Geospatial analysis of contagious infection growth and cross-boundary transmission in non-vaccinated districts of North-East Indian states during the COVID-19 pandemic

Mousumi Gupta,[a] Madhab Nirola,[a,*] Arpan Sharma,[a] Prasanna Dhungel,[a] Harpreet Singh,[b] and Amlan Gupta[c]

[a]Department of Computer Applications, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar, 737136, India
[b]Division of Biomedical Informatics, Indian Council of Medical Research, Delhi, 110029, India
[c]Department of Transfusion Medicine, Jay Prabha Medanta Super Speciality Hospital, Patna, 800020, India

Summary

Background During the initial phase of the COVID-19 pandemic, the Government of India implemented a nationwide lockdown, sealing borders across states and districts. The northeastern region of India, surrounded by three international borders and connected to mainland India by a narrow passage, faced particular isolation. This isolation resulted in these states forming a relatively closed population. Consequently, the availability of population-based data from Indian Council of Medical Research, tracked through national identification cards, offered a distinctive opportunity to understand the spread of the virus among non-vaccinated and non-exposed populations. This research leverages this dataset to comprehend the repercussions within isolated populations.

Methods The inter-district variability was visualized using geospatial analysis. The patterns do not follow any established grounded theories on disease spread. Out of 7.1 million total data weekly 0.35 million COVID-19-positive northeast data was taken from April 2020 to February 2021 including "date, test result, population density, area, latitude, longitude, district, and state" to identify the spread pattern using a modified reaction-diffusion model (MRD-Model) and Geographic Information System.

Findings The analysis of the closed population group revealed an initial uneven yet rapidly expanding geographical spread characterized by a high diffusion rate α approximately 0.4503 and a lower reaction rate β approximately 0.0256, which indicated a slower growth trajectory of case numbers rather than exponential escalation. In the latter stages, COVID-19 incidence reached zero in numerous districts, while in others, the reported cases did not exceed 100.

Interpretation The MRD-Model effectively captured the disease transmission dynamics in the abovementioned setting. This enhanced understanding of COVID-19 spread in remote, isolated regions provided by the MRD modelling framework can guide targeted public health strategies for similar isolated areas.

Funding This study is Funded by Indian Council of Medical Research (ICMR).

Keywords: COVID-19; Geospatial analysis; Lockdown; Northeast India; Modified Reaction-Diffusion model

Introduction

The severe acute respiratory syndrome caused by coronavirus SARS-CoV-2[1] spread as a pandemic throughout the globe during the years 2019, 2020, 2021 and 2022. There were periods when the vaccine was not available and then a period when majority of the people had been vaccinated during the different phases of pandemic. Disease transfer in isolated regions with minimal floating populations has not been studied in detail to understand the disease transfer dynamics. The northeast states of India which are connected to the mainland by an isthmic-like land strip and have impervious international borders. The dynamics of spread would give an idea on mitigation procedures that may be employed to contain any new similar viruses that appear in future. It is a good approach to see how pandemic proceeds through isolated regions during lockdown. Intuitively, if disease transmission between districts resembled a diffusion process, stringent border closures would likely have curbed its spread more

*Corresponding author.
  E-mail address: kishorniroula0@gmail.com (M. Nirola).

Research in context

Evidence before this study
A search was made in the PubMed with following query "("Geographic Information Systems" [Mesh] OR "Geographic Mapping" [Mesh]) AND "COVID-19/epidemiology" [Mesh]" resulting in 138 articles on 25-05-2024. Previous studies on the epidemiology of COVID-19 in India have primarily relied on national-level data to model transmission dynamics. While insightful, such analyses do not capture dynamics in isolated population groups with limited external connectivity. A few studies have used spatial modelling approaches to characterize COVID-19 spread across different regions within India. However, population-level data from northeastern states providing an epidemiologically closed setting had not been utilized to understand natural transmission patterns before the introduction of vaccines or mass immunity.

Added value of this study
To our knowledge, this is the first study to analyze COVID-19 testing data from the isolated population of northeastern India spanning April 2020–February 2021, before vaccines were widely available. By applying a geographic information system and modified reaction-diffusion modelling, we were able to characterize the inter-district spread and temporal evolution of SARS-CoV-2 in this unique epidemiological setting. Our analysis provides novel insights on the initial transmission dynamics and decline of the virus in a non-vaccinated population with limited external influence.

Implications of all the available evidence
Our findings have important implications for understanding COVID-19 transmission in isolated regions with minimal external connectivity. The rapid initial dissemination followed by slower growth trajectory identified in our models enhances theoretical frameworks for predicting spread in isolated settings. The successful application of geospatial and mathematical modelling approaches also demonstrates their utility for guiding public health strategies in similar populations worldwide. Ongoing and future studies incorporating these elements could further refine response plans for remote areas during epidemics.

effectively. However, if it followed a Brownian motion model, then prioritizing contact isolation would have been crucial. Such studies having a robust model framework would aid policymakers for informed policy decisions. Up to March 2022, there have been over 456 million confirmed cases of COVID-19 worldwide, including over 6 million deaths reported to WHO.[2] India has witnessed several waves of COVID-19 infection since the start of the pandemic.[3] These states of India are bounded by international borders of Bhutan, Nepal, Bangladesh, and Myanmar. The first reported cases of COVID-19 in the northeastern states of India occurred sometime later approximately 3 months later than in other parts of the country i.e., first case in India on January 27, 2020,[4] and first case in the Northeast was during April, 2020.[4] Four contiguous lockdowns were implemented in India from 25th March 2020 to 31 May 2020 which was also implemented by northeast India. Further, the lockdown was followed by 3 unlock phases and completely unlocked by 31st July.[5] The lag time form most new are novel viruses are not known in the start but can be calculated with a range in the early part of the pandemic as was done for Covid.[6] The incubation period was labelled as and still is labelled as 2–14 days and a median estimate was used as seven days. Our study revealed that incubation period did have an effect in the initial spread of sporadic disease but not significantly in the later stages of disease where a regional variation within the area studies.

This study is significant because we identified how the virus spreads in the absence of vaccination within closed population. Understanding how the virus spreads within such a small, closed population offers vital information about the virus's natural progression. This study emphasizes how important regional differences are in determining how the virus spreads, and how these variations are shaped by things like population density, regional mobility, and public health infrastructure. It also emphasizes how crucial it is to take COVID-19's incubation period into account when analyzing transmission dynamics since lag-time can facilitate cross-border spread even in the face of strict lockdown protocols. By analyzing data over a substantial period, this study offers valuable observations on the temporal impacts of containment and mitigation strategies, providing practical recommendations for enhancing public health responses in isolated and semi-closed populations globally.

Further, it has been observed that sub-national or state-tailored policies in Northeast India have not been effectively utilized for regional COVID-19 transmission control. Public health falls under the 'state list' as per Article 246 of the 7th Schedule of the Indian Constitution, although, it has been discerned that, since it was a global pandemic, sub-national policies focused more on a centralized approach. This centralized approach, supported by state governments, contrasted with the typically decentralized nature of public health governance in India.[7] Risk-benefit assessments and considerations by local communities on transmission spread were not widely considered due to the global nature of the pandemic, bypassing state-level governance. Notwithstanding any biases, almost all state governments agreed to follow nationalized protocols.

On the other hand, population density has been a critical factor in the transmission dynamics of COVID-19. The northeastern states have relatively lower

population density than other states of India. Studies from various regions, including South-East Asia and the Western-Pacific, have shown that densely populated urban areas experienced more rapid spread compared to rural areas. For instance, research conducted in countries like India and the Philippines indicated that metropolitan areas saw quicker outbreaks and higher infection rates due to the concentration of people in small areas.[8,9] In this purview, achieving critical vaccination thresholds is essential for controlling the spread of COVID-19, particularly in high-density regions. In densely populated areas, the threshold for herd immunity is higher because of the increased contact rates. Studies in the Western-Pacific region, such as in Singapore and Japan, have highlighted the importance of rapid and widespread vaccination campaigns to achieve these thresholds and effectively control outbreaks.[10,11]

As discussed, the northeastern region of India consists of five international boundaries which had limited connectivity and movement during the initial phases of the pandemic due to border closures.[12] Previous studies on COVID-19 spread in India have primarily focused on modelling transmission dynamics at national and subnational levels using large metropolitan datasets.[13] However, population-level data from isolated regions with minimal floating populations and limited external connectivity could provide novel insights into the natural transmission dynamics of SARS-CoV-2 in a non-vaccinated setting. Here, we analyze COVID-19 testing data from the northeastern states of India spanning April 2020–February 2021, before the wide availability of vaccines, to characterize the inter-district spread of SARS-CoV-2 using a geographic information system and modified reaction-diffusion modelling approach.

This analysis aims for a better understanding of the epidemiological characteristics of isolated regions and ensure more effective containment and mitigation strategies for similar populations worldwide. Specifically, the analysis will look at the patterns of COVID-19 transmission in the northeastern states, assess the effect of national-level lockdowns on local transmission rates, and provide actionable recommendations for modifying public health interventions to regional contexts.

Methods
Overview
As per the GATHER guidelines[14] this study looked at the population densities of the different districts in the 7 contiguous states and a single non-contiguous state (Sikkim) which falls under the northeast council of Indian Government. The Pan-India covid testing data was pooled in ICMR repositories right from the start of the pandemic India by government mandate, hence the ICMR data was acquired. The total population in this semi-closed population as per 2011 census data are:

Assam- 3,12,05,576; Tripura- 36,73,917; Meghalaya- 29,66,889; Nagaland- 19,78,502; Manipur- 28,55,794; Arunachal Pradesh- 13,83,727; Mizoram- 10,97,206; and Sikkim- 6,10,577. The district population distribution is given in the supplementary data (Appendix p 236). For transmission of contagious diseases, the age and sex details are irrelevant. The time frame studied was 1st week of March 2020–4th week of February 2021. In January 2021, the first lot of vaccines[15] had arrived and the transmission after the vaccination not been worked in this study. The source of data as well as the funding for this work including machine and software were provided by the Indian Council of Medical Research (ICMR) under the Granting head "2021-6393". The data was single source ICMR repository, JSON files were provided through application programming interface (API). Data were obtained in the JSON format, then converted to the table $m \times n$ format using Microsoft SQL Server. The data was uniquely identified as individuals based on Adhaar as a biometric identifier of all individual citizens of India. The Adhaar numbers were masked by ICMR to prevent identification of individuals by external sources so that individual privacy and confidentiality are maintained. It was passed as satisfactory by the institutional ethical committees. All positive data in the repository were considered for northeast India based analysis. Further, no exclusions were made, since it was a repository of all tests, no selection was made for sample size computation.

The data as a whole can be regarded as unbiased and representative, given that mandatory and multiple mechanisms were implemented to ensure that all tests, including antigen tests and self-test kits, which were uploaded to the ICMR database. Mechanisms were established for states utilizing their own apps for COVID testing data collection. It is acknowledged that symptomatic testing was primarily conducted during the early phases, owing to limited testing facilities and kits following the initial guidelines in India. However, starting from March 2020 onwards i.e., one month before the first case in northeast India, with the in-house development of testing kits and the introduction of private labs, testing was expanded to include a significant number of asymptomatic individuals as well. While there may be extremely few instances where tests were not uploaded, considering that the ICMR database served as the foundation for insurance and other legal procedures, it is presumed to be substantially unbiased and representative.

Data was collected week-wise for the study because (WU et al. 2022)[6] determined the incubation period for COVID-19 to be from 2 days to 22 days in the mean of 6.57 days. The incubation periods of COVID-19 caused by the Alpha, Beta, Delta, and Omicron variants were 5.00, 4.50, 4.41, and 3.42 days, respectively. Hence, the mean incubation was 7.43 days i.e., around one week. The week-wise data collected was then used for geospatial analysis and mathematical analysis. The entire

workflow diagram of our study is demonstrated in Fig. 1. The data used for our analysis were Date_Time, District_Name, State_Name, Final_Test_Result, Total Population, Population Density, Area, Latitude and Longitude. The Total Population, Population Density and Area these data were taken from the Indian official census website. Latitude and Longitude were generated using google API services. The software used for this study included Anaconda, ArcGIS Pro, Microsoft SQL Server, Microsoft Excel.

Mathematical modeling

As discussed, this study examines the week-wise cumulative data to analyze the spread of the initial COVID-19 wave across districts during the lockdown phase. The pattern was visualized using ArcGIS Pro software, and various models such as Brownian and Diffusion were tested using Python programming, but none yielded significant results. Different regression approaches were also explored, as discussed further in the "Regression Modelling and Sensitivity Analysis" and "Spatial Data Analysis" subsections below. However, one of the other candidate models to which this data was subjected for pattern of spread included a logistic growth model,[16] but the COVID-19 disease spread

growth rate predicted by the models was exceptionally low. Further, we tried to modify and implement reaction-diffusion model for identifying the Northeast COVID-19 spread pattern. The reaction-diffusion model is based on the Fisher-KPP equation, created by Fisher.[17] Though it is an early model with many intermittent modifications, we found this model reasonably matches with the pattern seen in consecutive visual display in maps. This Fisher-KPP equation explains how a group of organisms moves and grows in a particular area. It shows how they spread out by moving randomly and reproducing in their local surroundings. This equation provides a basic theoretical framework for modelling disease transmission dynamics.[18] The traditional Fisher-KPP is a basic model and might not consider all the complexities involved in the real world, making it less accurate in certain situations like actual disease outbreaks. To address this, we used the Fishers-KPP equation with slight modifications as per our data availability. In (Appendix pp 1 and 2) we have shown a detailed elaboration of how we modified the Fishers-KPP equation. After modifying the Fishers-KPP equation we call it Modified Reaction-Diffusion Model (MRD Model).

The MRD model is particularly found useful because it examines COVID-19 transmission dynamics across isolated regions with limited connectivity in Northeast India. Specifically, the carrying capacity term used in our MRD model $(1-u/K)$ allows the growth rate to saturate as the infected population approaches the maximum capacity K set by environmental constraints. By incorporating geographic factors through the diffusion term and accounting for finite carrying capacity, this MRD model provides a more accurate representation of disease spread compared to simple compartmental models or uncontrolled exponential growth models. This enhances our ability to interpret the epidemiological patterns observed in this isolated population. The MRD model is an appropriate and effective theoretical framework for analyzing the district-level COVID-19 transmission dynamics in this study region. It addresses some limitations of basic reaction-diffusion formulations while retaining their advantages for spatial–temporal modelling of infectious disease spread.

Data analysis

Initially, the weekly segregated data was encoded as integer encoding, where each unique category value is assigned an integer value. Then the encoded data was used to find a relationship between population density and positive COVID-19 cases. So, we performed regression analysis. After performing the regression analysis, we applied simulation or looping, to simulate the progression of the disease within a specified region over time (week-wise). The progression of the disease was simulated using our derived MRD model.
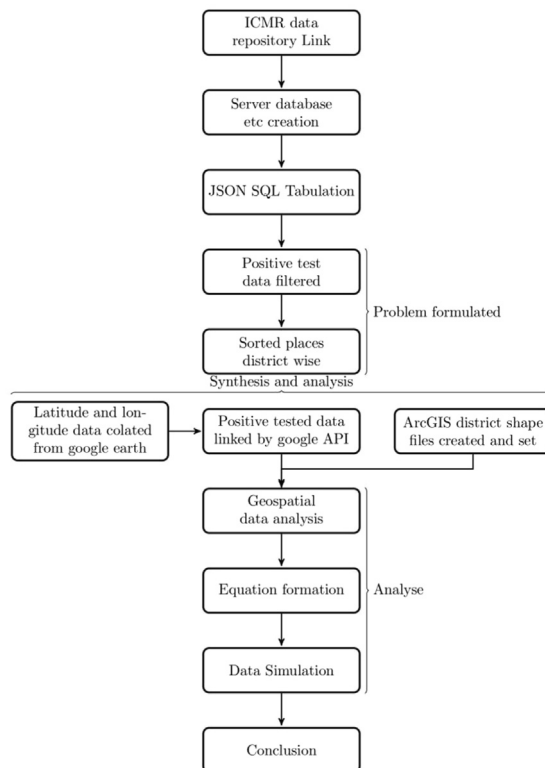

Fig. 1: **This graphical representation shows the different stages of the workflow.**

The looping method is integral to our analysis for parameter estimation within the Modified Reaction-Diffusion (MRD) model. This method was embedded within the simulation framework to iteratively optimize the model parameters (α, β, and K) for the best fit to our observed COVID-19 case data. The simulation begins with initial guesses for the diffusion rate (α), reaction rate (β), and carrying capacity (K). These initial values were $\alpha = 0.1$, $\beta = 0.01$, and $K = 1000.0$, chosen based on preliminary analysis and literature review. The looping method iterates through each time step (weekly intervals) to calculate the infected population density using the MRD equation. The equation incorporates both spatial diffusion and local reaction dynamics. At each iteration, the infected population density is updated based on the density from the previous week. This iterative approach allows the model to simulate the temporal evolution of the infection across the geographic area. For each iteration, the negative log-likelihood is computed to measure the fit between the simulated infected density and the observed data. This is crucial for assessing the accuracy of the model at each time step. The formula used to calculate the negative log-likelihood is based on the assumption of normally distributed errors and is given in (Appendix p 2).

After this process the differential evolution function is used to find the parameters (α, D, K) that minimize the negative log-likelihood, effectively maximizing the likelihood of the observed data given the model. The optimization algorithm explores the parameter space (within defined bounds) to find the optimal values. The Differential evolution function performs mutation by creating a mutant vector V based on three randomly chosen individuals (A, B, C). Where A, B, and C are three distinct individuals (sets of parameters) from the district population. The mutant vector (V) is calculated for each parameter separately using the formula present in (Appendix p 2).

After this, a crossover operation combines the original values generated by the simulation loop with the mutant vector to create a trial candidate solution. This is done for each parameter and follows a binomial crossover process. For each parameter it generates a random number between 0 and 1 if this random number is less than the crossover probability (a user-defined parameter), take the corresponding parameter from the mutant vector; otherwise, keep the original parameter. Finally, it performs a selection operation to determine which candidate solutions survive based on their fitness (negative log-likelihood). The better-fit solutions are chosen to be part of the updated population. It also compares the negative log-likelihood (fitness) of each trial candidate solution with that of the corresponding original candidate solution and if the trial candidate solution has a lower negative log-likelihood (better fitness), it replaces the original candidate solution. The mutation, crossover, and selection steps are performed for each candidate solution in the population, and the updated population becomes the values for the next iteration. This iterative process continues for a certain number of iterations or until a convergence criterion is met. The result of the optimization is the set of parameters (α, D, K) that minimize the negative log-likelihood, giving us the best-fit model for the observed data.

Spatial data analysis

In this spatial analysis various mapping techniques[19] were employed, including Quantile Maps, Normalized Maps, Bivariate Maps, Spatial Autocorrelation, and Empirical Bayesian Kriging (EBK), to effectively visualize and interpret the data. Quantile maps were generated to display the spatial distribution of COVID-19 cases by categorizing data into quantiles. This allows for the comparison of case numbers across different districts. Each color on the map represents a quantile range, making it easier to identify patterns and disparities in case distribution. Normalized maps were created to illustrate the concentration of positive COVID-19 cases relative to the total population across various districts in Northeastern India. These maps were created to provide a clear visual representation of how the number of cases varies in proportion to the population size, helping to identify areas with higher relative infection rates. Bivariate maps were employed to visually represent the distribution of COVID-19 positive cases in relation to the total population across different districts in Northeastern India. This method provides a dual-variable visualization, highlighting critical areas that require attention and offering insights that can inform public health policies and response strategies.

We performed spatial autocorrelation[20] to identify the Moran's Index value, z-score, p-value in our weekly COVID-19 Northeast positive data to visualize if data are dispersed, clustered or random. The results indicated a random distribution. Hence, Simultaneous Autoregressive Model (SAR), and Conditional Autoregressive Model (CAR) were excluded from the analysis. We used our whole Northeast COVID-19 positive data to compute Empirical Bayesian Kriging. Empirical Bayesian Kriging (EBK)[21] is a method used to predict how certain variables are spread out in space. Unlike traditional methods, EBK considers the uncertainty in the model used to describe how data points relate to each other over distance. It uses Bayesian statistics to create a reliable estimate of spatial patterns, making it especially useful for data that doesn't follow a consistent pattern or has varying degrees of similarity between points. The results generated by the EBK method are attached in appendix as Supplement 7 (eEBKmaps and results).

Regression modelling and sensitivity analysis

Regression analysis was also considered in this study to understand how the MRD model may be significant or

less significant in an isolated population. Different polynomial and exponential orders to identify the relationship between the population density and COVID-19-positive cases was considered initially. Further, Ordinary Least Squares (OLS) regression[22] was considered due to less computational complexity. OSL analysis is a statistical method used to estimate the relationship between one or more independent variables and a dependent variable. In OLS regression analysis, number of positive COVID-19 cases was considered as the dependent variable and the population of each district as the independent variable. Further, Geographically Weighted Regression (GWR)[22] was considered which is a spatial analysis technique used to model spatially varying relationships in data rather than fitting a single, global regression model to the entire dataset. GWR creates separate regression equations for each target feature based on the dependent and explanatory variables of nearby features within a defined neighborhood. This localized modelling approach allows the relationships between variables to vary across different locations, resulting in a more detailed and contextualized understanding of the data's underlying patterns, particularly when there are significant geographic or spatial dynamics at work. The equations for both OLS and GWR are explained in (Appendix p 3).

At last, sensitivity analysis and uncertainty analysis were performed in this analysis using Sobol' sensitivity analysis[23] in Modified Reaction-Diffusion (MRD) model. This approach quantified the impact of input parameter uncertainties on the model's predictions, specifically evaluating how variations in the diffusion rate (α), reaction rate (β), and carrying capacity (K) influenced the forecasted infected population.

Role of the funding source

Results

Initially, this study tried to identify the relationship between COVID-19-positive cases and population density using regression analysis. Different polynomial and exponential orders to identify the relationship between the population density and COVID-19-positive cases were considered. No relationship was observed between these two factors as shown in Fig. 2.

Ordinary Least Squares (OLS) regression model was implemented against weekly Northeast COVID-19 data. The model included population as an explanatory variable for the number of COVID-19 infections (INFECTED_P). The results indicate that the model did not effectively explain the variability in infection rates at the initial months of COVID-19 spread. The model's explanatory power was low, the overall model was not
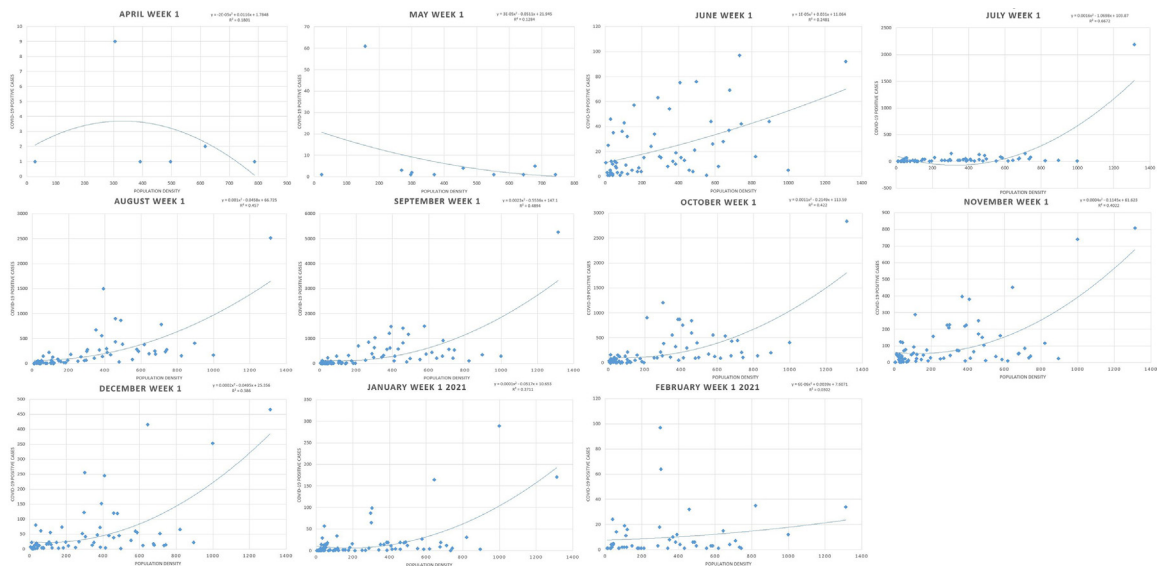

Fig. 2: This figure shows the monthly first week polynomial relationship between the population density and the COVID-19 positive cases. The rest of the weekly charts are available in Supplement 1 (ePlots).

statistically significant, as indicated by the Joint F-Statistic and the Joint Wald Statistic at the initial COVID-19 spread period. The results of OLS are provided in appendix (Appendix pp 234). The Geographically Weighted Regression (GWR) model was also implemented on the same northeast data. During the initial period of COVID-19 spread, the R^2 values were found to be very low in the early weeks of April 2020 and May 2020. The corresponding Adjusted R^2 values were −0.0266 and −0.0252, respectively. These negative Adjusted R^2 values suggest that the model performed poorly in explaining the spatial variability of COVID-19 cases, indicating a random disease pattern initially. GWR analysis results are available in appendix (Appendix pp 234). Since, initial regression analysis didn't yield any results, Spatial autocorrelation was considered to assess whether the distribution of COVID-19 cases was random, dispersed, or clustered across Northeastern India. This analysis involved calculating the Moran's Index value, z-score, and p-value to determine the spatial patterns in the data. The results indicated a random distribution of cases as shown in Fig. 3.

After regression analysis and spatial autocorrelation analysis were performed, the MRD Model was employed to identify the spread of COVID-19 cases in Northeastern India. It was observed that the MRD model proved to be highly effective in understanding the spatial distribution and temporal dynamics of infectious diseases. During the analysis, it was found that the diffusion rate (α) was approximately 0.4503. This parameter signifies the speed at which the virus spreads spatially within the region. A higher α indicates a more rapid geographic expansion of the infection, while a lower value suggests a slower spread, possibly influenced by factors such as geographical barriers or containment measures.

Furthermore, the reaction rate parameter (β) was determined to be around 0.0256. β plays a crucial role in modeling the growth of the infected population. A higher β signifies a faster increase in the number of cases, assuming no constraints, while a lower β value implies slower growth, possibly due to factors such as public health interventions.

The modified reaction-diffusion model was then implemented, allowing the creation of a 3D plot, as shown in Fig. 7. These figures illustrate the weekly increase in COVID-19 cases, with the x and y axes representing longitude and latitude, respectively, and the z-axis indicating the weeks of a given month. Within the figure, overlapping four-color point surfaces are observed, reflecting the spread of COVID-19 in a specific week and indicating which latitude and longitude values experienced the greatest COVID-19 spread. On the right side of each graph, a color bar labeled "infected density" illustrates the density value, with higher values indicating greater spread in specific latitude and longitude coordinates. Additional 3D graphs are available in Supplement 6 (e3D-Charts).

It was observed in weekly COVID-19 spread maps that the spread was not rapid during the initial weeks of April and May 2020, as depicted in Fig. 4. These patterns of uneven spread are consistent with the α and β values predicted by MRD-model. Quantile maps are created to visualize these spread patterns as shown in Fig. 4. Each color on the map represents a specific quantile range, making it easier to identify patterns and disparities in case distribution. The legends in these maps indicate the range of positive cases, with white indicating zero cases or very lower cases and other shades representing varying levels of infection. The spatial distribution of COVID-19 cases for every first week from April 2020 to February 2021 is shown in Figs. 5 and 6. In Fig. 5, normalized map has illustrated the concentration of positive COVID-19 cases relative to the total population across various districts in Northeastern India. The color legends represent the concentration of cases; darker the shade, higher is the concentration of cases. This weekly analysis provides valuable insights for health authorities to prioritize areas for resource deployment and to implement targeted containment measures. The distinct color-coded representation facilitates quick identification of hotspots, thereby aiding efficient decision-making in public health management. In Fig. 6, bivariate maps visually represent the distribution of COVID-19 positive cases in relation to the total population across various districts in the northeastern region of India. The x-axis of the color legend represents the number of COVID-19 positive cases (Positive), ranging from lighter shade of pink to darker shade for COVID-19 positive case occurrence intensity. The y-axis represents the total population (Total_P), ranging from low (light blue) to high (dark blue). This bivariate map is convenient for visualizing and analyzing the relationship between COVID-19 positive cases and the total population in northeastern India. It highlights critical areas requiring attention and provides insights that can inform public health policies and response strategies. Weekly normalized maps and bivariate maps from April 2020 to February 2021, are available in Supplement 3 (eNormalizedMaps) and Supplement 4 (eBivariateMaps).

To understand the MRD model analysis, both uncertainty and sensitivity parameters were estimated. The sensitivity analysis of the reaction-diffusion model provided both first-order and total-order sensitivity indices for the parameters α, β, and K. The first-order sensitivity indices were −1.5438 for α, −1.2607 for β, and −1.5386 for K. These values suggest that each parameter has a moderate inverse relationship with the model output, meaning that increases in these parameters lead to a decrease in the model output. The total-order sensitivity indices were 2.1026 for α, 1.8423 for β, and 2.0426 for K. These indices indicate that each parameter, along with
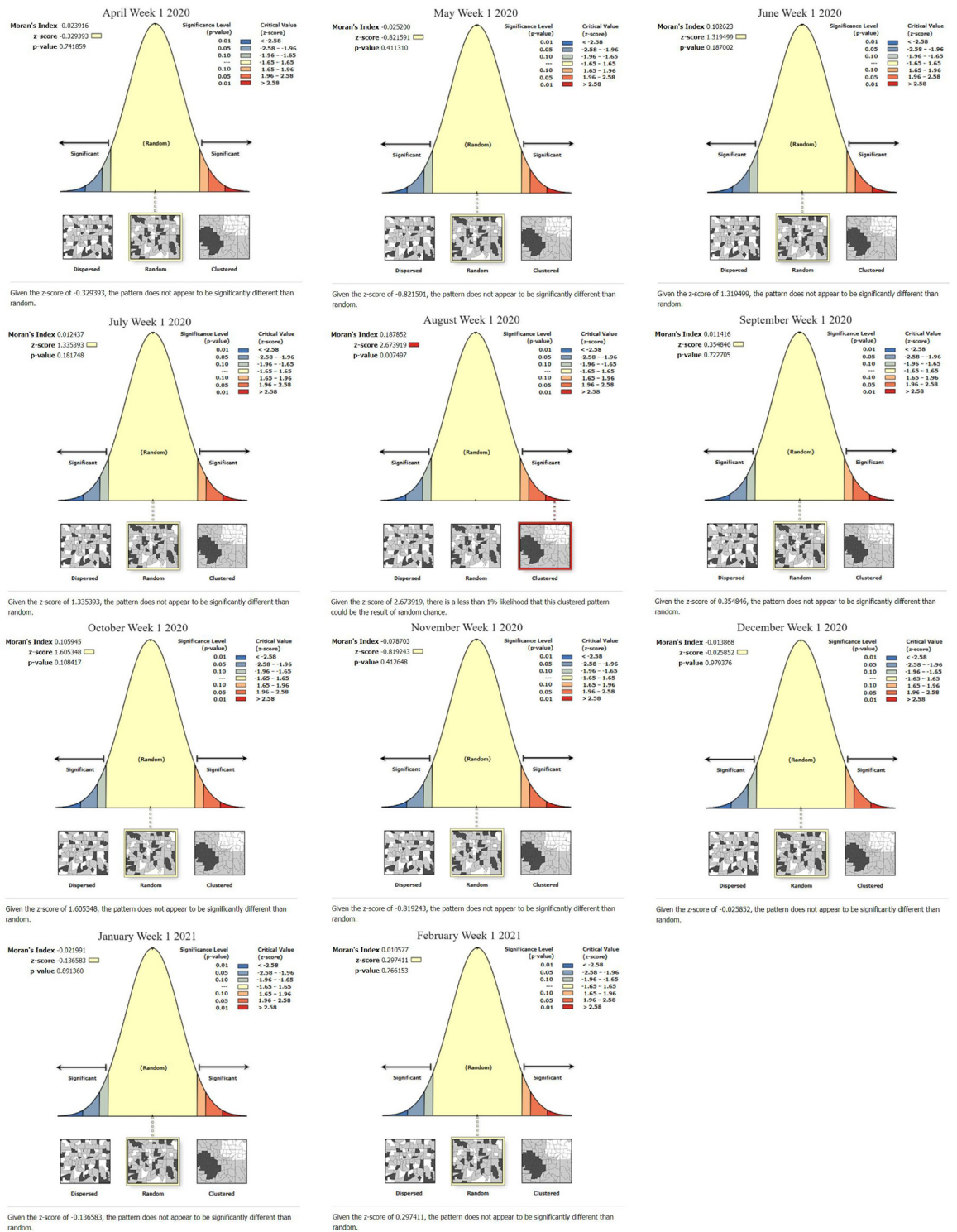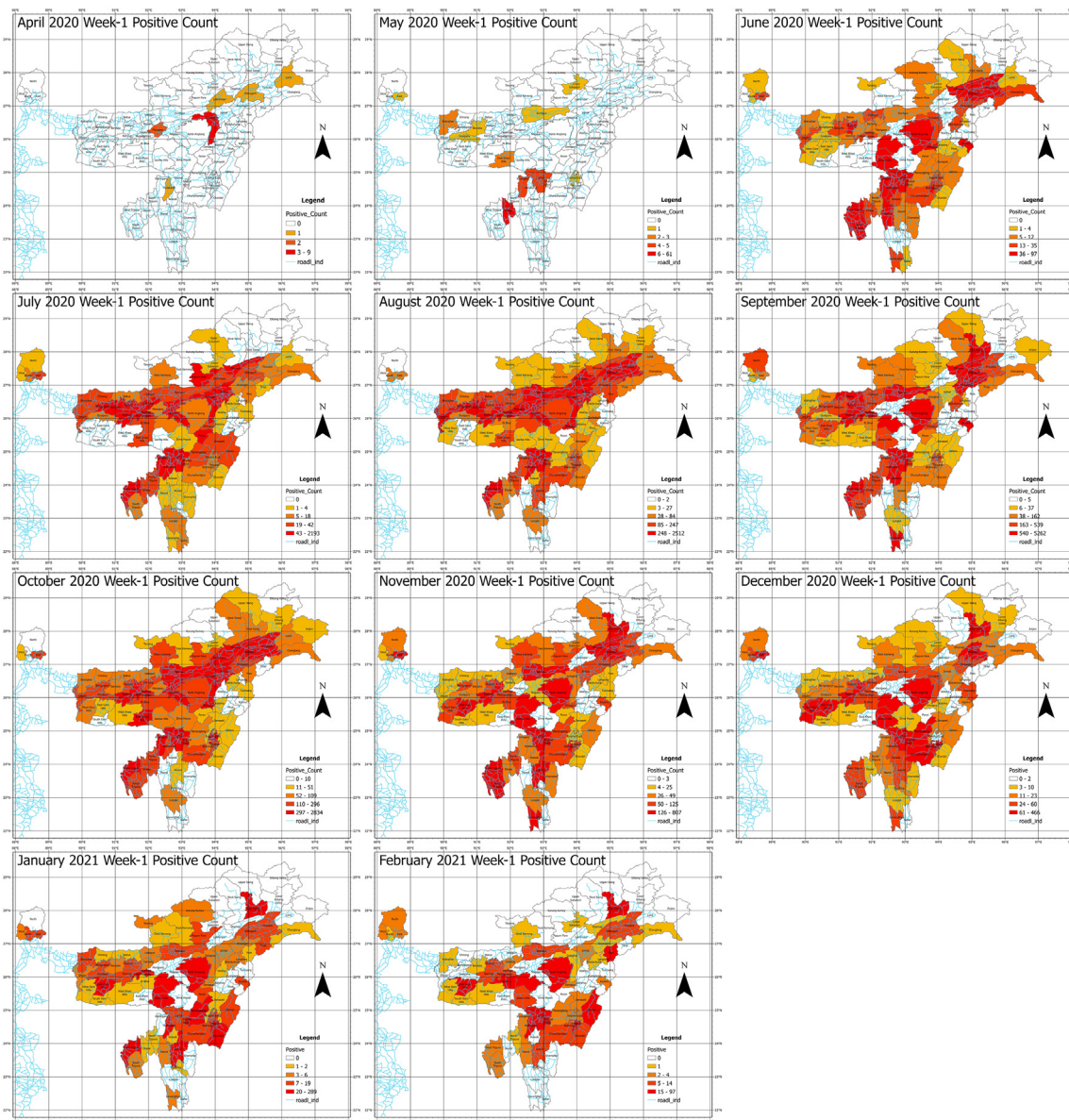
Fig. 3: This figure shows the monthly first week spatial autocorrelation relationship of COVID-19 positive cases. The rest of the weekly results are available in Supplement 2 (eSpatialPlots).

Fig. 4: **This figure contains the monthly first week quantile maps of COVID-19 positive cases. The rest of the weekly results are available in Supplement 3 (eQuantileMaps).**

its interactions with other parameters, significantly influences the model output variance.

Uncertainty analysis was also done using Sobol' sensitivity analysis to see how variations in input parameters affect the model's predictions. The results showed that the combined effects of the parameters and their interactions are significant, as seen in the high total-order indices. Even though the first-order indices were negative, the high total-order indices mean that the parameters and their interactions greatly influence the model. This means the model is very sensitive to

changes in the parameters, especially when considering their interactions. Understanding this is important for accurately adjusting parameter estimates to improve the model's accuracy and reliability.

Discussion

During the COVID-19 pandemic, the Government of India implemented a strict nationwide lockdown,[24] which included sealing state and district borders. The northeastern states of India, with their impervious
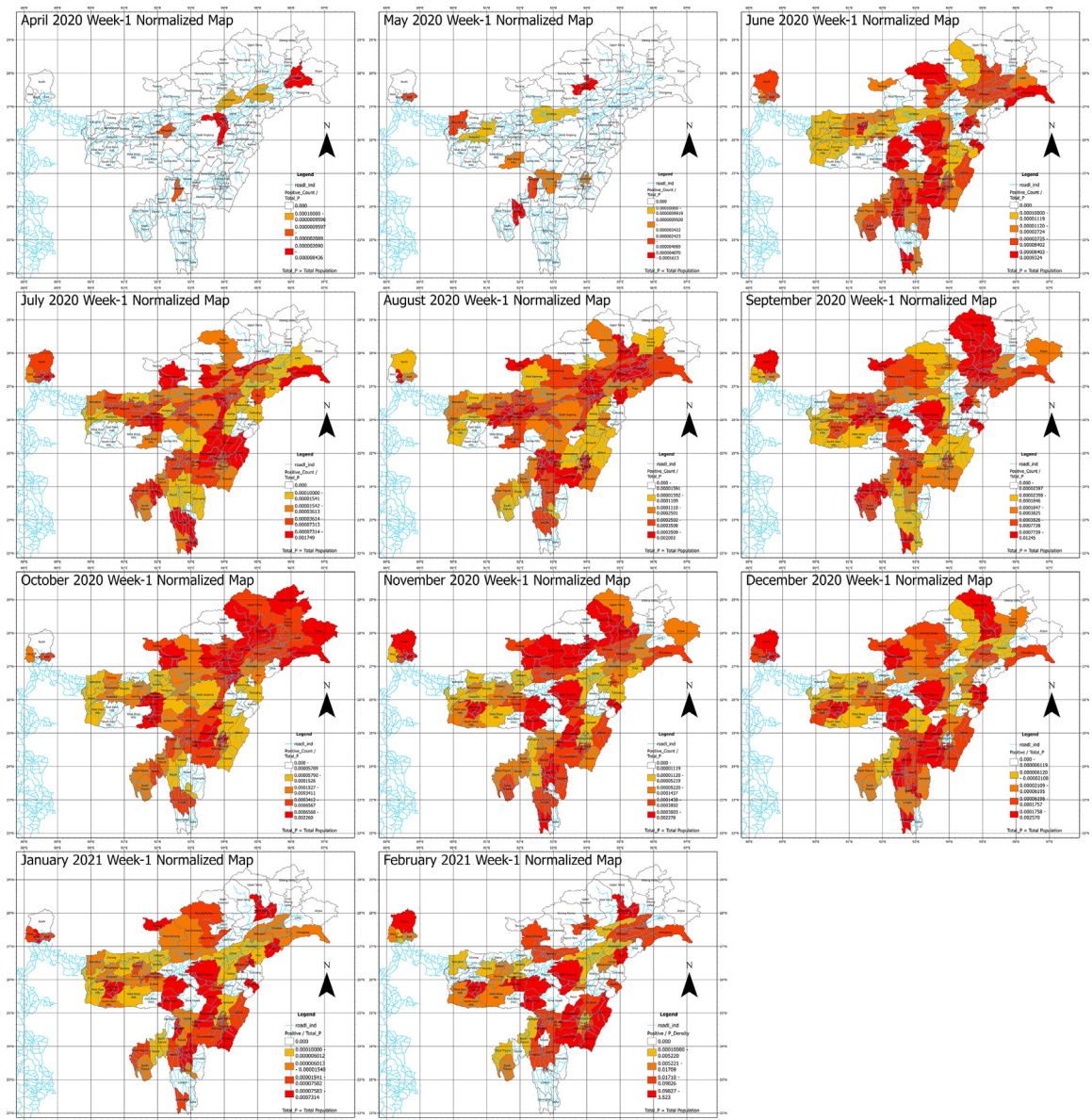
Fig. 5: **This figure contains the monthly first week normalized maps with district wise population of COVID-19 positive cases. The rest of the weekly results are available in Supplement 4 (eNormalizedMaps).**

international borders and limited connectivity to mainland India, provide a unique opportunity to study the spread of the virus in isolated regions during a lockdown. In India, many research articles were published related to the spread of COVID-19. All the articles used open-source data available on the Internet. In literatures, the disease spread analysis in India was done using various machine learning and AI models however no authors presented any mathematical basis for disease spread in a closed population which was a major research gap. Wang et al. proposed an analysis of the COVID-19 epidemic in India[25] using machine learning

and they have used COVID-19 data from several sources. In a recent study,[25] the authors used a Machine learning model with different combinations of input features, in which the Transformer is proven to be the most precise one. In another study,[26] recurrent neural networks such as long short-term memory (LSTM), bidirectional LSTM, and encoder-decoder LSTM models for multi-step (short-term) COVID-19 infection forecasting were used to capture the COVID-19 hotspots of the first wave and the second wave in India. Similar to the above study, performances of deep learning methods,[27] including the hybrid convolutional neural
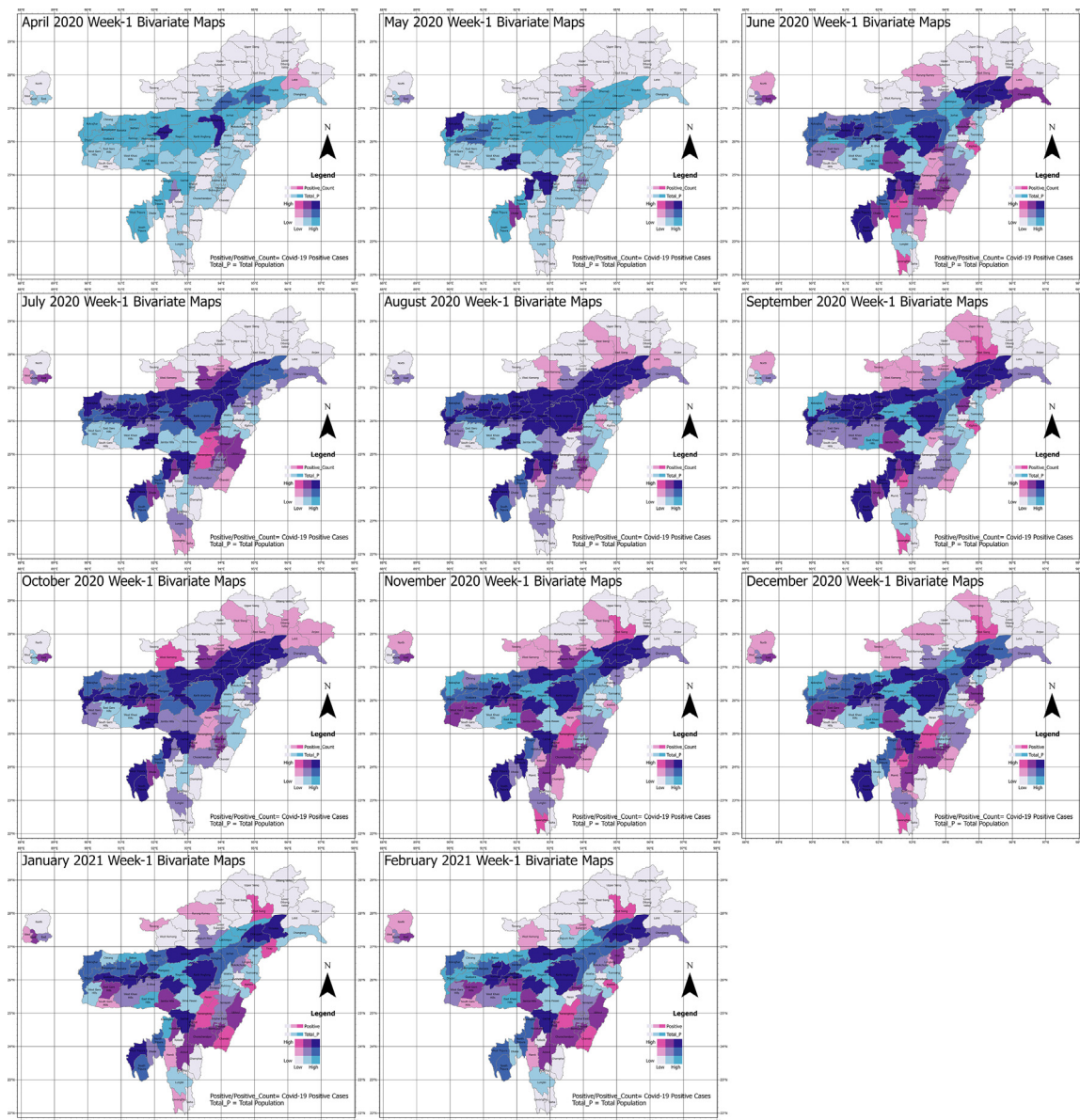
Fig. 6: This figure contains the monthly first week bivariate maps with total population and positive cases of northeast India. The rest of the weekly results are available in Supplement 5 (eBivariateMaps).

networks short-term memory (LSTM-CNN), the hybrid gated recurrent unit-convolutional neural networks (GAN-GRU), GAN, CNN, LSTM, and Restricted Boltzmann Machine (RBM), as well as baseline machine learning methods, namely logistic regression (LR) were used for COVID-19 transmission forecasting. Machine learning has provided valuable insights in both the quantified demographic and medical data, however, quantified demographic data like population density still lags proper research at large.

Few studies have used population density to understand the disease spread in various regions.

Ganasegeran et al. (2021)[28] conducted an ecological study in Malaysia, utilizing spatial analytics to create choropleth maps and hierarchical cluster analysis to identify potential clusters based on population density. They found a strong correlation between high population density and increased COVID-19 cases, especially in urban areas (r = 0.912).[28] Such methods can be relevant for studying similar patterns in Northeastern India. Ganasegeran et al. (2024) further examined the impact of vaccinations on pandemic indicators using geographic visualizations and regression models. The study confirmed the effectiveness of vaccinations in
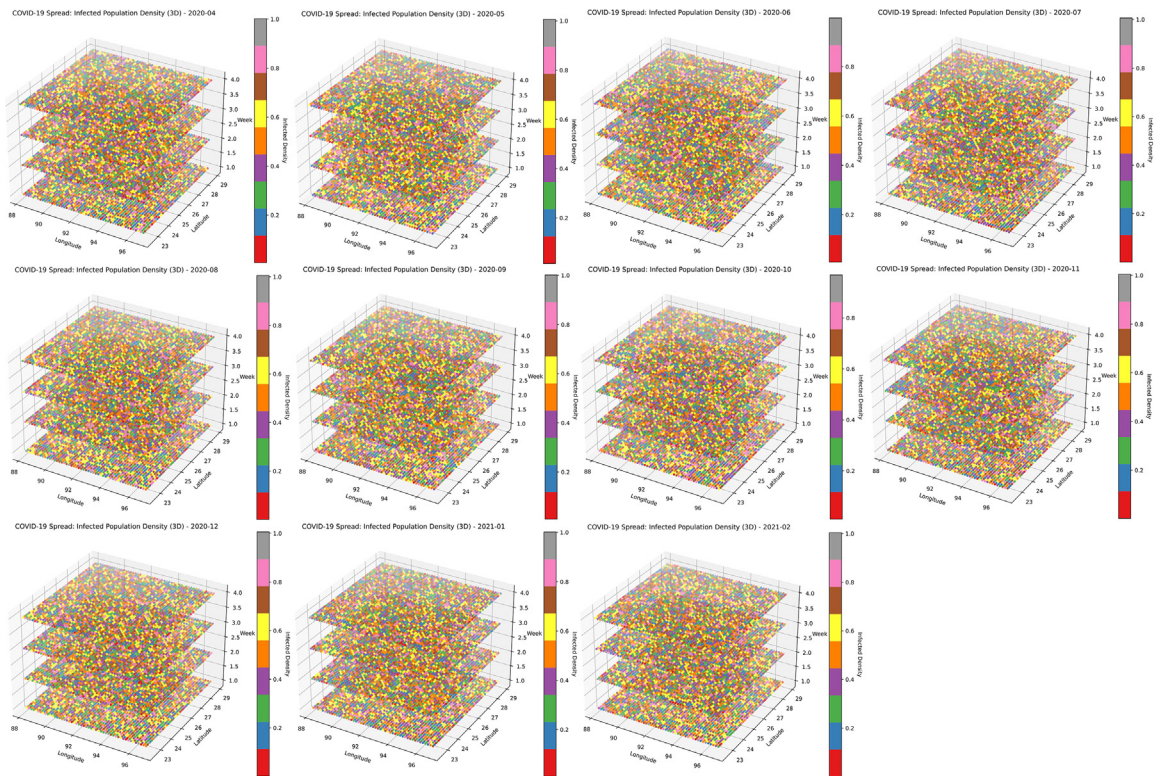
Fig. 7: In this figure it contains the MRD-Model 3D plot of weekly spread analysis based with latitude, longitude and time.

reducing infection rates and other indicators across different regions.[29] These techniques and the approach used in this study can be adapted to analyze regional variations in COVID-19 spread in India, considering its diverse population density. An important aspect while considering population density, is herd immunity. It is critical in understanding and managing the spread of infectious diseases, including COVID-19. Herd immunity occurs when a sufficient proportion of the population becomes immune to an infection, either through vaccination or previous infections, thereby reducing the likelihood of disease spread. Several models and algorithms have been developed to estimate the critical ratio of vaccination needed to achieve herd immunity.[30] Achieving herd immunity is crucial for controlling the spread of infectious diseases like COVID-19. The critical ratio of vaccination required to reach herd immunity can be estimated using the reproduction number (Rt), which reflects the average number of secondary infections caused by an infected individual at a given time. The formula $P_{crit} = 1 - \frac{1}{Rt}$, adjusted for vaccine efficacy, provides a robust estimator for the required vaccination coverage. This method is consistent with prior studies such as Ganasegeran et al. (2021),[30] which allows for dynamic assessment based on real-time data.

Meanwhile in this present study we have developed a modified reaction-diffusion mathematical model to predict the initial spread of the COVID-19 virus in isolated populations, using the Indian Council of Medical Research data utilizing quantified demographic data. This MRD model was chosen rather than traditional machine learning algorithms or other approaches discussed above due to its specific advantages in capturing the dynamics of disease transmission in isolated regions with limited connectivity. The analysis of the data revealed interesting patterns of disease spread that did not align with established grounded theories. Out of the total 7.1 million weekly data collected from April 2020 to February 2021, including information such as date, test result, population density, area, latitude, longitude, district, and state, the spread pattern was identified using the modified reaction-diffusion model and Geographic Information System (GIS).These data allowed us to understand the truth behind the spread of the COVID-19 virus in Northeastern states of India. The simulation approach has been overlooked in India; hence, this study developed a novel approach by integrating Geographic Information Science and a Modified Reaction-Diffusion System to simulate the initial disease pattern. The model predicts a higher rate of COVID-19 spread across different regions of Northeastern states of India and a lower number of COVID-19 cases in a particular region. Along with it, this study will also lay the foundation for the studies that try to understand the grounded theories

related to this pandemic or any future pandemic/epidemic in isolated regions. The simulation results based on the MRD model showed the spatial spread of COVID-19 over time and inferred how the infected population density evolved within the simulated region. This makes this study particularly important compared to other studies by those who employed Machine Learning or GIS-based approaches. The modified reaction-diffusion model effectively captured the dynamics of disease transmission in this unique setting.

It is important to note that this study was funded by the Indian Council of Medical Research (ICMR), which monitored serological testing in India and compiled the data in their servers. The utilization of population-based data, individualized by national identification cards, further strengthens the insights gained from this study. The implications of these findings are significant. Understanding COVID-19 transmission in isolated regions with limited external connectivity is crucial for developing effective public health strategies. The rapid initial dissemination followed by a slower growth trajectory observed in this study can inform theoretical frameworks for predicting the spread of the virus in isolated settings. Additionally, the successful application of geospatial analysis and mathematical modelling approaches demonstrates their utility in guiding public health strategies in similar populations worldwide. In conclusion, this study provides valuable insights to the spread of COVID-19 in isolated regions during lockdown. The analysis of the data using geospatial analysis and a modified reaction-diffusion model has enhanced our understanding of the transmission dynamics in these settings. The findings from this study have significant implications for public health policy and pandemic control strategies in India and other regions with similar demographic and geographic characteristics. The MRD model's ability to incorporate geographic factors and carrying capacity constraints makes it particularly suited for predicting disease spread in regions with limited connectivity and varying population densities. This allows for a more adaptive approach to public health strategies, for instance, areas identified as potential hotspots can be prioritized for vaccination drives, testing, and resource allocation, reducing the overall burden on healthcare systems. Also, the identification of uneven spread patterns, particularly during the initial weeks of the pandemic, underscores the importance of early detection and localized containment measures. Public health policies can thus be designed to implement rapid response teams in districts showing early signs of increased transmission, thereby preventing wider outbreaks. The predictive accuracy of the MRD model also supports the deployment of preemptive measures in regions predicted to experience future spikes in cases. This aids the policymakers to enhance their preparedness and response capabilities, ultimately mitigating the impact of COVID-19 and future pandemics.

This study has a few limitations such as inconsistencies in data reporting during certain periods might introduce biases. It also lacked detailed data on population movements within and between districts, which could impact the understanding of how human mobility influenced the spread of COVID-19. The accuracy of geographic coordinates varied, potentially leading to slight inaccuracies in spatial representations. The parameters estimated by the MRD model, such as diffusion rate, reaction rate, and carrying capacity, involve uncertainties. The uncertainty analysis using Sobol' sensitivity analysis revealed that variations and interactions in these parameters significantly influence the model's predictions. These parameters might not capture all complexities of disease spread, highlighting the importance of accurately calibrating them to improve the model's reliability.

The coronavirus (COVID-19) started infecting the Indian population in early March 2020 and it spread across the nation very quickly. The study conducted on the spread of COVID-19 in Northeastern India employs a Modified Reaction-Diffusion (MRD) model to understand the spatial and temporal dynamics of the disease within this geographically isolated region. Utilizing data provided by the Indian Council of Medical Research (ICMR) from April 2020 to February 2021, the study incorporates geospatial mapping and Geographic Information System (GIS) analysis to visualize the inter-district variability and the spread patterns of COVID-19. The MRD model revealed an initial rapid geographical spread with a high diffusion rate and lower reaction rate, indicating a slower growth trajectory of case numbers. Sensitivity and uncertainty analyses using Sobol' sensitivity method were conducted to quantify the impact of input parameter. The uncertainties on model predictions showed significant combined effects of the parameters on the model's output variance. Geospatial analyses, including quantile maps, normalized maps, bivariate mapping, and Empirical Bayesian Kriging, were employed. This analysis provided a detailed spatial representation of COVID-19 spread, highlighting areas with higher infection rates relative to the population size. The study's findings underscore the importance of MRD modelling and geospatial analysis in developing effective public health strategies and targeted interventions. Analysis like this can aid in managing disease spread in isolated regions with minimal external connectivity. Limitations include potential biases from data inconsistencies and lack of detailed population movement data, but overall, the insights gained are critical for improving pandemic response and policy decisions in similar settings. Hence, this study provides a detailed geospatial analysis of COVID-19 spread in closed, non-vaccinated districts of Northeast Indian states, highlighting the effectiveness of lockdowns in preventing intra-cluster transmission. These findings have significant implications for pandemic

preparedness and response strategies. Future research should focus on developing robust surveillance systems that integrate geospatial data and real-time Rt estimations for early outbreak detection. Along with it a focus can be made on optimizing region-specific vaccination strategies to ensure equitable distribution and evaluating the impact of non-pharmaceutical interventions across different regions to create adaptable frameworks. Fostering multidisciplinary approaches that include public health, data science, and social sciences, and translating scientific insights into actionable policies, will enhance global health security. Further, by aligning our findings with the global expert consensus on research priorities, this study contributes to the collective effort to manage current outbreaks and proactively respond to future pandemics.

References
1. Guarner J. Three emerging coronaviruses in two decades. *Am J Clin Pathol*. 2020;153:420–421.
2. Barber RM, Sorensen RJD, Pigott DM, et al. Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. *Lancet*. 2022;399:2351–2380.
3. Selvavinayagam TS, Somasundaram A, Selvam JM, et al. Author correction: contribution of infection and vaccination to population-level seroprevalence through two COVID waves in Tamil Nadu, India. *Sci Rep*. 2024;14:5141.
4. Andrews M, Areekal B, Rajesh K, et al. First confirmed case of COVID-19 infection in India: a case report. *Indian J Med Res*. 2020;151:490.
5. Soni P. Effects of COVID-19 lockdown phases in India: an atmospheric perspective. *Environ Dev Sustain*. 2021;23:12044–12055.
6. Wu Y, Kang L, Guo Z, Liu J, Liu M, Liang W. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains. *JAMA Netw Open*. 2022;5:e2228008.
7. Choutagunta A, Manish GP, Rajagopalan S. Battling COVID-19 with dysfunctional federalism: lessons from India. *South Econ J*. 2021;87:1267–1299.
8. Talabis DA S, Babierra AL, Buhat CA H, Lutero DS, Quindala KM, Rabajante JF. Local government responses for COVID-19 management in the Philippines. *BMC Public Health*. 2021;21:1711.
9. Jain N, Hung I-C, Kimura H, et al. The global response: how cities and provinces around the globe tackled Covid-19 outbreaks in 2021. *Lancet Reg Health Southeast Asia*. 2022;4:100031.
10. Kuniya T, Tokuda Y, Nakamura H, Moromizato T, Shibuya K. Estimating COVID-19 herd immunity in Japan at the beginning of the seventh wave: transitioning from a pandemic to an endemic. *J Gen Fam Med*. 2022;23:295–298.
11. Goh FT, Chew YZ, Tam CC, Yung CF, Clapham H. A country-specific model of COVID-19 vaccination coverage needed for herd immunity in adult only or population wide vaccination programme. *Epidemics*. 2022;39:100581.
12. https://pib.gov.in/PressReleasePage.aspx?PRID=1625332. Accessed February 23, 2024. Lock-down measures in all the NE States.
13. Sengupta P, Ganguli B, SenRoy S, Chatterjee A. An analysis of COVID-19 clusters in India. *BMC Public Health*. 2021;21:631.
14. Stevens GA, Alkema L, Black RE, et al. Guidelines for accurate and transparent health estimates reporting: the GATHER statement. *Lancet*. 2016;388:e19–e23.
15. Kuter BJ, Offit PA, Poland GA. The development of COVID-19 vaccines in the United States: why and how so fast? *Vaccine*. 2021;39:2491–2495.
16. Tsoularis A, Wallace J. Analysis of logistic growth models. *Math Biosci*. 2002;179:21–55.
17. Kolmogorov A, Petrovskii I, Piskunov N. *Study of a Diffusion Equation That Is Related to the Growth of a Quality of Matter and Its Application to a Biological Problem*. Moscow University Mathematics Bulletin; 1937;1:1–26.
18. Murray JD. *Mathematical biology*. New York, NY: Springer New York; 2002. https://doi.org/10.1007/b98868.
19. Andrews B. Techniques for spatial analysis and visualization of benthic mapping data: final report. http://hdl.handle.net/1834/20024; 2003.
20. Getis A. Spatial autocorrelation. In: *Handbook of applied spatial analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:255–278.
21. Krivoruchko K. Empirical bayesian kriging. *ArcUser Fall*. 2012;6:1145.
22. Scott LM, Janikas MV. Spatial statistics in ArcGIS. In: *Handbook of applied spatial analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:27–41.
23. Nossent J, Elsen P, Bauwens W. Sobol' sensitivity analysis of a complex environmental model. *Environ Model Software*. 2011;26:1515–1525.
24. Saha J, Barman B, Chouhan P. Lockdown for COVID-19 and its impact on community mobility in India: an analysis of the COVID-19 community mobility reports, 2020. *Child Youth Serv Rev*. 2020;116:105160.
25. Wang Q. COVID-19 epidemic analysis in India with multi-source state-level datasets. *Biomed Res Int*. 2022;2022:2601149.
26. Chandra R, Jain A, Singh Chauhan D. Deep learning via LSTM models for COVID-19 infection forecasting in India. *PLoS One*. 2022;17:e0262708.
27. Dairi A, Harrou F, Zeroual A, Hittawe MM, Sun Y. Comparative study of machine learning methods for COVID-19 transmission forecasting. *J Biomed Inform*. 2021;118:103791.
28. Ganasegeran K, Jamil MFA, Ch'ng ASH, Looi I, Peariasamy KM. Influence of population density for COVID-19 spread in Malaysia: an ecological study. *Int J Environ Res Public Health*. 2021;18:9866.
29. Ganasegeran K, Jamil MFA, Ch'ng ASH, Looi I, Rusli N, Peariasamy KM. COVID-19 pandemic indicators and variation with vaccinations in Malaysia: a regional-based geo-visualization and geo-ecological regression study. *GeoJournal*. 2024;89:114.
30. Ganasegeran K, Ch'ng ASH, Looi I. What is the estimated COVID-19 reproduction number and the proportion of the population that needs to be immunized to achieve herd immunity in Malaysia? A mathematical epidemiology synthesis. *COVID*. 2021;1:13–19.