Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER
PUBLISHER

Research Article

Spatial Aspects of Acute Respiratory Disease Syndrome: An Application of Scan Statistics Using SaTScan in Identification and Analysis of Hotspot in India

Priyanka Subramani, Kalpanapriya Dhakshnamoorthy*

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore-632 014, Tamil Nadu, India
E-mail: dkalpanapriya@vit.ac.in

Received: 6 May 2024; **Revised:** 12 August 2024; **Accepted:** 28 August 2024

Abstract: Respiratory illnesses like acute respiratory disease syndrome (ARDS) have been a persistent issue throughout history, rather than being a contemporary concern and have likely afflicted humans since ancient times, these have become increasingly more prevalent during recent decades and these diseases are among the leading causes of global deaths. This study investigates the spatial distribution of ARDS in India. In the present study, we acquired ARDS disease data from the year 2006 to 2021 and environmental, demographic, and meteorological data from various government sources. Using spatial scan statistics, we detected significant clusters of ARDS cases, highlighting areas with unusually high incidences. To further comprehend the factors contributing to these clusters, we employed regression modeling incorporating a Box-Cox transformation to identify and analyze potential explanatory variables. This transformation was crucial for stabilizing variance and making the data more suitable for regression analysis. Comprehensive and integrated diagnostics are being implemented to validate the model and derive judicious implications. For robust computing, we have used the software MS Excel, MS Solver, R, and Python. Our findings reveal critical insights into the spatial dynamics of ARDS and underline the importance of spatial analysis and appropriate statistical transformations in public health research. These results can guide resource allocation and policy-making aimed at mitigating ARDS in India, ultimately contributing to better health outcomes.

***Keywords*:** acute respiratory disease syndrome, scan statistics, hotspot, most likely clusters, multiple regression

MSC: 65L05, 34K06, 34K28

Abbreviation

ARDS	Acute Respiratory Disease Syndrome
MERS	Middle East Respiratory Syndrome
SARS	Severe Acute Respiratory Syndrome
TB	Tuberculosis
LRI	Lower Respiratory Infection
COPD	Chronic Obstructive Pulmonary Disease

DOI: https://doi.org/10.37256/cm.5320244880

NFHS	National Family Health Survey
GIS	Geographic Information System
PM	Particulate Matter
SD	Standard Deviation
CV	Coefficient of Variation
BLUE	Best Linear Unbiased Estimator
LLR	Loglikelihood Ratio
RR	Relative Risk
MLC	Most Likely Clusters
AIC	Akaike Information Criterion
SBIC	Schwarz Bayesian Information Criterion
MSE	Mean Square Error
MAD	Mean Absolute Deviation
MAPE	Mean Absolute Percentage Error

1. Introduction

Acute respiratory disease syndrome (ARDS) is a leading cause of morbidity and mortality globally, the World Health Organization (WHO) [1] reports that respiratory diseases were responsible for more than 8 million deaths annually. Improvements in the knowledge of respiratory disorders occurred during the 19th century, especially with the advent of microbiology. The discovery of bacteria and viruses, made possible by researchers like Louis Pasteur and Robert Koch, opened the door to our understanding of infectious respiratory diseases. Infectious respiratory diseases have surfaced in the 20th and 21st centuries, leading to epidemics and outbreaks. Examples include the Middle East respiratory syndrome (MERS) in 2012, the severe acute respiratory syndrome (SARS) in 2003, and the COVID-19 pandemic that started in late 2019 and was brought on by the new coronavirus SARS-CoV-2. In India, respiratory infections are the combined death rate for pneumonia and influenza positions these diseases as the top leading cause of deaths. The identification of risk factors for ARD in the community is important for developing effective policies and strategies to interrupt transmission and improve health outcomes. Studies in developing countries have reported that growth, the poorest regions of the world have the greatest disease burden from chronic respiratory diseases. Risk factors such as smoking, environmental pollution, and body weight also play a key role, nutritional factors, and parental smoking were associated with ARDS [2–4]. However, due to the differences in both living condition and environmental circumstances, these results cannot be directly applied to industrialized countries like India. According to research by Xie and colleagues [5, 6], between 1990 and 2017, the overall number of patients with chronic respiratory conditions grew by 39.5%. The spatial and temporal distribution of ARDS and the environmental elements influencing that distribution have also been successfully mapped using geographic information systems (GIS) [7]. Kim et al. [8] confirmed that air pollution increases the risk of the most common respiratory diseases (RDs) such as lung cancer, asthma, tuberculosis (TB), lower respiratory infections (LRIs), and chronic obstructive pulmonary disease (COPD). Research had examined that spatial cluster detection is an important problem in spatial epidemiology among the various statistical methods, spatial scan statistics is highly recommended to find out the geographical disease surveillance over the collection of scanning windows [9]. Table 1 provides a summary of relevant research and literature. After an extensive literature review, we found that the spatial analysis using scan statistics is often used for finding the hotspots. However, there's a limited research on Acute respiratory disease cases in India with respect to specific reasons. To fill this gap, this study is directed towards achieving the subsequent goals of this study:

- Investigating the incidence of ARDS in India through exploratory and descriptive statistics.
- Finding out the probabilistic hotspots and coldspots of ARDS.
- Develop the suitable statistical model to fit the ARDS cases for future forecasting.
- Using inferential measures diagnostic and hypothesis testing to validate the model.

Through the use of a spatial analysis, this study offers insightful information for better hotspot prediction with reference to certain causes. The primary causes of the illness are its causes, which improve the precision and effectiveness of ARDS prevention strategies by allowing for the customization of interventions, the allocation of resources, the development of policies, and the implementation of public awareness campaigns that target the particular cause of this annoyance. This study makes it possible to eliminate the illness with concentration and initiative.

Table 1. Literature review of different approaches and time frame

Author	Context	Domain	Time frame	Method
Anjali et al. [10]	Spatio-temporal	Crime, four major cities of Tamil Nadu	2011-2019	Hotspot analysis and modelling on suicidal cases
Balasubramani et al. [11]	Spatial analysis	Health, India	2015-2016	Hotspot and statistical analysis
Fatima et al. [12]	Spatial distribution	Health, Southern Punjab, Pakistan	2010-2012	Inverse distance weighted (IDW) by spatial interpolation, spatial autocorrelation, cluster outlier analysis
Wang et al. [13]	Spatio-temporal varying coefficient model	Health, Taiwan	2019-2020	Bayesian mapping model, model fitting
Kaindal et al. [14]	Spatio-temporal	Health, India	2010-2020	Hotspot analysis
Katale and Gemechu [15]	Spatio-temporla	Health, North Namibia	2018-2020	Spatial autocorrelation, Bayesian CAR model
Kumar and Anjali [16]	Spatial analysis of multivariate factors	Crime, India	2012-2021	Welch t-test, exponential smoothing, MANOVA
Manish et al. [17]	Spatial analysis	Crime, Delhi	2005-2015	Hotspot, ANOVA, model fitting
Saravag [18]	Spatio-temporal	Crime, Rajasthan	2014-2020	Hotspot and time series forecasting
Tesfaye et al. [19]	Spatial patterns	Health, Ethiopia	2005, 2011, 2016	Statistical and spatial analysis, spatial regression analysis
Abolhassani et al. [20]	Spatial, spatio-temporal	Health	1936-2021	Scan statistics
Mondal et al. [21]	Geo-spatial	Different crimes	2012-2015	Kernel density estimation (KDE), and Getis-Ord Gi*
Montnemery et al. [22]	-	Health	-	Multiple logistic regression
Saran et al. [23]	Geo-spatial	Health	2019-2020	Infectious disease modelling
Sukhija et al. [24]	Spatial visualization	Crime	2010-2014	K-means clustering, spatio-temporal generalized additive local spatio-temporal generalized additive

2. Data & research design

We collected data for our research from reliable secondary sources, including government agencies and various public and private websites. Instead of conducting primary surveys, we made this choice to ensure the accuracy and validity of our dataset. Conducting surveys would have been challenging due to limited control over respondents and potential issues with scientific validity. Our aim was to optimize both time and cost while leveraging the wealth of information available digitally. We tailored our data structure, processing, and design to suit our computational algorithm, employing techniques like normalization through z-scores and box plots. Our research design combines exploratory, descriptive, and causal elements, incorporating spatial and temporal data analysis to identify hotspots and cold spots. Ultimately, our design allows us to draw inferential conclusions and develop effective policy frameworks for tackling the eradication of ARDS cases. We focused on analyzing data at the state level in India.

3. Methodology

3.1 *Exploratory and descriptive analysis*

The initial stage of analysis is exploratory, where our focus is on uncovering surface-level attributes of the dataset. Here, we conduct basic statistical operations such as data processing, identifying missing values, and organizing the data into a customized format. We also look for outliers and transform raw data into normalized data using z-scores. Visualizing general trends helps us understand the monotonicity of attributes, whether they are increasing or decreasing. Additionally, we categorize variables and parameters into nominal, ordinal, interval, and ratio scales for subsequent computations. The logic established in this stage forms the foundation for descriptive and inferential analyses. Our goal is to understand and explore surface-level visualizations or general patterns, often utilizing various plots and charts. Trend fitting allows us to track the progression of ARDS cases over the years, while identifying outlier districts provides specific insights. Descriptive analysis involves characterizing data using absolute and relative measures. For instance, the coefficient of variation (CV) is a relative tool, whereas measures like sample mean and variance are absolute. When comparing basic characteristics, we use absolute measures, relational measures for comparison, and relative tools for model building. Descriptive statistics play a key role in statistical modeling, expressing our prior understanding of the probability experiment that generated the observed data.

3.2 *Hotspot analysis*

The SaTScan tool, developed by Kulldorff [25, 26], was employed to implement and estimate the parameters of scan statistics. A Monte Carlo simulation procedure was used to compute the log-likelihood ratio value for each window or circle and determine the corresponding p-values. In this study, we examined and identified hotspot regions of ARDS cases scattered across the country, aiming to highlight the most probable clusters using the discrete Poisson model. The method involves a scanning window with a moving column, where the base represents a geographic area and the height indicates time. By comparing actual and theoretical incidences inside and outside the scanning window, we calculate the log likelihood ratio (LLR). The theoretical incidence for each window is determined based on the overall incidence and the number of individuals within it. The window with the highest LLR is selected as it indicates the strongest clustering, while other windows with statistically significant LLR values contribute to defining secondary clusters, refining the regional distribution of ARDS notification rates. The likelihood function for a particular window under the Poisson model is proportional to the equation (1).

$$L(\upsilon)=\left(\frac{\upsilon}{E[\upsilon]}\right)^{\upsilon}\left(\frac{\Upsilon-\upsilon}{\Upsilon-E[\upsilon]}\right)^{\Upsilon-\upsilon}K(\upsilon) \tag{1}$$

In this context, Υ represents the total number of ARDS cases, υ stands for the observed number of ARDS cases within the window, and $E[\upsilon]$ denotes the covariate-adjusted expected number of cases within the window under the null hypothesis. The term Υ-$E[\upsilon]$ signifies the expected number of cases outside the window, considering the constraint that the total observed cases are accounted for. In equation (1), the function $K(\upsilon)$ serves as an indicator, taking the value 1 if the recorded cases within the window exceed what is expected under the null hypothesis, and 0 otherwise. Equation (2) outlines the Relative Risk, which is the ratio of the estimated risk within the cluster to the estimated risk outside the cluster.

$$RR = \frac{\upsilon/E[\upsilon]}{\Upsilon - \upsilon/E[\Upsilon] - E[\upsilon]} = \frac{\upsilon/E[\upsilon]}{\Upsilon - \upsilon/\Upsilon - E[\upsilon]} \tag{2}$$

The count of cases detected within the cluster is denoted as υ, while the total number of cases is represented by Υ. As the analysis is limited to the total observed cases, which remains constant at Υ, it follows that the expected value of Υ, denoted as $E[\Upsilon]$ equals Υ.

3.3 *Inferential analysis*

Inferential analysis is the process of drawing inferences from data using a variety of statistical techniques. Estimating and evaluating the model's parameters was our primary concern. After estimating these parameters, we verified the estimated values by statistical testing. The estimator converges to the true parameter if it meets the four requirements of unbiasedness, consistency, efficiency, and sufficiency. We used both point and interval estimates to help us make decisions.

3.4 *Regression modelling*

Regression modelling for ARDS cases at state levels, we have attempted to implement multiple linear regression model consists of average ARDS cases during the period 2006-2021. The model is being represented by equation (3),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_i X_i + \varepsilon_i; \quad i = 1,\ 2,\ 3,\ \cdots \tag{3}$$

$k =$ number of independent regressors in the model, such that $p = k+1$; where p denotes the number of parameters in the model, ε_i is the error term. For estimating the parameter of the model illustrated above we have implemented the least square method (LSM). As per Gauss Markov's theorem, we will able to estimate these β's known as best linear unbiased estimator (BLUE) that is likely to satisfy the criteria of unbiasedness, consistency, efficiency, and sufficiency as defined above. For purpose of proper statistical analysis and drawing the inferentials were need to compute the R^2, $R^2_{adjusted}$ measures. The role of error analysis play vital role in regression modelling especially to diagnose the problem of multicollinearity, autocorrelation and homoscedasticity of the model using different approach. Further more, error analysis enable us to find out the suitable solution of these problems.

The akaike information criterion (AIC) is given by the equation (4) used to assess each model's relative superiority for a particular data set when comparing statistical models fitted by maximum likelihood (ML) to the same data, often for non-nested models, to statistical models fitted by other methods. As a result, the statistic penalizes for the number of predictors included in the model and considers model parsimony: It is obvious that the second term is a deviation and the first is a penalty for the number of parameters. The statistical model with the lowest AIC value is the most successful.

$$AIC = n\ln(SSE) - n\ln(n) + 2p. \tag{4}$$

Using the AIC, a statistical model's comparative excellence for a particular set of data is evaluated. The AIC provides an unbiased method for identifying the most economical model among several rival models. When compared to the AIC values of rival models, the meaningless AIC values for a specific set of data become meaningful. The optimal (ideal) model for the given data set is the one with the lowest AIC value among rival models. The AIC penalizes the addition of parameters since it is used to choose a model with few parameters that fits the data well.

The Schwarz Bayesian information criterion (SBIC), shown in the equation (5) a distinct estimate of model fit for a given set of data across several non-nested model types, has the following formula:

$$SBIC = n\ln(SSE) - n\ln(n) + p\ln(n). \tag{5}$$

When additional variables are added to a model, model selection statistics like the residual sum of squares always go lower. This is known as overfitting, and it is solved by Mallows's C_p (6).

$$C_P = \frac{SSE_P}{\hat{\sigma}^2} + 2(p+1) - n. \tag{6}$$

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}. \tag{7}$$

$$MAD = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}. \tag{8}$$

$$MAPE = \left(\frac{\sum_{i=1}^{n} \frac{(|y_i - \hat{y}_i|)}{|y_i|}}{n} \right) \times 100. \tag{9}$$

The above equations (7)-(9) are commonly computed measures. The decision rule (guiding principle) can be presented as models with smaller mean square error (MSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE) are better for prediction purpose. MAPE is a relatively measure and shows percentage deviations of forecast from the original datasets. Though, average relative measure prefers under estimated values and as such is somewhat biased, still considered relevant for an effective analysis. The advantage of using such a approach is that the models are tested on the data that are not used in the fitting (model estimation process). This approach provides an independent assessment of models predictive ability. We compute PRESS statistic and the PRESS residuals by the equation (10) & (11) for the prediction model. R^2_{adj} maintains its value at the target R^2 of the model, despite increasing the variable, while R^2_{PRESS} decreases as variable increases. A method of assessing prediction quality is to use PRESS statistic. PRESS stands for Prediction Sum of Squares and is defined as

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{i,\ -1})^2, \tag{10}$$

where $\hat{y}_{i,\ -1}$ represent the prediction obtained from a model estimated with one of the sample observations deleted. If there are n-observations in the sample, then there are n-different prediction $\hat{y}_{i,\ -1}$. The quantities $(y_i - \hat{y}_{i,\ -1})$ often are called PRESS residuals because they are similar to the actual residuals $(y_i - \hat{y}_i)$. The prediction of sum of squares is also equal to the error sum of squares (SSE). We use the statistic,

$$R^2_{PRESS} = 1 - \frac{PRESS}{SST}. \tag{11}$$

Here, the guiding decision rule is larger values of R^2_{PRESS} (or smaller values of PRESS) suggest the models of greater predictive ability.

4. Results and discussions

4.1 *Exploratary and descriptive results*

The distribution of ARD cases across different states in India is not uniform, with some areas experiencing higher concentrations of cases compared to others. To focus our analysis, we collected data from all states of India, forming our sample size for hotspot detection using saTScan. Our preliminary descriptive analysis, detailed in Table 2, provides essential statistics for further examination. Notably, we found a high kurtosis (> 3) indicating a leptokurtic curve, suggesting a pronounced clustering of ARDS cases throughout the study period. This highlights the need for thorough investigation into the underlying reasons for such heterogeneity. Additionally, a skewness value exceeding 2 indicates a positively skewed distribution of ARDS cases nationwide, prompting consideration for fitting a negative binomial distribution model. This raises the possibility of identifying both hotspots and coldspots of ARDS cases. Our approach for hotspot detection leans towards utilizing scan statistics, particularly with a Poisson model deemed suitable for our spatio-temporal dataset. Figure 1 presents a trend line visualization plot during the 2006-2021, enable us to have rough idea of growth pattern of ARDS cases and its deaths in India. From 2006 to 2009, there was an upward trend in the number of cases, followed by a decline from 2010 to 2012. However, from 2012 to 2019, there was a sharp increase in cases, which then saw a significant drop in 2020 and 2021. The number of deaths varied slightly between 2006 and 2014 and then remained steady for the next four years. Since 2019, there has been a noticeable increase in the number of deaths. From the graphs depicting incidences and deaths, we can observe a noticeable shift in the pattern during the COVID-19 period.

Table 2. Exploratory and descriptive statistics for ARDS cases of the year 2006-2021

Summary statistics	2006	2011	2016	2019	2020	2021
Mean	729,615.5	730,561.3	1,089,274	1,100,537.27	631,110	468,448.3
Standard deviatio	1,407,559	1,053,292	1,318,310	1,505,516.857	866,612.6	766,384.1
Sample variance	1.98 E + 12	1.11 E + 12	1.74 E + 12	2.26658 E + 12	7.51 E + 11	5.87 E + 11
Kurtosis	19.76398	7.44579	3.650089	3.393249914	3.527156	7.941491
Skewness	4.087382	2.49463	2.605902	2.742220422	2.903537	2.649839
Minimum	12,602	25,441	19,537	21,633	15,228	5,494
Count	37	37	37	37	37	37

Descriptive analysis is crucial for informing inferential methods. We've calculated coefficient of variations and rankings at the state level to understand the significance of each data element. These findings serve as the foundation for building a model and enable comparison with hotspot states identified through saTScan statistics. To detect outliers, we've computed z-scores for each state and identified outlier states on a yearly basis. The boxplot generated from the dataset covering the years 2006-2021 provides a visual summary of ARDS case data for each state given in Figure 2. Years are defined by the x-axis, while incident cases are defined by the y-axis in lakhs. The majority of states exhibit typical data distribution, but outliers are noticeable in states such as Kerala, Ladakh, and Rajasthan, which fall outside the expected range. Furthermore, the data consistently shows skewness to the right, indicating a consistent trend across all years.

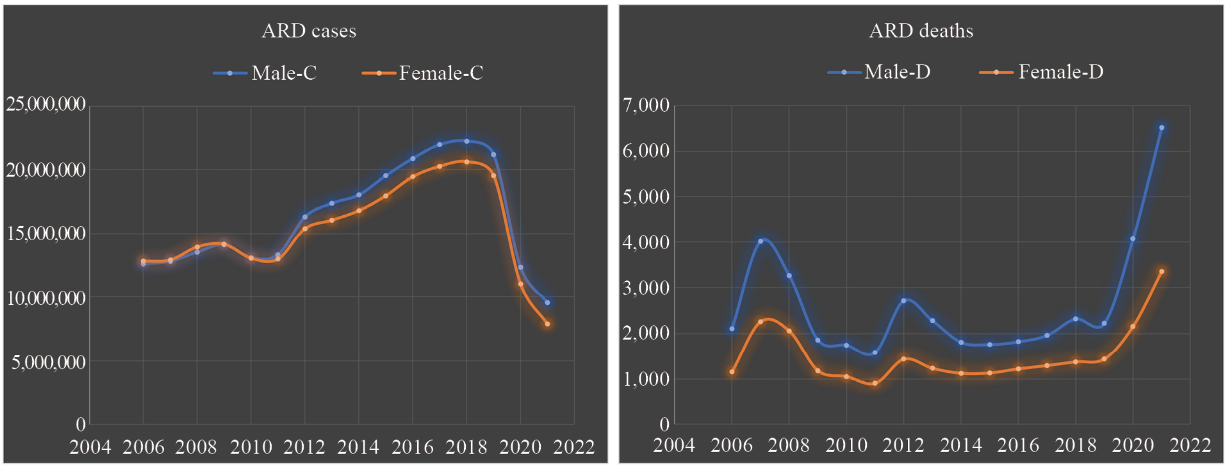


Figure 1. Trendline of ARDS cases and deaths in India

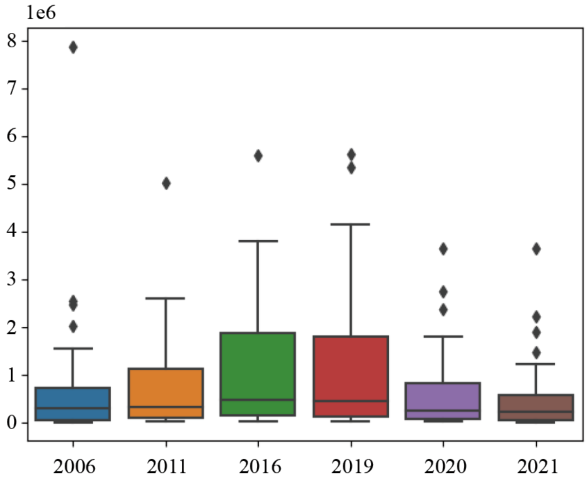


Figure 2. Boxplot of ARDS cases in India

4.2 *Hotspot analysis*

This study aimed to identify hotspots and coldspots using scan statistics methodology. The computational equation used for hotspot calculation via the saTScan software is represented. Figure 3 illustrates the primary hotspot of ARDS cases in 2021, although due to space constraints, visual figures for other years are not presented but were computed similarly. Table 3 provides inferential statistics for hotspot characterization of TB cases, including relative risk (RR), log-likelihood ratio (LLR), and *p*-values for primary clusters. Statistical values for additional clusters such as secondary, tertiary, quaternary, quinary, etc were intentionally excluded as they were not deemed statistically significant. Table 4 and 5 shows the top 5 and least 5 in consistent ranking of ARDS cases, where Table 6 gives the primary and secondary hotspots of ARDS cases in each year under study.

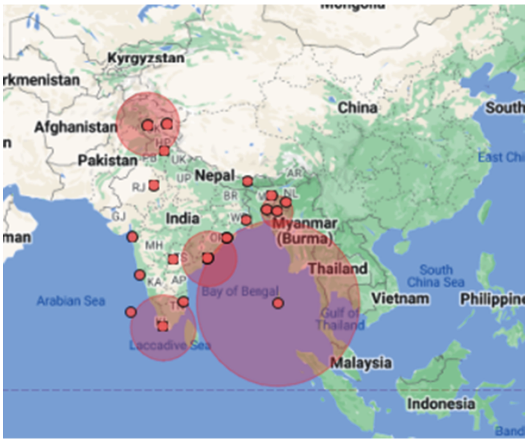


Figure 3. Hotspot of ARDS cases in India

Table 3. Hotspot analysis of ARDS

Parameters	Primary hotspot
Gini cluster	Yes
Number of cases	859,661
Expected cases	1,464.19
Annual cases	77,866.2
Observed/expected	587.12
Relative risk	604.74
Log likelihood ratio	4,634,968.80
P-value	< 0.00000000000000001

Table 4. Top 5 most consistent ranking of ARDS

S.No	States	Mean	SD	CV	Rank
1	Andhra Pradesh	2,546,594.5	589,991.92	23.16	1
2	Punjab	499,670.17	159,995.67	32.02	2
3	West Bengal	2,775,941.5	890,882.7	32.09	3
4	Himachal Pradesh	1,291,527	445,221.43	34.47	4
5	Nagaland	21,213	7,610.12	35.87	5

Table 5. Bottom 5 most consistent ranking of ARDS

S.No	States	Mean	SD	CV	Rank
1	Lakshadweep	859,661.33	2,045,282.7	237.91	1
2	Manipur	119,303.5	222,497.74	186.49	2
3	Sikkim	238,674.67	418,093.43	175.17	3
4	Assam	72,281.33	119,335.53	165.09	4
5	Mizoram	72,594.5	109,737.65	151.16	5

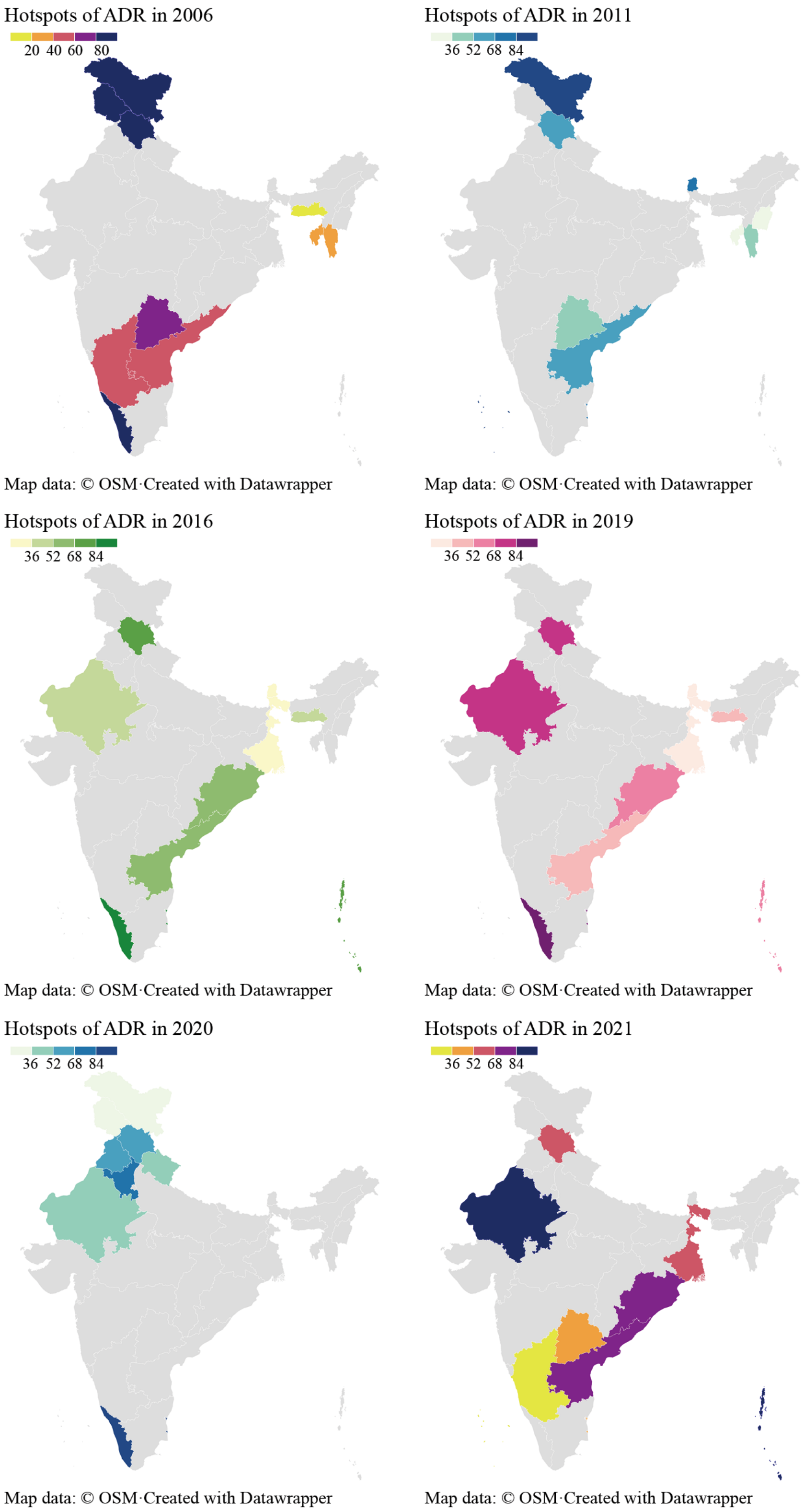
Table 6. Primary and secondary hotspots of ARDS

Year	Primary hotspots	Secondary hotspots
2006	Kerala, Puducherry	Jammu Kashmir, Ladakh, Himachal Pradesh
2011	Lakshadweep	Ladakh
2016	Kerala, Puducherry	Himachal Pradesh
2019	Kerala, Puducherry	Himachal Pradesh
2020	Kerala, Puducherry	Chandigarh, Haryana, Punjab, Himachal Pradesh
2021	Rajasthan	Andaman Nicobar, Odisha, Andhra Pradesh, West Bengal

After assessing whether hotspots remain consistent or change annually, we proceeded with hypothesis testing to determine whether distinct strategies, policies, and planning are needed for ARDS cases, or if an integrated approach suffices. Employing the z-test, we obtained a computed z-value of 4.8129, with a critical z-value of 1.6448 at $\alpha = 0.05$. Since the computed z-value exceeds the critical value, we reject the null hypothesis, indicating a significant difference between hotspot types. Consequently, we recommend implementing separate policy frameworks for each type of ARDS hotspot. This finding is crucial as it provides scientific evidence to guide policymakers in effectively addressing both types of ARDS cases. Figure 4 gives the visualization idea of the hotspots of ARDS cases in India throughout the period of 2006-2021.

4.3 *Statistical modelling and inferential analysis*

We aim to construct and train a statistical model to aid policymakers and program implementers in predicting and managing ARDS cases in India at the state level. To meet our statistical requirements, we utilized multiple linear regression, carefully selecting variables. We conducted thorough data processing to address outliers, missing values, and any extraneous information. The fitted regression model is expressed by the equation below. To enhance the model's performance and control variability, we explored and applied two sets of transformations: $\log y \sim \log x$ and $\sqrt{y} \sim \sqrt{x}$, respectively, on two sets of variables. One set comprises twelve factors, while the other has six factors which was shown in the Table 7. Our intent is to evaluate the model's performance with twelve regressors against the same model with six regressors.


Hotspots of ADR in 2006

Map data: © OSM·Created with Datawrapper

Hotspots of ADR in 2011

Map data: © OSM·Created with Datawrapper

Hotspots of ADR in 2016

Map data: © OSM·Created with Datawrapper

Hotspots of ADR in 2019

Map data: © OSM·Created with Datawrapper

Hotspots of ADR in 2020

Map data: © OSM·Created with Datawrapper

Hotspots of ADR in 2021

Map data: © OSM·Created with Datawrapper

Figure 4. Hotspots based on ARDS cases during 2006-2021

Table 7. List of variables for provided trasformations

S.No	Factors	Variables
1	12 regressors	Per capita income, population density, malnutrition, temperature, urbanisation, humidity, air pollutants (PM (2.5), PM (10), SO_2, CO, O_3, NO_2).
2	6 regressors	O_3, NO_2, population density, malnutrition, temperature, humidity.

The fitted regression for $\sqrt{y} \sim \sqrt{x}$ transformation of 12 variables is

$$\begin{aligned} y = 5.9028 + (0.4244)x_1 + (-0.4442)x_2 + (-0.0735)x_3 + (0.0281)x_4 + (0.0425)x_5 + (0.2602)x_6 \\ + (-0.0012)x_7 + (-0.3346)x_8 + (0.3874)x_9 + (0.8389)x_{10} + (-0.5486)x_{11} + (0.5348)x_{12} \end{aligned} \tag{12}$$

The fitted regression for $\log(y) \sim \log(x)$ transformation of 12 variables is

$$\begin{aligned} y = 6.5418 + (1.1685)x_1 + (-2.5479)x_2 + (-0.2984)x_3 + (0.9946)x_4 + (0.0013)x_5 + (0.4212)x_6 \\ + (-0.0668)x_7 + (-0.9008)x_8 + (0.0795)x_9 + (0.4239)x_{10} + (-2.8722)x_{11} + (0.0247)x_{12}, \end{aligned} \tag{13}$$

The fitted regression for $\sqrt{y} \sim \sqrt{x}$ transformation of 6 variables is

$$y = 1.8632 + (0.0489)x_1 + (0.1594)x_2 + (-0.0141)x_3 + (-0.0780)x_4 + (0.6518)x_5 + (-0.4206)x_6, \tag{14}$$

The fitted regression for $\log(y) \sim \log(x)$ transformation of 6 variables is

$$y = 3.0674 + (0.2011)x_1 + (0.2741)x_2 + (-0.1823)x_3 + (-0.0113)x_4 + (1.5048)x_5 + (-2.6851)x_6. \tag{15}$$

The above equations are suitable for short-term and medium-term forecasts, but they are not advisable for long-term forecasts. Long-term forecasting may be impacted by various time series components and uncontrolled dynamics across the country, and potentially even beyond in global regions. We observe from above fitted models that five out of the twelve regressors namely x_2, x_3, x_7, x_8 and x_{11} impact y in reverse direction whereas model fitted with six regressors witnessed the reverse impact by x_3, x_4 and x_6. This imply that if we attempt to increase these factors by one unit that are result into corresponding decrease in ARDS cases. The above equations may be used for forecasting ARDS cases with $100(1-\alpha)$ confidence level.

4.3.1 *Model diagnosis and validations*

We have attempted to investigate the validation of model using residual analysis, test of linearity, normality, homoscedasticity and independent of errors in the model using error analysis and error plots. Since normal probability plot gives a straight line pattern; hence normality seems to be preserved as shown in Figure 5. An overall view of the standardized error plots displayed by Figure 6 may be obtained by examining the standard residual versus fitted value plot. The absence of systematic patterns suggested the randomness of distribution of errors. Also, we do not found any cone shaped patterns. These facts support our models premises that the error terms are randomly distributed and preserve constant variance (homoscedastic property). Table 8 presents computed values of Q-statistic, d-statistic (Durbin Watson test statistic) and variance inflation factor (VIF) for all four transformed models. A hypothesis testing is performed for each models. Since computed Q values as shown in the column 2 of Table 8 for respective models are smaller than the critical value of Z at $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis in all four cases which means that there is no violation of assumption of homoscedasticity (constant variance) and absence of heteroscedasticity is confirmed. The values of Durbin-Watson test statistic is four transformed models shows absence of problem of autocorrelation in all the four models under the study. Since in all four cases our VIF values are below 10. Therefore, we do not preserve serious multicollinearity in analysis of ARDS modelling.

Table 8. Test statistic and VIF value for transformed models

Transformations	Q-statistic	(DW) d-statistic	VIF
$\log y \sim \log x$ (12 variables)	-0.3416	1.9791	2.2646
$\log y \sim \log x$ (6 variables)	-0.9089	1.9587	1.4598
$\sqrt{y} \sim \sqrt{x}$ (12 variables)	0.1030	1.9541	1.6841
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	-0.2517	2.0000	1.3300

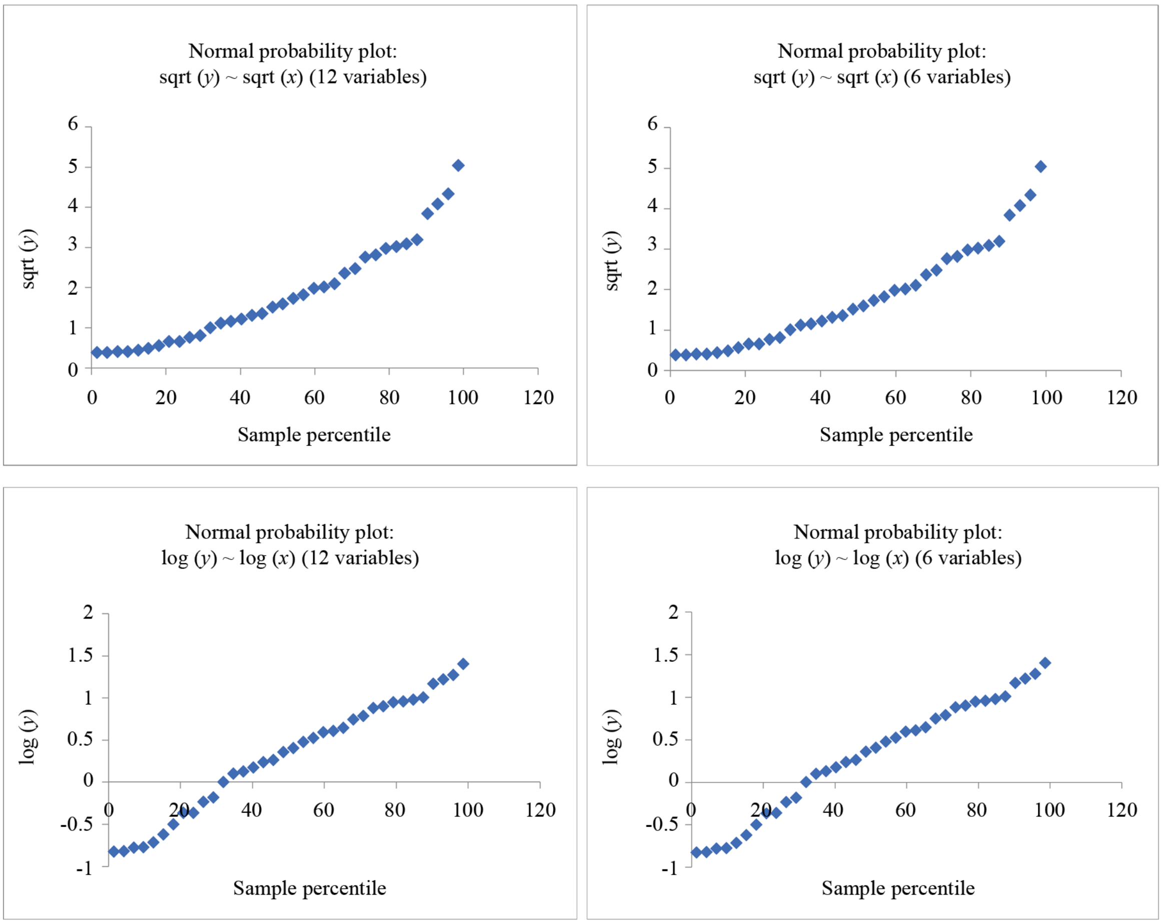


Figure 5. Normal probability plot for all transformed models

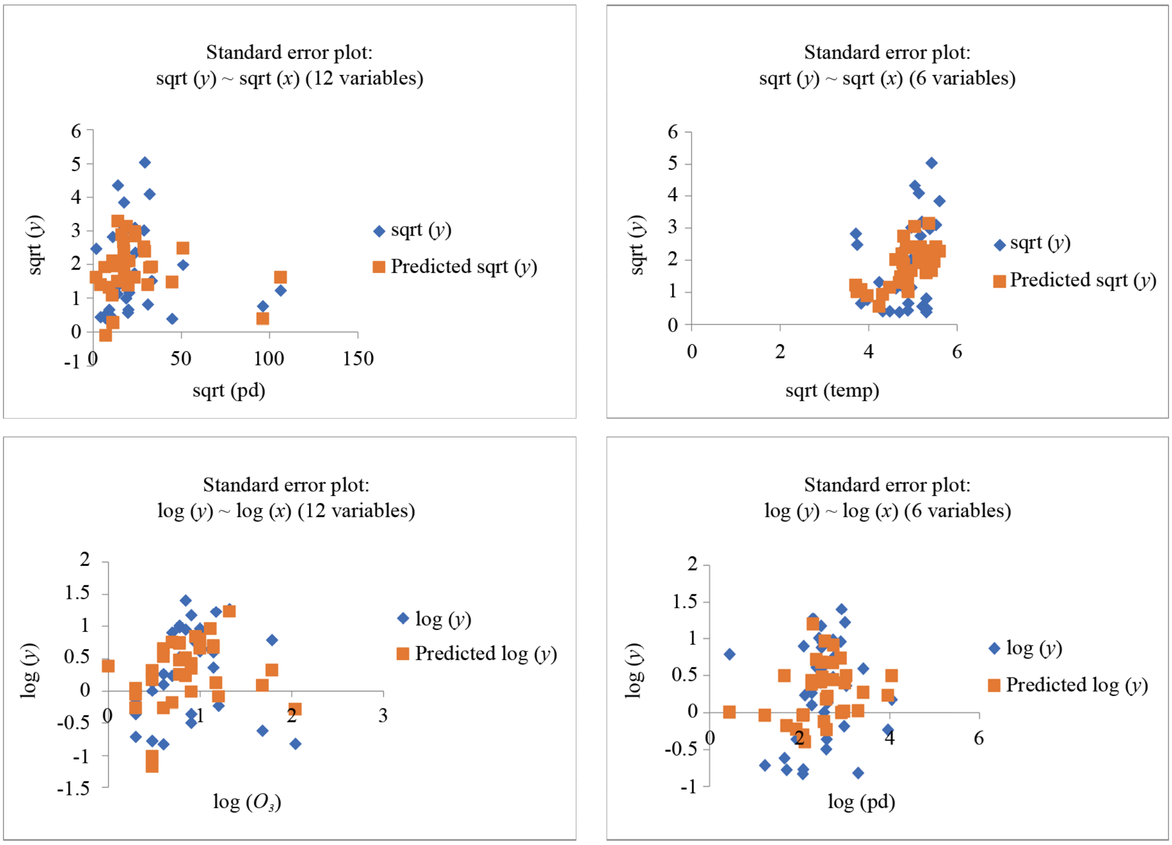


Figure 6. Standard error plot

4.4 *Model performance criteria*

Table 9 presents the computed values of R^2 and adjusted R^2, PRESS statistic and R^2 for predicted of various transformation model. Table 10 presents the model selction values of standard statistical criteria namely akaike information criterion (AIC), schwarz bayesian information criterion (SBIC), C_p (Mallow's C_p), mean square error (MSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE). The value of spearman's rank correlation between two attributes such as goodness of fit and model performance is 0.7 lies between 0 & 1, that shows the statistically significant relationship between the goodness of fit of the model and performance of the model for ARDS cases, keeping in the view of above discussion though any of the transformed model will a good job.

Table 9. Computed values of statistics for transformed models

Transformations	R^2	R^2_{adj}	PRESS	R^2_{PRESS}
log$y$ ∼ log$x$ (12 variables)	0.5584	0.3280	43.3629	-1.7022
log$y$ ∼ log$x$ (6 variables)	0.3149	0.1732	26.7111	0.6645
$\sqrt{y} \sim \sqrt{x}$ (12 variables)	0.4062	0.0964	132.5126	-1.3755
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	0.2481	0.0925	104.7099	-0.8783

Table 10. Model selection criteria

Transformations	MSE	MAD	MAPE	AIC	SBIC	C_p
log$y$ ~ log$x$ (12 variables)	0.1968	0.3521	3.7514	-32.5164	-11.9305	15
log$y$ ~ log$x$ (6 variables)	0.3053	0.4408	2.7455	-28.7087	-17.6241	9
$\sqrt{y} \sim \sqrt{x}$ (12 variables)	0.9194	0.7873	0.7023	22.9777	43.5635	15
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	1.1642	0.8653	0.7829	19.4744	30.5591	9

Table 11. Confidence interval and confidence length for the estimated coefficients

Name of variables	Modelling*	Standard error	Lower 95.0%	Upper 95.0%	Confidence length	FR*	OFR*
sqrt (PM (2.5))	1	0.2153	-0.0210	0.8699	0.8910	4	9
sqrt (PM (10))	1	0.2302	-0.9205	0.0320	0.9526	5	10
sqrt (SO_2)	1	0.4168	-0.9358	0.7888	1.7246	10	22
sqrt (CO)	1	0.0341	-0.0425	0.0988	0.1413	2	3
sqrt (O_3)	1	0.1708	-0.3960	0.3109	0.7069	3	5
sqrt (NO_2)	1	0.2939	-0.3479	0.8684	1.2163	7	16
sqrt PD	1	0.0157	-0.0338	0.0313	0.0652	1	2
sqrt (Urbanisation)	1	0.2449	-0.8413	0.1721	1.0134	6	13
sqrt (Malnutrition)	1	0.3407	-1.0923	0.3175	1.4099	9	19
sqrt (Temperature)	1	0.5635	-0.3269	2.0047	2.3316	11	28
sqrt (Humidity)	1	0.3287	-1.2287	0.1314	1.3601	8	17
sqrt (PCI)	1	0.8276	-2.2470	1.1772	3.4243	12	31
log PM (2.5)	2	0.6093	-0.0920	2.4291	2.5211	9	30
log PM (10)	2	1.0139	-4.6454	-0.4504	4.1950	10	32
log (SO_2)	2	0.4576	-1.2451	0.6481	1.8932	5	25
log CO	2	0.4166	0.1327	1.8565	1.7237	3	21
log (O_3)	2	0.3298	-0.6810	0.6836	1.3647	2	18
log (NO_2)	2	0.4669	-0.5447	1.3872	1.9319	7	27
log PD	2	0.2028	-0.4864	0.3528	0.8392	1	8
log Urbanisation	2	0.5666	-2.0730	0.2712	2.3443	8	29
log Malnutrition	2	0.4651	-0.8826	1.0417	1.9243	6	26
log Temperature	2	1.4267	-2.5275	3.3753	5.9029	12	36
log Humidity	2	1.2865	-5.5337	-0.2107	5.3230	11	35
log PCI	2	0.4491	-0.9043	0.9538	1.8581	4	24
sqrt (O_3)	3	0.1132	-0.1825	0.2805	0.4631	2	4
sqrt (NO_2)	3	0.2372	-0.3256	0.6445	0.9702	4	11
sqrt (PD)	3	0.0110	-0.0367	0.0084	0.0452	1	1
sqrt (Malnutrition)	3	0.1849	-0.4564	0.3002	0.7566	3	6
sqrt (Temperature)	3	0.4528	-0.2744	1.5780	1.8524	6	23
sqrt (Humidity)	3	0.2765	-0.9863	0.1450	1.1313	5	15
log (O_3)	4	0.2659	-0.3427	0.7451	1.0878	3	14
log (NO_2)	4	0.3540	-0.4500	0.9983	1.4483	4	20
log PD	4	0.1954	-0.5820	0.2174	0.7994	1	7
log Malnutrition	4	0.2458	-0.5141	0.4913	1.0054	2	12
log Temperature	4	1.2697	-1.0919	4.1016	5.1936	6	34
log Humidity	4	1.2269	-5.1945	-0.1757	5.0187	5	33

Modelling* 1 is log$y$ ~ log$x$ (10 variables), 2 is log$y$ ~ log$x$ (6 variables), 3 is $\sqrt{y} \sim \sqrt{x}$ (10 variables), 4 is $\sqrt{y} \sim \sqrt{x}$ (6 variables), FR* is factor ranking for different models, OFR* is overall factor ranking

Table 11 presents the lower and upper confidence intervals for coefficients associated with the independent regressors for each type of transformed model and corresponding ranks have been assigned. Also, we have attempted to find out the ranking of each of the factors through four transformed models such a ranking is useful for understanding factor-wise significant under each of the four categories of the transformed model. The non-pathogenic factors responsible for ARDS cases and their growth in India could be PM (2.5), CO, relative humidity, NO_2, and SO_2. Therefore, our analysis provides strong evidence to support the theory that the role of air pollutants and relative humidity followed by population density, increasing ozone level (O_3), temperature and malnutrition.

5. Conclusion

To conclude with, we may say that there are almost top ten states of India that is the hotspot of ARDS cases. The geographical location of the hotspots are visualised in the figure. As such, we have pinpointed the locations witnessing highest concentration of ARDS cases during the period of study 2006-2021; the resultant hotspot states needs thorough investigations on urgent priority to achieve optimal as well as feasible solutions of the problems at the moment. We found twelve factors namely population density, urbanisation, temperature, relative humidity, per capita income, malnutrition, NO_2, PM (2.5), O_3, SO_2, PM (10), CO out of which six factors viz population density, temperature, relative humidity, malnutrition, O_3 and NO_2 are the pertinent one. It is being evidenced through our modelling process that with sincere efforts we may control these factors and able to apply the break on ARDS infections and its growth. We become successful in obtaining the best-fit model as shown in the equations (12)-(15). Model performance has been explained satisfactorily and presented in the Table 9. Though, the current work is quite comprehensive in nature but have a complete clarity in regards to factors identification and hotspot determination. The obtained results suggest that we should tackle the primary clusters (MLC) states first and then concentrate on states under secondary and tertiary clusters too. We learned that there is a need of launching an effective ARDS awareness programme through out the country. Also, we should focus on implementation of resource utilization unbiasedly, efficiently, consistently, and sufficiently. With a higher probability confidence level, we may foresee the fact and findings of our work can be generalized in any part or at least in some part of the world. In conclusion, we may recommend strongly that ARDS cases in India are exclusively due to man-made factors and not by chance, hence can be eliminated effectively with the highest possible accuracy and reliability. This will boost our confidence and morals for achieving complete success towards the set goal. We may anticipate that the hotspot states of ARDS cases if tackled effectively, would greatly help achieve our goal of the complete eradication of ARDS cases in India.

Future scope

The present work provide ample opportunity to pursue further in-depth research at micro levels (districts, sub-divisions, blocks, panchayats and villages). The people from expert domain (bio-medical professionals) and health care specialist will derive, frame, and develop an effective decision making tools to combat ever alarming ARDS cases from India.

Acknowledgement

The Vellore Institute of Technology, Vellore, has been acknowledged by the authors for giving us the resources we needed to complete this research project effectively.

Funding

There was no particular grant awarded for this research by governmental, private, or non-profit funding organizations.

Data availability

The datasets used and analyzed in this study are available from the corresponding author upon reasonable request.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. *International Journal of Stroke*. 2022; 17(1): 18-29. Available from: https://doi.org/10.1177/17474930211065917.

[2] Rushworth A, Lee D, Mitchell R. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-Temporal Epidemiology*. 2014; 10: 29-38. Available from: https://doi.org/10.1016/j.sste.2014.05.001.

[3] Yuan Q, Qi B, Hu D, Wang J, Zhang J, Yang H, et al. Spatiotemporal variations and reduction of air pollutants during the COVID-19 pandemic in a megacity of Yangtze River Delta in China. *Science of the Total Environment*. 2021; 751: 141820. Available from: https://doi.org/10.1016/j.scitotenv.2020.141820.

[4] Wu Y, Song P, Lin S, Peng L, Li Y, Deng Y, et al. Global burden of respiratory diseases attributable to ambient particulate matter pollution: findings from the global burden of disease study 2019. *Frontiers in Public Health*. 2021; 9: 740800. Available from: https://doi.org/10.3389/fpubh.2021.740800.

[5] Leining LM, Gatchalian SR, Gunter SM, Castillo-Carandang NT, Mandalakas AM, Cruz AT, et al. Geospatial and hot spot analysis of pediatric tuberculosis infection in Bohol, Philippines. *Epidemiology & Infection*. 2020; 148: e89. Available from: https://doi.org/10.1017/S0950268820000795.

[6] Xie M, Liu X, Cao X, Guo M, Li X. Trends in prevalence and incidence of chronic respiratory diseases from 1990 to 2017. *Respiratory Research*. 2020; 21: 49. Available from: https://doi.org/10.1186/s12931-020-1291-8.

[7] Dlamini SN, Beloconi A, Mabaso S, Vounatsou P, Impouma B, Fall IS. Review of remotely sensed data products for disease mapping and epidemiology. *Remote Sensing Applications: Society and Environment*. 2019; 14: 108-118. Available from: https://doi.org/10.1016/j.rsase.2019.02.005.

[8] Kim D, Chen Z, Zhou LF, Huang SX. Air pollutants and early origins of respiratory diseases. *Chronic Diseases and Translational Medicine*. 2018; 4(2): 75-94. Available from: https://doi.org/10.1016/j.cdtm.2018.03.003.

[9] Jung I. Spatial scan statistics for matched case-control data. *Plos One*. 2019; 14(8): e0221225. Available from: https://doi.org/10.1371/journal.pone.0221225.

[10] Anjali B, Kumar J. Spatio-temporal aspect of suicide and suicidal ideation: An application of SaTScan to detect hotspots in four major cities of Tamil Nadu. *Journal of Scientific Research*. 2021; 65(9): 7-18. Available from: https://doi.org/10.37398/JSR.2021.650902.

[11] Balasubramani K, Prasad KA, Kodali NK, Abdul Rasheed NK, Chellappan S, Sarma DK, et al. Spatial epidemiology of acute respiratory infections in children under 5 years and associated risk factors in India: District-level analysis of health, household, and environmental datasets. *Frontiers in Public Health*. 2022; 10: 906248. Available from: https://doi.org/10.3389/fpubh.2022.906248.

[12] Fatima M, Khattak RM, Grady SC, Butt I, Arshad S, Ittermann T, et al. Spatial and temporal analysis of acute respiratory infections (Aris) in southern Punjab, Pakistan. *Spatial Information Research*. 2022; 30: 477-487. Available from: https://doi.org/10.1007/s41324-022-00447-4.

[13] Wang F, Duan C, Li Y, Huang H, Shia BC. Spatiotemporal varying coefficient model for respiratory disease mapping in Taiwan. *Biostatistics*. 2024; 25(1): 40-56. Available from: https://doi.org/10.1093/biostatistics/kxac046.

[14] Kaindal S, Venkataramana B, Kumar J. Cancer hotspot identification and analysis: A scan statistics approach. In *International Conference on Information Technology*. Singapore: Springer; 2023. p.13-28.

[15] Katale RN, Gemechu DB. Spatio-temporal analysis of malaria incidence and its risk factors in North Namibia. *Malaria Journal*. 2023; 22: 149. Available from: https://doi.org/10.1186/s12936-023-04577-4.

[16] Kumar BR, Anjali. Spatial analysis of multivariate factors influencing suicide hotspots in Urban Tamil Nadu. *Journal of Affective Disorders Reports*. 2024; 16: 100741. Available from: https://doi.org/10.1016/j.jadr.2024.100741.

[17] Manish GV, Simran, Kumar J, Choubey DK. Identification of hotspot of rape cases in NCT of Delhi: A data science perspective. In *International Conference on Information Systems and Management Science*. Cham: Springer; 2021. p.485-496.

[18] Saravag PK. An application of scan statistics in identification and analysis of hotspot of crime against women in Rajasthan, India. *Applied Spatial Analysis and Policy*. 2024; 2024: 1-20. Available from: https://doi.org/10.1007/s12061-024-09572-z.

[19] Tesfaye SH, Seboka BT, Sisay D. Spatial patterns and spatially-varying factors associated with childhood acute respiratory infection: data from Ethiopian demographic and health surveys (2005, 2011, and 2016). *BMC Infectious Diseases*. 2023; 23(1): 293. Available from: https://doi.org/10.1186/s12879-023-08273-1.

[20] Abolhassani A, Prates MO. An up-to-date review of scan statistics. *Statistic Surveys*. 2021; 15: 111-153. Available from: https://doi.org/10.1214/21-SS132.

[21] Mondal S, Singh D, Kumar R. Crime hotspot detection using statistical and geospatial methods: A case study of Pune City, Maharashtra, India. *GeoJournal*. 2022; 87: 5287-5303. Available from: https://doi.org/10.1007/s10708-022-10573-z.

[22] Montnemery P, Nihlén U, Göran Löfdahl C, Nyberg P, Svensson Å. Prevalence of self-reported eczema in relation to living environment, socio-economic status and respiratory symptoms assessed in a questionnaire study. *BMC Dermatology*. 2003; 3: 4. Available from: https://doi.org/10.1186/1471-5945-3-4.

[23] Saran S, Singh P, Kumar V, Chauhan P. Review of geospatial technology for infectious disease surveillance: use case on COVID-19. *Journal of the Indian Society of Remote Sensing*. 2020; 48: 1121-1138. Available from: https://doi.org/10.1007/s12524-020-01140-5.

[24] Sukhija K, Singh SN, Kumar J. Spatial visualization approach for detecting criminal hotspots: An analysis of total cognizable crimes in the state of Haryana. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. Bangalore, India: IEEE; 2017. p.1060-1066.

[25] Kulldorff M. A spatial scan statistic. *Communications in Statistics-Theory and Methods*. 1997; 26(6): 1481-1496. Available from: https://doi.org/10.1080/03610929708831995.

[26] Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2001; 164(1): 61-72.