APPLICATION OF SCAN STATISTICS AS AN INFERENTIAL AND MULTIVARIATE TOOL TO ANALYSE CARDIOVASCULAR CASES IN INDIA: A DATA SCIENCE APPROACH

Saumya Sinha
Department of Mathematics
M.Sc. Data Science
Vellore Institute of Technology,
Vellore, India
saumya.sinha2023@vitsudent.ac.in

Sriya S Pillai
Department of Mathematics
M.Sc. Data Science
Vellore Institute of Technology,
Vellore, India
sriya.spillai2023@vitstudent.ac.in

Dr. Jitendra Kumar
Associate Professor
Department of Mathematics
Vellore Institute of Technology,
Vellore, India
jitendra.kumar@vit.ac.in

***Abstract*: Cardiovascular Diseases continue to be the leading causes of death across all regions of India, including less affluent states and rural areas. Medical professionals and experts note that heart-related issues among Indians have nearly doubled over the past decade, affecting even younger individuals. This increase is attributed to present lifestyles which contribute to conditions like high blood pressure, obesity, and diabetes, ultimately leading to heart problems. Early identification of signs of cardiovascular diseases and consistent medical treatment can reduce both mortality rates and the number of individuals affected. In the realm of spatial epidemiology, employing Scan Statistics as an inferential and multivariate analytical tool assists in determining which clusters of alarms warrant further investigation and which are likely random occurrences. Calculations have been performed with SaTScan Version 10.1.2. The research presented in this paper incorporates data from various states and union territories of India from the years 2017-2021.**

Keywords*: *Cardiovascular Diseases, Spatial epidemiology, Scan Statistics.

I. INTRODUCTION

In India, Cardiovascular Diseases are emerging as a major public health concern. The age-standardized death rate for CVD in India (282 deaths/100,000 (264–293)) is higher compared with global levels (233 deaths per 100,000 (229–236)). Urban areas have a higher prevalence of CVD compared to rural areas. The aim of this research paper is to detect cardiovascular disease hotspots across India. The problem of CVD hotspot detection is modelled using the Spatial Scan Statistics for finding every state/UT where there is higher density of cardiovascular cases than normal at a particular period of time. For calculations, SaTScan, a software that analyses spatial, temporal and space-time data and detect clusters, has been used.

Further, to analyse relation between different factors affecting CVD, the data has been modelled using linear regression. The project work is relevant to all stake holders ranging from common people to policy makers. In India, there is an approximate ratio of 17 nursing and midwifery professionals per 10,000 residents, while the doctor-to-patient ratio stands at 9 per 10,000 inhabitants. Hence, it is crucial to find out the feasible solution and exploring the hotspots. It enables in precise decision making and efficient resource allocation.

II. LITERATURE SURVEY

This section discusses some of the related works conducted in the field of cardiovascular disease analysis.

This research [1] works to find out the best prediction system for heart disease detection among the seven supervised machine learning techniques, namely KNN, SVM, Naïve Bayes, Random Forest, AdaBoost, ANN and Logistic Regression. The proposed work found LR to be the promising predictive model.

The research of authors Fiaidhi, J. and Mohammed S. [2] show that the strength of qualitative analytics lies in data thickness, involving patient characteristics such as socioeconomic status, family background, age group, gender & lifestyle risk factors, and their weights in a particular clinical practice. A Fuzzy C-Means algorithm is presented as technique to identify risk groups associated with CVD conditions, compared to ML Algorithms: Logistic Regression and Decision Tree.

In this research a stacking ensemble method is used [4] to produce an optimal predictive model, by combining several single classifiers, for the prediction of CVDs. EDA indicates that CVD was more common in males and diabetic individuals. Furthermore, people above the age of 65 were more susceptible to CVDs.

III. METHODOLOGY

DATASET: For scan statistics, the geographical data, the estimated population and the no. of CVD cases in each state and union territory of India has been collected to locate the hotspots. For modelling, data for the age group 45 and above has been collected from Longitudinal Ageing Study in India (LASI) Wave 1 for every State & UT. The dataset contains cases of different factors

relating to CVD such as, blood pressure, diabetes, alcohol and tobacco consumption, along with other necessary details. For comprehensive study, it is recommended to use the data from each district of the country.

S.No.	STATE/UT	Estimated Population	Estimated Adult Population (>45)	LATITUDE	LONGITUDE	Cardiovascular diseases (CVDs)	Hypertension or high blood pressure
1	Andaman & Nicobar	415547	97446	10.2188344	92.5771329	26993	25141
2	Andhra Pradesh	53139259	15171258	15.9240905	80.1863809	6022989	5840934
3	Arunachal Pradesh	1538603	340031	28.0937702	94.5921326	120031	112550
4	Assam	34837717	7594622	26.4073841	93.2551303	1564492	1496141
5	Bihar	120976512	22985537	25.6440845	85.906508	7079545	6757748
6	Chandigarh	1134647	266075	30.72811	76.77065	64656	58537

Table1: A glimpse of data (37 rows X 16 cols)

After collecting and processing the data, Scan Statistics, Descriptive analysis and Regression analysis is done, using SatScan and data analysis tool in excel.

MULTIPLE LINEAR REGRESSION MODEL:

For a given regressed variable, multiple linear regression model involves two or more regressor variables. It is created by calculating the regression coefficients, as indicated in eq1. The inference of the regressor variable on the regressed variable is suggested by these regression coefficients. Afterwards, prediction of regressed variable can be performed for a specific collection of regressor variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

where,

y = dependent/regressed variable (CVD cases)

β_0= y-intercept

$\beta_1, \beta_2, \ldots, \beta_k$ = regression coefficients

$x_1, x_2, \ldots, x_k$= independent/ regressor variables

Model Assumptions:

1. There is no Autocorrelation. Random error terms or disturbances are iid distributed.
2. $E[\varepsilon_i] = 0$
3. $V[\varepsilon_i] = \sigma^2$ There's assumption of Homoskedasticity. Variance of each disturbance term conditional on the chosen value of independent variable is equal i.e. σ^2

Descriptive analysis of various factors is performed. The mean and standard deviation is utilised to calculate Coefficient of Variance of each independent variable. Ranking is done on the basis of CV.

	Mean	Standard Deviation	Coefficient of Variance	Ranking
Hypertension or high blood pressure	2390211.5	2775212.926	116.1074202	1
Obesity by Anthropometric Indicators	656620.62	784902.9944	119.5367566	2
Diabetes or high blood sugar	1020418.2	1287862.081	126.2092439	3
Prevalence of heavy episodic drinking	327204.27	413584.8982	126.3996029	4
Anaemia	345212.86	484537.42	140.3590275	5
High Cholesterol	206192.73	294001.0353	142.5855488	6
Currently smoking	986662.86	1529223.861	154.9895021	7
Currently consuming tobacco	1779047.9	2959404.932	166.3476847	8

Table2: Ranking variables on the basis of CV

The top 5 variables, namely, High BP, Obesity, Diabetes, Alcohol Consumption and Anaemia. are considered for regression analysis. The summary output of regression analysis includes regression statistics, ANOVA, residual output and probability output.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	3.21E+14	6.41E+13	6077.573	2.18289E-45
Residual	31	3.27E+11	1.06E+10		
Total	36	3.21E+14			

Table3: ANOVA Table-1

Similarly, Proportional Ranking has been calculated for each independent variable.

			Ranking
High Cholesterol	7629131	0.026738	1
Prevalence of heavy episodic drinking	12106558	0.04243	2
Anaemia	12772876	0.044766	3
Obesity by Anthropometric Indicators	24294963	0.085147	4
Currently smoking	36506526	0.127946	5
Diabetes or high blood sugar	37755473	0.132323	6
Currently consuming tobacco	65824771	0.230699	7
Hypertension or high blood pressure	88437826	0.309951	8
	285328124		

Table4: Proportional Ranking

The top 5 variables, namely, High Cholesterol, Alcohol Consumption, Anaemia, Obesity and Smoking, are considered for regression analysis.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	5	2.99E+14	5.97E+13	82.35583	5.86347E-17
Residual	31	2.25E+13	7.25E+11		
Total	36	3.21E+14			

Table5: ANOVA Table-2

Regression Plots of important variables are as follows:

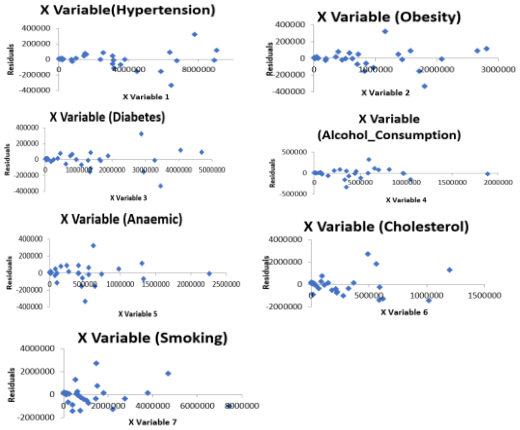

Fig1: Residual Plots

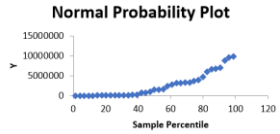

Fig2: Normal Probability Plot

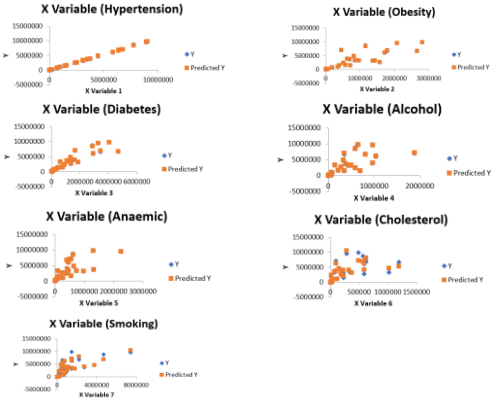

Fig3: Line Fit Plots

Statistical Criteria	Model based on CV	Model based on PR
R Square	0.998980896	0.929987668
Adjusted R Square	0.998816524	0.918695356
Standard Error	102722.416	851418.6308
p-value	2.18289E-45	5.86347E-17

Table6: Comparative Study of models based on Consistency ranking and that based on Proportional ranking.

SCAN STATISTICS: The excel data has been used as input into SatScan application. Case file includes the number of cases of independent variables, taken separately, coordinate file contains the latitude and longitude, and lastly, population file contains the number of CVD cases of every State & UT. Discrete Poisson based model is used to identify the hotspots.
Clusters on the basis of:

Fig4: CVD

Fig5: Hypertension/BP

Fig6: Alcohol Consumption

Fig7: Obesity

Fig8: Diabetes

Fig9: Anaemia

IV. RESULT AND DISCUSSION

The main objective of this research is to evaluate the statistical significance of disease cluster alarms and find the hotspots. Here, Discrete Poisson based model is used to identify the hotspots. After the completion of the calculations, a standard text-based results file is automatically generated, containing information about the clusters detected, computing time and the analysis parameters chosen. Prominent clusters on the basis of CVD cases are as follows: Primary-Gujarat, Odisha, Chhattisgarh, Jharkhand, West Bengal, Bihar, Telangana, AP, Tripura. Secondary- J&K, Ladakh, HP, Punjab. On the basis of: Hypertension- AP, Telangana, Puducherry, Karnataka. Alcohol Consumption- Sikkim, Bihar, Meghalaya, Assam. Obesity- Karnataka, Goa, Kerala, Maharashtra, TN, AP, Lakshadweep, Telangana, Puducherry, Daman & Diu, Dadra Nagar Haveli, Gujarat, Chhattisgarh. Diabetes- Karnataka, Goa, Kerala, Maharashtra, TN, AP, Lakshadweep, Telangana, Puducherry, Daman & Diu, Dadra & Nagar Haveli, Gujarat, Chhattisgarh. Anaemia- Delhi, Haryana, Chandigarh, Uttarakhand, Punjab, HP, Rajasthan, UP, Madhya Pradesh, J&K.

The multiple linear regression equation considering top 5 variables based on CV:

$$y = -6990.78 + 1.130775x_1 + (-0.02882)x_2 + (-0.037)x_3 + (-0.23315)x_4 + (-0.0423)x_5 \quad (2)$$

The multiple linear regression equation considering top 5 variables based on PR:

$$y = -155527 + 1.438701x_1 + 2.478916x_2 + (-0.05643)x_3 + 1.444067x_4 + 0.676285x_5 \quad (3)$$

There is no pattern followed in the error plots(Fig1), which implies that 3rd assumption is valid.
MODEL VALIDATION:
Hypothesis testing
1. H_0: $\beta_0 = \beta_1 = \cdots = \beta_5$
2. H_1: Atleast one variable is significant
Let $\alpha = 0.05$
Test statistic:
Suppose we apply p-value criteria. In both the models p-value is less than α (Table5,6). Hence, we reject H_0
Similarly, suppose we apply F-statistic criteria.
$F_{table(5,31,0.05)} = 2.5336$, $F_{stat} = 6077.57$ & 82.35 for both the models respectively (Table5,6). Since $F_{stat} > F_{table}$, we reject our null hypothesis H_0.
This implies that both our models are valid.

Comparing the 2 models (Table5), model based on coefficient of variance gives better results as R square value i.e. 0.998 is higher than that of the counterpart. Even the standard error is lesser in comparison.

V. CONCLUSION

This project involves examining cardiovascular data from 37 states and union territories. The dataset's descriptive analysis reveals a predominantly positive skew, which is also highlighted from the NPP(Fig3). Regression analysis indicates a strong fit with 99% accuracy when considering major factors such as Hypertension, Diabetes, Alcohol Consumption, Obesity and Anaemia. It has yielded a multiple linear regression equation (eq2). Hotspots, identified through SatScan, majorly include regions like Gujarat, Andhra Pradesh, Telangana, Bihar, Karnataka, Maharashtra, Kerala, Chhattisgarh, Punjab and Tamil Nadu. These analyses can be used to design a model for proper resource allocation in these critical regions and prevent increase in cardiovascular diseases.

VI. REFERENCES

[1] Kumar, Vinod, Lalotra, Gotam Singh. "*Predictive Model Based on Supervised Machine Learning for Heart Disease Diagnosis".* 2021 IEEE International Conference on. :1-6 Dec, 2021

[2] Fiaidhi, J., Mohammed S. *"Prognosis analysis of thick data: Clustering heart diseases risk groups case study".* Department of Computer Science, Lakehead University, Ontario, Canada: June 2021

[3] Lohaj, Oliver, Pella, Zuzana, Paralic, Jan. *"Data analytics methods for analyzing the impact of factors on early detection of cardiovascular risk"*. 2022 IEEE 20th Jubilee World Symposium (SAMI): Mar, 2022

[4] Rustamov, Zahiriddin, Rustamov, Jaloliddin, Sultana, Most Sarmin, Ywei, Jeanne, Balakrishnan, Vimala, Zaki, Nazar. *"Cardiovascular Disease Prediction using Ensemble Learning Techniques: A Stacking Approach"*. 2023 19th IEEE International Colloquium (CSPA): Mar, 2023