Spatial Analysis of Crime in Uttar Pradesh and Identification of crime "hotspots"

Rishav Raj Agrawal (12577)

Abstract

This study aims to conduct a spatial analysis of cross-sectional data using economic and socio-demographic variables to investigate the impact of regional proximity on crime rate thus identifying crime hotspots.

1 Motivation

Uttar Pradesh (UP), the most populous state in India has been historically famous for its high crime rate and lawlessness. In a 2012 survey, UP has been termed as the one of the **"worst states"** in India in terms of law and order by the National Crime Records Bureau (NCRB). UP is also notorious for its illegal arms trafficking and several illegal arms factories exist. I aim to find the determinants of crime in UP and explore whether spatial patterns exist in the distribution of crime over the state and thus highlight major crime *"hotspots"* in the process.

2 Introduction

Crime opportunities are neither uniformly nor randomly organized in space and time [5]. As a result, we can find spatial patterns and strive for a better theoretical understanding of the role of geography, as well as tailor practical crime prevention solutions for specific places. I propose an analysis of crime at city level, to capture important inter-regional differences using *Explanatory Spatial Data Analysis* (ESDA). ESDA allows us to detect some important geographical dimensions and to distinguish crucial macro- and micro-territorial aspects of offenses.

This study will use a *Spatial Model* for cross-sectional data using economic and socio-demographic variables to investigate the determinants of crime in UP cities for 2011 and its *"neighbouring"* effects, measured in terms of geographical and relational proximity.

3 Review of Literature

The earliest studies that explicitly explored the role of geography in the distribution of crime noted various spatial relationships [5]. Both Guerry (1833) and Quetelet (1842) examined nationwide statistics for France, the latter identifying that higher property crime rates were reported in more affluent locations, and that seasonality had a role to play in crime occurrence. British government studies followed, but data were only collected for large administrative units, and local crime data at the neighborhood (or smaller) level were not available. With more data being available, econometrists realized socio-demographic characteristics from one area can influence the volume of crime in another area. The spatial analysis of crime has demonstrated that the location of illegal activity can supply relevant insights about the exploration of crime dynamics [8].

In Anselin [8], spatial econometrics is defined as "the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models." These spatial effects (spatial autocorrelation and spatial heterogeneity) are captured using a weight matrix [9].

More recently, Cracolici [1] found that the empirical results obtained by using different spatial weights matrices highlighted that socioeconomic variables have a relevant impact on crime activities in Italy. Similar studies have been done by Delbecq [2] who analysed crime for Chicago and Pavlo, [3] for Ukraine. A study by Ahamed [7] has been done on the spatial patterns of crime in India but the results were not very concrete.

Delbecq [2] also talks about theory of social disorganization which states that "in urban areas, delinquency is not randomly distributed in space but tends to be concentrated in poor and socially excluded places." i.e. individuals are subject to neighborhood effects and factors such as income, schooling, etc play an important role in determining the crime rates in a place. Similarly Cracolici [1] mentions Becker's crime economic model (CEM) (1968) the illegal behaviour of individuals could be explained by means of the theory of rational behaviour under uncertainty.

Thus, the approach I follow here is inspired by Cracolici[1]. The study is unique as such a profound approach has never been undertaken in the Indian context especially at the state level.

4 Hypothesis

- H_0: Crime shows no spatial autocorrelation and is not affected by the GDP, Population Density, Unemployment rate and Literacy and Health Index.
- H_A: At least one of the above mentioned factors has a substantial effect on the crime rate in UP.

5 Data Sources

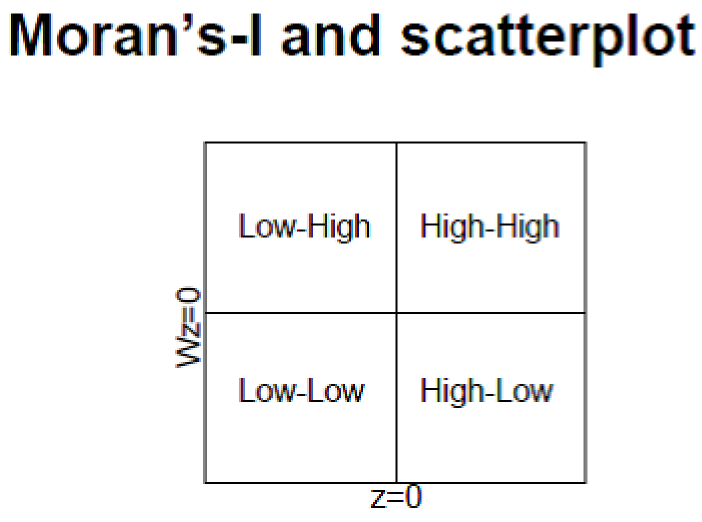
- World Bank Data on Uttar Pradesh
- National Crime Records Bureau Data

6 Model and Methodology

6.1 Spatial Autocorrelation

Spatial autocorrelation can be loosely defined as the coincidence of value similarity with locational similarity. In other words, high or low values for a random variable tend to cluster in space (positive spatial autocorrelation) or locations tend to be surrounded by neighbors with very dissimilar values (negative spatial autocorrelation). Following tests are used to test for Spatial Correlation:

- **Moran's I test:** The most common test for the existence of spatial autocorrelation is due to Patrick Moran, and is usually referred to as Moran's-I. It captures the "global" spatial autocorrelation, i.e. if provinces with high crime rate are clustered nearby or not, and ranges between -1 and $+1$.
- **Moran Plot:**



Category	Autocorrelation	Interpretation
high-high	positive	"I'm high and my neighbors are high."
high-low	negative	"I'm a high outlier among low neighbors."
low-low	positive	"I'm low and my neighbors are low."
low-high	negative	"I'm a low outlier among high."

- **Geary's C:** This identifies the presence of spatial autocorrelation, and is used to describe differences in small neighbourhoods, if its value is less than 1 there is a positive spatial autocorrelation, if higher than 1 there is negativepatial autocorrelation. Geary's statistic is more sensitive to local autocorrelation.

6.2 LM Tests

If residuals are spatial autocorrelated (Moran's I), then use the Langrange Multiplier diagnostic to determine appropriate model

- **Regression residuals (LM-Error)**

  Mis-match of process and spatial units systematic errors, correlated across spatial units

  y = βx + ϵ

  ϵ= λ W ϵ + υ

- **Dependent variable (LM-Lag)**

  Underlying process has led to clustered distribution of variables influence of neighboring values on unit values

  y= λ Wy + βx + ϵ

Where:
W is the Weight Matrix
ϵ is the error

6.3 Weight Matrix (W)

The basis for most models is an indicator of whether one region is a spatial neighbour of another; or equivalently, which regions are neighbours of a given region. This is a square symmetric RXR matrix with *(i,j)* element equal to 1 if regions i and j are neighbours of one another (or more generally, are spatially related), and zero otherwise. By convention, the diagonal elements of this "spatial neighbours" matrix are set to zero. The matrix is then row-standardised in which the rows of the neighbors matrix are made to sum to unity.

As LeSage (1998) points out, there is an embarrassingly large number of ways to construct such a matrix. I am using Rook contiguity matrix (W_c) as defined in Viton [4].

Rook contiguity (so called after the movement of the chess piece): two regions are neighbors if they share (part of) a common border (on any side).

6.4 Other variables

The other attributes studied from each of the cities are GDP, unemployment rate (U), Population density (R), Literacy (L) and Health Index (H).

- **U** is the unemployment and is a proxy for opportunity cost of legal and illegal activities; a positive sign of its coefficient should indicate that people excluded from labour market tend to commit a crime.

- **GDP** is the gross domestic product per capita as proxy for legal and illegal income opportunity; the excepted sign of the coefficient is negative.

- **L is the Literacy Rate** High dropout rate i.e. low literacy higher crime rate. (Galster, 1998)

- **H is the Health Index**

7 Results

7.1 OLS estimates

The model is

$$y = \lambda Wy + \beta_1 U + \beta_2 GDP + \beta_3 R + \beta_4 L + \beta_5 H + \epsilon \quad (1)$$

The estimates are:

Coefficients

	Estimate Std.	Error	t value	$Pr(>\|t\|)$
(Intercept)	-4.228e+02	2.320e+02	-1.822	0.0731 .
gdp	-9.640e-05	3.678e-03	-0.026	0.9792
unemployment	-1.192e+03	5.314e+02	-2.242	0.0284 *
literacy	7.540e+02	2.907e+02	2.594	0.0118 *
health	2.343e+02	4.394e+02	0.533	0.5957
density	77.811e-02	3.442e-02	2.269	0.0266 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129 on 64 degrees of freedom
Multiple R-squared: 0.272, Adjusted R-squared: 0.2151
F-statistic: 4.782 on 5 and 64 DF, p-value: 0.0008936

We can see that only population density, unemployment and literacy are significant and rest all the variables are insignificant while performing OLS. The F statistic indicated that the model is overall significant.

7.2 Moran I Test

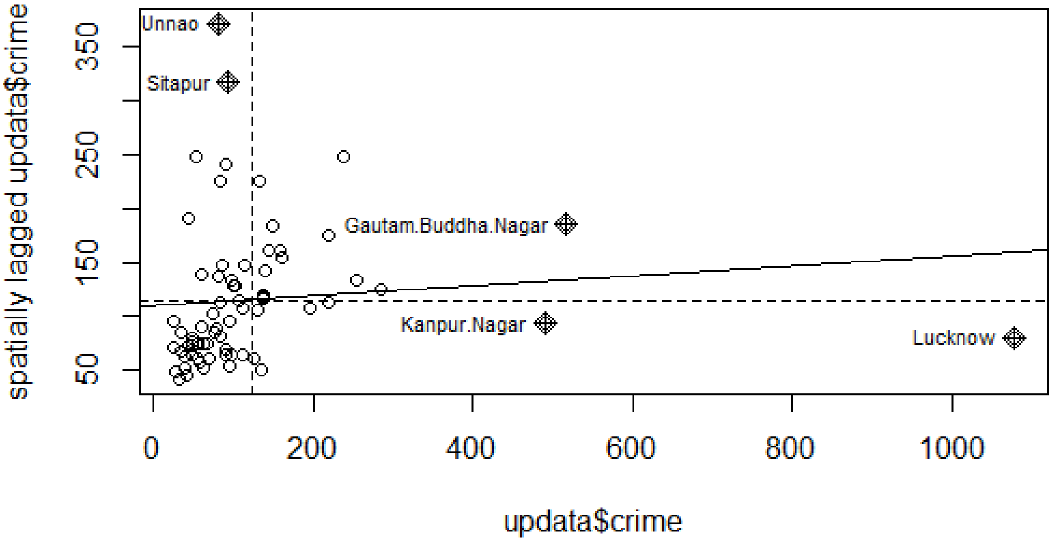
Following are the results of the Moran I test:

Moran I statistic standard deviate = 1.0007, p-value = 0.1585
alternative hypothesis: greater

Moran I statistic	Expectation	Variance
0.046586722	-0.014492754	0.003725639

We can clearly see that the value of the coefficient is very low and the p value is greater than .05.

Moran Plot



From the above graph and the interpretation mentioned in the previous section, we can say that most of the points lie in the "low-low" zone.

7.3 Geary C Test

Geary C statistic standard deviate = 1.3562, p-value = 0.08751		
Alternative hypothesis: Expectation greater than statistic		
Geary C statistic	Expectation	Variance
0.82976142	1.00000000	0.01575586

Again we can see that the value of Geary C coefficient is close to 1 so there is weak positive correlation but p value is still greater than 0.05.

We can saw that there are some local effects but there is *no spatial autocorrelation with respect to crime.*

7.4 LM Tests

We still perform LM tests to confirm whether we can apply any spatial model here or not.

Model	Statistic	Parameter	p-value
LMerr	2.85980	1	0.090819 .
LMlag	0.34341	1	0.557867
RLMerr	8.80285	1	0.003008 **
RLMlag	6.28647	1	0.012166 *
SARMA	9.14626	2	0.010326 *

We can see that both LMerr and LMlag are insignificant but the both RLMerr and RLMlag coefficient are significant.

So, we choose the LMerr model as the coefficient is bigger.

7.5 Spatial Autoregressive Error Model

The model is

$$C = +\beta_1 U + \beta_2 GDP + \beta_3 R + \beta_4 L + \beta_5 H + \epsilon \tag{2}$$

$$\epsilon = \lambda W \epsilon + \upsilon \tag{3}$$

The estimates are

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-7.3453e+02	2.3669e+02	-3.1033	0.001914
gdp	-1.3121e-03	3.5435e-03	-0.3703	0.711180
unemployment	-1.1540e+03	5.1636e+02	-2.2348	0.025428
literacy	9.7788e+02	3.1714e+02	3.0834	0.002046
health	5.5371e+02	4.1780e+02	1.3253	0.185073
density	6.5800e-02	3.3058e-02	1.9904	0.046544

Lambda: 0.37868, LR test value: 3.9154, p-value: 0.047845
Asymptotic standard error: 0.14342 z-value: 2.6404, p-value: 0.0082811
Wald statistic: 6.9716, p-value: 0.0082811

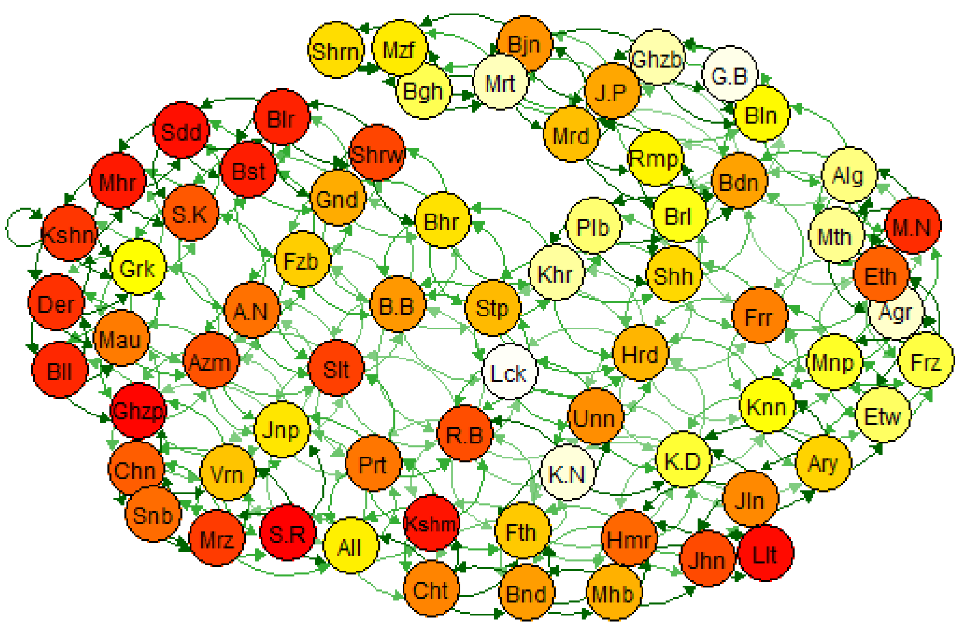
Log likelihood: -434.4218 for error model
ML residual variance (sigma squared): 13901, (sigma: 117.9)
Number of observations: 70
Number of parameters estimated: 8
AIC: 884.84, (AIC for lm: 886.76)

Again Literacy, Unemployment and Population Density are significant but the coefficients are larger indicating that the spatial model gives a better estimate of the model.

We can see that unemployment has a negative impact on crime which was as expected from the social disorganization theory.

The interesting point to note here is that literacy has a positive impact on crime indicating that the main reason for greater crime is not education but perhaps unavailability of basic commodities. Another explanation could be the fact that crimes are usually reported in large cites and crime in smaller districts often goes unnoticed. As the data concerns only "reported" crimes, the resultes may not give the true nature of crime in UP.

7.6 Heat Map to find hotspots



The heat map maps the areas with the highest crime with light colours and the shade increases with decrease in crime rate.

From the graph we can clearly see that there are no ”hotspots” and only low crime cities are clustered together (as seen by the Moran Plot).

We can see that the high crime cities like Lucknow, Gaziabad, Kanpur are far way and as we see from the Moran Plot, these are outliers. While very low crime zones like Basti show some correlation.

The more urban districts have higher population density,

8 Conclusion

After applying the various tests stated above I conclude that there is low spatial correlation with respect to crime in UP.

The crime rate is aptly explained by the socio-economic variables but since data collection is only possible in urban cities we might not get the entire picture of crime in UP.

9 Future Scope

Although the results of this study show weak spatial correlation, this is still a very interesting study in itself. We can clearly see that there are some local effects which can be further seen by applying some other weight matrix as described in LeSage and Pace. Also, acquiring a more robust dataset might help in extracting more prominent results.

A study of this nature is unique and promising for future research.

References

[1] Cracolici, Maria Francesca, and Teodora Erika Uberti. ”Geographical distribution of crime in Italian provinces: a spatial econometric analysis.”Jahrbuch für Regionalwissenschaft 29.1 (2009): 1-28.

[2] Delbecq, Benoît, Rachel Guillian, and Diègo LEGROS. ”Analysis of crime in Chicago: new perspectives to an old question using spatial panel econometrics.”

[3] Pavlo, Iavorskyi. Distribution of Crime Across Ukraine: Panel and Spatial Analysis. Diss. Kyiv School of Economics, 2011.

[4] Viton, Philip A. ”Notes on spatial econometric models.” City and regional planning 870.03 (2010): 9-10.

[5] Ratcliffe, Jerry. ”Crime Mapping: Spatial and Temporal Challenges.” Handbook of Quantitative Criminology. By David Weisburd. New York: Springer, 2010. N. Print.

[6] LeSage, James P., and R. Kelley. Pace. Introduction to Spatial Econometrics. Boca Raton: CRC, 2009. Print.

[7] Ahamed Shafeeq B, Dr. Binu V. “Spatial Patterns of Crimes in India using Data Mining Techniques”. International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 11, May 2014

[8] Anselin, Luc. ”Spatial econometrics.” A companion to theoretical econometrics 310330 (2001).

[9] Anselin, Luc, et al. ”Spatial analyses of crime.” Criminal justice 4.2 (2000): 213-262.