773.2006.00684.x PMid:17286625 PMCid:PMC1955280
2. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. The qualitative researcher's companion. 2002 Mar 19; 573(2002):305–29.
3. Castro F, Vellido A, Nebot A, Mugica F. Applying data mining techniques to e-learning problems. In Evolution of teaching and learning paradigms in intelligent environment. Springer Berlin Heidelberg. 2007; 183–221. https://doi.org/10.1007/978-3-540-71974-8_8
4. Jongman RH, Ter Braak CJ, van Tongeren OF. Data analysis in community and landscape ecology. Cambridge University Press. 1995 Mar 2.
5. Hsieh WW, Tang B. Applying neural network models to prediction and data analysis in meteorology and oceanography. Bulletin of the American Meteorological Society. 1998 Sep; 79(9):1855–70. https://doi.org/10.1175/1520-0477(1998)079<1855:ANNMTP>2.0.CO;2
6. Brefeld U, Gartner T, Scheffer T, Wrobel S. Efficient co-regularised least squares regression. In Proceedings of the 23rd International Conference on Machine Learning, ACM. 2006 Jun 25. p. 137–44. https://doi.org/10.1145/1143844.1143862
7. Rahimi A, Recht B. Unsupervised regression with applications to nonlinear system identification. Advances in Neural Information Processing Systems. 2007; 19:1113.
8. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.
9. Berry MJ, Linoff G. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc. 1997 Jun 1.
10. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.

11. Piatetsky-Shapiro G. Advances in knowledge discovery and data mining. Fayyad UM, Smyth P, Uthurusamy R, editors. Menlo Park: AAAI press; 1996 Mar.
12. Mosteller F, Tukey JW. Data analysis and regression: a second course in statistics. Addison-Wesley Series in Behavioral Science: Quantitative Methods. 1977 Jan.
13. Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD. 1996 Aug 2; 96:82–8.
14. Seber GA, Lee AJ. Linear regression analysis. John Wiley & Sons. 2012 Jan 20.
15. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. Journal of the American Statistical Association. 2006 Mar 1; 101(473):119–37. https://doi.org/10.1198/016214505000000628

Citation:
Amita Jain, Amishapriya Singh, Devendra Kumar Tayal and Sonakshi Vij
"Amalgamating Data Analytics and Machine Learning for Predicting Sex Ratio and Infant Mortality Rate to Improve Gender Composition",
Global Journal of Enterprise Information System. Volume-8, Issue-3, July-September, 2016. (http://informaticsjournals.com/index.php/gjeis)
Conflict of Interest:
Author of a Paper had no conflict neither financially nor academically.