Spatial Autocorrelation of Crime in India

Mayuri Salunke, Shambhavi Goenka, Wong Kelly

ABSTRACT- This paper investigates the spatial patterns of crime in India and explores the degree of spatial autocorrelation present in the data. The study uses exploratory data analysis (EDA) techniques, including visualization methods to examine and identify any significant patterns or trends. Local and global autocorrelation tests, including Moran's I and Geary's C have been employed to assess the degree of spatial dependence in the data. In addition to local indicators of spatial autocorrelation (LISA) cluster maps, Getis and Ord's G statistics have been used to generate hotspots to detect spatial patterns. The results show that overall there is low spatial autocorrelation but certain regions do exhibit higher likelihood of specific crimes.

Problem & Motivation

India has a high crime rate of 445.9 per 100,000 individuals [1] spanning across different types from violent to property crime. Based on the crime-sheeting rate, Gujarat is at 98.3% and whereas Arunachal Pradesh is at 53.6% [2]. The two states are under the same political party and Gujarat is above the poverty line while Arunachal Pradesh is below it [3] which led to the question of spatial autocorrelation of crime.

In order to visualize the crime patterns in India, the research focuses on conducting a geospatial analysis, specifically in Global and Local Measures of Spatial Autocorrelation. This involves identifying the crime rates in the various districts of India as well as the distribution of crime types among districts in India.

Knowledge of the spatial patterns of crime would assist policymakers in understanding the type and severity of crime rates in various Indian districts, they would be better equipped to conduct investigations and enact more stringent laws in the districts that have been identified as having high crime rates. Additionally, it intends to increase awareness of safety and crime issues in India so that visitors can make more informed decisions about where they want to vacation after learning about India's crime patterns.

I. Data Sources

Geospatial Data: Kaggle – contains the SHP file of India 2014 District [4].

Aspatial Data:

1. Districtwise IPC Crimes 2021 [5]: The dataset consists of crimes punishable under the Indian Penal Code (IPC) on a district level including the 28 states and 8 union territories. There are 142 such crimes that can be further combined.

2. Districtwise SLL Crimes 2021 [5]: Similarly, this dataset consists of crimes punishable under Special and Local Laws (SLL) on a district level including the 28 states and union territories. There are 92 such crimes that can be further combined.

The data contains multipolygons of districts associated with the number of cases of crime. The datasets initially consisted of high cardinality so most of the information has been made more coarse-grained by merging into 27 types of crime. Furthermore, due to several changes in naming conventions and boundaries, the number of districts studied has been reduced to 686 to match the 2014 geospatial and 2021 aspatial data.

II. Analysis Method

The analysis methods in this study are divided into the following stages:

1. **Exploratory Data Analysis (EDA):** The first stage of the analysis involves conducting EDA using statistics and graphical techniques to gain insights into the distribution of different types of crimes across districts in India. This analysis uses choropleth maps to visualize the geographic distribution of crime rates for different types of crimes.

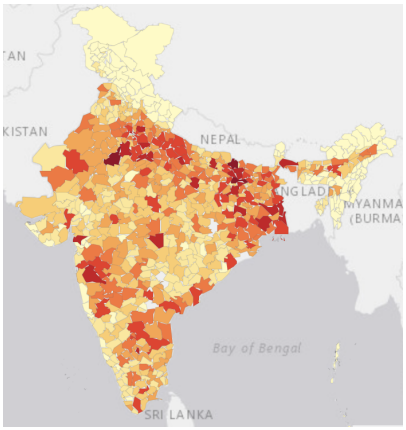

Figure 1: Choropleth mapping using Jenks style of murders

From the figure above, the analysis shows that the distribution of different types of crimes varies widely across districts in India. For example, the murders in India have a higher rate near the northeastern region of Bihar and Uttar Pradesh but are generally spread across the nation. The central and eastern regions have a lower rate of murders. Although the Jammu and Kashmir region shows a low rate, the data for that region is a bit unreliable due to high terrorist activity there.

Another example is that of attempted murders below.

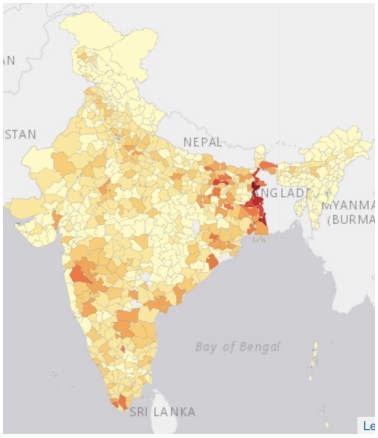

Figure 2: Choropleth mapping using Jenks style of attempted murders

While murders and attempted murders are expected to follow a similar trend, attempted murders seem to be highly clustered in the state of West Bengal, especially the districts bordering Bangladesh.

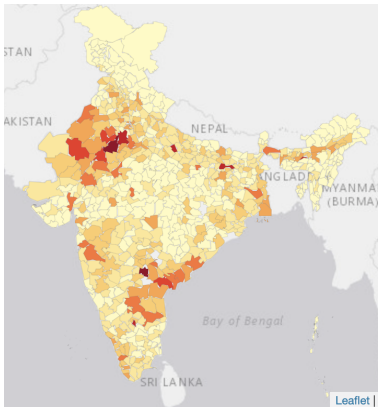

Figure 3: Choropleth mapping using Jenks style of document fraud

Lastly on document fraud which consists of all forms of counterfeiting, forgery, cheating and fraud, these crimes are clustered around the north western region of Delhi, that is the national capital and the southeastern state of Andhra Pradesh. As these are more developed regions, this visualization links that development to higher white collar crimes.

2. **Global Spatial Autocorrelation:** The next stage of the analysis involves conducting global spatial autocorrelation tests to determine the overall spatial patterns of crime in India. Two widely used measures of global spatial autocorrelation are Moran's I test and Geary's C test.

   The Moran's I test measures the degree of similarity between neighboring districts with respect to the incidence of a particular type of crime. The test produces a score between -1 and +1, where positive values indicate spatial clustering (i.e., similar values tend to be found in neighboring districts), negative values indicate spatial dispersion (i.e., dissimilar values tend to be found in neighboring districts), and a score of 0 indicates spatial randomness (i.e., no spatial pattern).

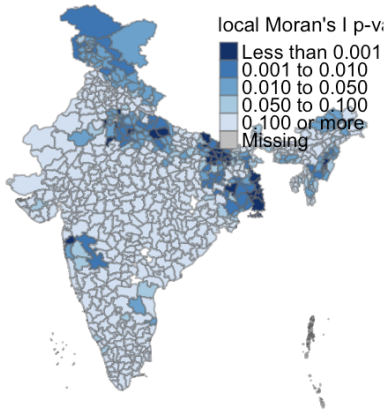

Figure 4: P-values of local Moran's I values of murders

The figure above shows the p-values of local Moran's I test. Most of the values are insignificant but the districts in Bihar, Uttar Pradesh and northeastern India have highly significant values indicating the likelihood of extremes of murder and lack thereof.

The significance of the Moran's I score is assessed using a permutation test, which involves simulating random permutations of the data and calculating the Moran's I score for each permutation. The observed Moran's I score is compared to the distribution of the simulated scores to determine whether the observed score is significantly different from what would be expected under the null hypothesis of no spatial autocorrelation.

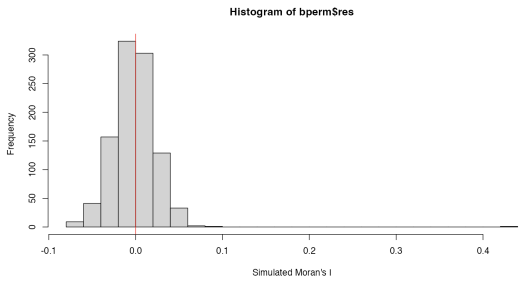

Figure 5: Simulation of permutations of global Moran's I of murders

In this example of the murders, as it is fairly distributed across most regions, the clusterings are mostly randomized as shown by the distribution around 0.

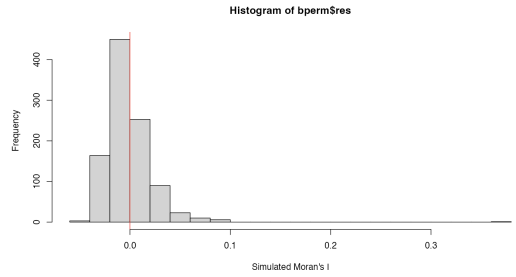

Figure 6: Simulation of permutations of global Moran's I of attempted murders

Expectedly, based on the earlier EDA, attempted murders although similarly distributed had a higher clustering in West Bengal. This is observed here based on the right skewed distribution.

The Geary's C test is another measure of global spatial autocorrelation that is based on differences in the incidence of a particular type of crime between neighboring districts. The test produces a score between 0 and 2, where values close to 0 indicate spatial clustering and values close to 2 indicate spatial dispersion. The significance of the Geary's C score is also assessed using a permutation test, which involves simulating random permutations of the data and calculating the Geary's C score for each permutation. The observed Geary's C score is compared to the distribution of the simulated scores to determine whether the observed score is significantly different from what would be expected under the null hypothesis of no spatial autocorrelation.

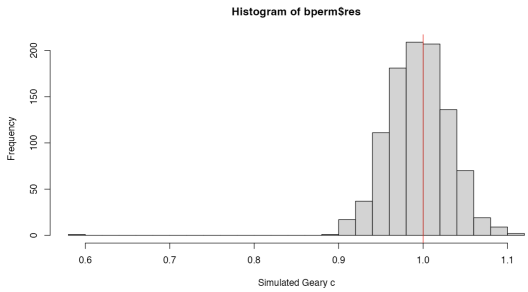

Figure 7: Simulation of permutations of global Geary's C of murders

With similar results as the Moran's I test, the Geary's C test shows randomness in the spatial autocorrelation of murders with its values centered around 1.

The spatial lag values are also plotted against the standardized residuals to examine the relationship between the incidence of a particular type of crime in a district and the incidence of the same type of crime in neighboring districts.

3. **Cluster & Outlier Analysis with Local Indicators of Spatial Association (LISA) maps:** In addition to the global spatial autocorrelation tests, it is also important to identify specific areas of high and low crime rates in India. This can be done using local spatial autocorrelation measures, such as the Local Moran's I statistic, which identifies clusters of districts with similar crime rates and outliers with significantly different crime rates compared to their neighboring districts.

The Local Moran's I statistic measures the spatial autocorrelation of a particular type of crime for each district, taking into account the incidence of the same type of crime in neighboring districts. The statistic produces a score between -1 and +1, where positive values indicate clusters of high-high (HH) or low-low (LL) values, negative values indicate clusters of high-low (HL) or low-high (LH) values, and a score of 0 indicates no significant clustering or outliers. The significance of the Local

Moran's I score is assessed using a permutation test, similar to the global spatial autocorrelation tests.

The significant outliers identified in the Local Moran's I analysis can be linked to the quadrants of the LISA cluster map. The HH and LL clusters are represented in the upper right and lower left quadrants, respectively, while the HL and LH outliers are represented in the upper left and lower right quadrants, respectively. The quadrants can be used to further interpret the results of the Local Moran's I analysis and identify the specific areas of high and low crime rates in India.

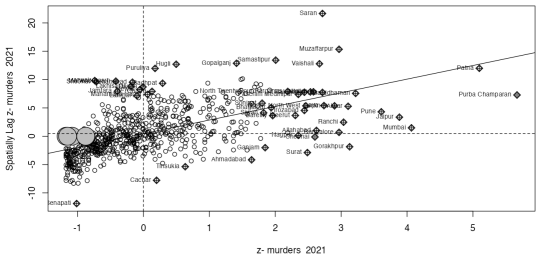

Figure 8: LISA cluster map of murder rates in districts

Based on the LISA cluster map of the Moran's scatterplot (after standardization), the highest districts with the highest murders (top right quadrant) - Patna, Purba Champaran and Saran are all in Bihar.

Overall, the cluster and outlier analysis using Local Moran's I provides important insights into the specific areas of high and low crime rates in India and can inform the development of more targeted crime prevention and control strategies. The analysis highlights the need for a more nuanced approach to crime prevention that takes into account the unique spatial patterns of different types of crimes in India.

4. **Hot & Cold Spot Analysis:** In addition to the cluster and outlier analysis, it is important to identify areas of significant spatial clustering or dispersion of crime rates in India. This can be done using the Getis-Ord's G statistic, which measures the degree of spatial clustering or dispersion of a particular type of crime across districts.

   The Getis-Ord's G statistic produces a score for each district, representing the degree of spatial clustering or dispersion of a particular type of crime relative to its neighboring districts. Positive scores indicate hotspots, or areas with high crime rates and significant spatial clustering, while negative scores indicate coldspots, or areas with low crime rates and significant spatial clustering. The significance of the Getis-Ord's G score is assessed using a permutation test, similar to the global and local spatial autocorrelation tests.

   The results of the Getis-Ord's G analysis are presented visually in the form of a hotspot and coldspot map, which shows the spatial distribution of areas with significant spatial clustering or dispersion of crime rates for each type of crime. The map highlights areas with significant hotspots or coldspots as well as areas with no significant spatial clustering or dispersion of crime rates.

   The analysis uses fixed and adaptive weights for the distance matrix computation. The example below is of the hotspots and coldspots for murder in India based on the adaptive bandwidth method.

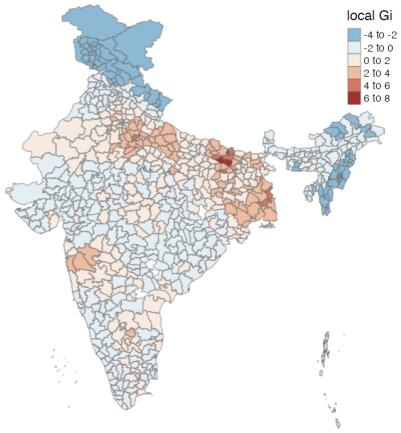

Figure 9: Local Gi values of hotspots and coldspots of murders

The visualization gives a better description of what was earlier observed in the EDA and the local Moran's I tests. The hotspots (indicated in red) for murder are primarily in Bihar and extend into Uttar Pradesh. The coldspots (indicated in blue) i.e., lower likelihood of murders are observed in northeastern states.

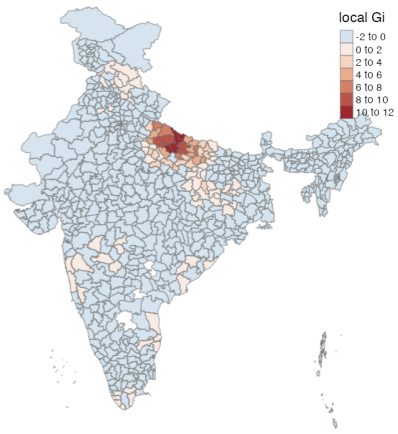

Figure 10: Local Gi values of hotspots and coldspots of crimes against scheduled tribes and scheduled castes

In the figure above, it is observed that the crimes against the scheduled tribes and scheduled castes are primarily in Uttar Pradesh. This is because the state has the highest demographic of scheduled tribes and scheduled castes.

III. Design Framework

The UI design of CrimeWatch application was inspired by the work of Greg Pilgrimname [5] that conducted a study on looking at college swimming programs to help incoming student athletes in selecting schools and programs that would fit them athletically, academically, locationally, and financially. The application provides interactive maps and charts to display programs sorted by swimming times/events, etc. It also allows users to filter the data to narrow down to the areas of interest. CrimeWatch also follows a similar design.

At the top of the application is the navigation bar to show the different tabs: Overview, EDA, Global Spatial Autocorrelation, Cluster and Outlier Analysis & LISA Cluster Maps, and Hot and Cold Spot Analysis. The overview tab offers users a walk-in comprehension of the goal of this visual application by providing a macro picture of the study area, including motivation and problem statement. Below the overview section, users can find the raw datasets for the Indian Criminal Code (IPC) and the Special and Local Laws (SLL), allowing them to always resort to this information in order to make more educated judgments based on accurate and recent data.

The main features of the analytical tools used to in each tab of analysis method are:

1. EDA
    a. Filter by crime law - dropdown
    b. Filter by crime type - dropdown
    c. Filter by classification method - dropdown
    d. Filter by color scheme – dropdown
    e. Filter by number of classes – Slider
2. Global Spatial Autocorrelation
    a. Filter by crime law – dropdown
    b. Filter by crime type – dropdown
    c. Filter by spatial autocorrelation test - dropdown
3. Cluster & Outlier Analysis & LISA Cluster Maps
    a. Filter by crime law – dropdown
    b. Filter by crime type – dropdown
    c. Filter by visualizing map – dropdown
    d. Filter by type of map – dropdown
4. Hot & Cold Spot Analysis
    a. Filter by crime law – dropdown
    b. Filter by crime type – dropdown
    c. Filter by distance weight matrix - dropdown

In our application, dropdown buttons were frequently utilized as a UI element since they provide users with lists of alternatives that are simple to read at a glance. Dropdown buttons also save screen real estate because they occupy so little of it, leaving more room for visualization.

An interactive slider is used to filter the number of classes in EDA. It is especially helpful because it is intuitive and offers users a tactile and visual approach to input and modify value. When altering settings, it additionally offers a high level of precision. Customers can change in modest, gradual steps to get the precise value they want.

A combination of these filter methods along with the different maps and statistical plots presented in our application, we hope to provide users with a user-friendly and comprehensive place to search and visualize the information they need.

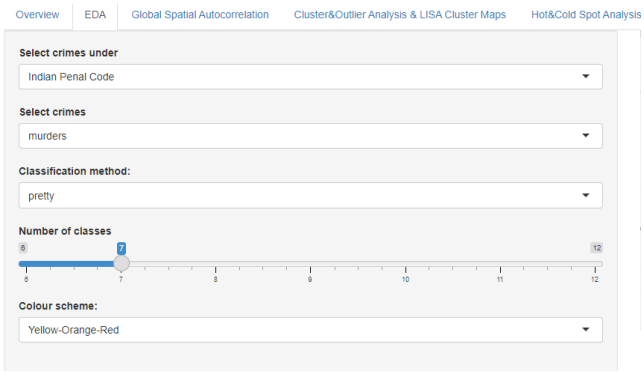

Figure 11 - Navigation and Filters

IV. Reviews and Critics

Prior to beginning the project, we also conducted some web research to study previous sample works and gather ideas for what we can accomplish as well as how we can improve on their work to make it better for users. We picked out two past geospatial analysis work and here are the reviews and critics:

Crime Mapping and Hot Spot Analysis using Geospatial Techniques: A case of Ajmer City

By Monika Kannan, Kasina V.Rao, Mayur Raj and P. Alok Verma [6]

The objective of the project is to identify crime prone zones through crime mapping with probability of occurrence based on the past incidence of various crime locations

Methodology used are:

1. Euclidean distances of Crimes from Police station and Highways
2. Cluster Analysis
3. Hotspot Analysis
4. Interpolation (co-Kriging): To predict the susceptible crime regions based on the correlation with other non spatial attributes

Learning points:

1. The use of co-Kriging total crime with socioeconomic factors as an estimation method
2. The use of weighted overlay analysis of selected variables to predict crime prone areas

Areas for improvement:

1. Local Gi* statistics could have been used to better filter the data based on crime levels
2. Euclidean distance mapping was unnecessary. Alternates like scatter plots with regression lines could have been used to visualise the distance better.
3. The correlation coefficients for co-Krining were really high (0.9+) and could have caused high collinearity.

Analysis of Criminal Spatial Events in GIS for predicting hotspots

By Abbas F. Mohammed and Wadhah R. Baiee [7]

To recognize the hotspots for crime data like Shooting, Homicide and Assault by threat in Baltimore, Maryland

Methodology used are:

1. Getis-Ord Gi* statistics
2. Hot Spot Analysis

Learning points:

1. The use of a small dataset and region was beneficial to get a granular understanding of the crimes rates over time

Areas for improvement:

1. There could have been a final aggregation of the crimes as a final visualization
2. Additional data on the strength of each police district could have improved the predictive value of crime rates

V. Conclusion

In conclusion, we hope that the statistical techniques and tools employed in this project can be used to examine India's crime trends better and provide variable insights to a range of interested parties.

The application has a lot of potential to be extended and enhanced further. For real-time analysis and exploration, the stored criminal data can be replaced by a stream of real-time data in the future. The data approach is adaptable to crime data from various nations as well.

REFERENCES

[1] National Crime Records Bureau (Ministry of Home Affairs) Government of India, *Crime in India 2021*, https://ncrb.gov.in/sites/default/files/CII-2021/CII%202021%20SNAPSHOTS%20STATES.pdf.

[2] National Crime Records Bureau (Ministry of Home Affairs) Government of India, *Crime in India 2021*, https://ncrb.gov.in/sites/default/files/CII-2021/CII_2021Volume%202.pdf.

[3] The Global Statistics. (February 20, 2021). Leading states and union territories with the largest share of people living below the poverty line in India in 2021 [Graph]. In Statista. Retrieved April 16, 2023, from https://www-statista-com.libproxy.smu.edu.sg/statistics/1269976/india-population-living-below-national-poverty-line-by-state/

[4] DEVAKUMAR, K. P. "India 2020 District Level Shape Files." https://www.kaggle.com/datasets/imdevskp/india-district-wise-shape-files. Kaggle.

[5] Pilgrimname, Greg. "NCAA Swimming Team Finder for Incoming College Athletes." Shiny, https://shiny.rstudio.com/gallery/ncaa-swim-team-finder.html.

[6] Sharma, Ravi & Palria, Sarvesh & Bhalla, Parul. (2016). Crime Mapping & Analysis of Ajmer City-A GIS Approach. Journal of Geomatics. 10. 96-101.

[7] F. Mohammed, Abbas, and Wadhah R. Baiee. "Analysis of Criminal Spatial Events in GIS for Predicting Hotspots." *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, 2020, p. 032071., https://doi.org/10.1088/1757-899x/928/3/032071.