International Journal of Business Analytics and Intelligence
8 (1) 2020, 01-03
http://publishingindia.com/ijbai/

# Analytically Yours:

# Spatial Data Analysis

**Arnab Kumar Laha***

In this article, we discuss a special kind of data that occurs quite routinely that is referred to as "Spatial Data". In this kind of data, there is a geographical location associated with every data element. For example, one may be interested in studying the variation of temperature across different cities of India. Note that the data here is of the form (Variable, Geographical location) where Variable = Temperature of a city and Geographical location = Name of the city. Another example is the number of confirmed COVID-19 cases in a state on July 22, 2020. Here the Variable is the number of confirmed COVID-19 cases in a state on July 22, 2020, and Geographical location = Name of state. Spatial data analysis refers to a set of techniques that are designed to find pattern and test hypotheses and theories, based on spatial data.

Spatial data abound in our daily lives. Every day when we open our newspapers we find information from around our state, country and the world. If you look at these news items carefully you will notice lots of information that can be properly described as spatial data. Some examples are (a) rainfall in different cities of India on the previous day (b) atmospheric pollution index in different cities of India in the previous day and (c) traffic accidents in different parts of the city that occurred the previous day etc. Other examples of spatial data that are of some interest to business and industry are daily labor rate in different parts of India, sales of anti-malarial medication in the different districts of a state, monthly sale of automobiles in different states of India etc.

Visualization of spatial data is typically done using a "map". By a map, we mean a visual representation on a flat surface of the whole or a part of an area. One of the main aims of map construction is to facilitate the visual examination of the values of the attribute variable at different locations. A popular way to do this is to use a colour-coding scheme; different areas of the map are colored differently based on the value of the attribute variable. Such figures are popularly referred to as "heat map". Fig. 1 below gives a simplified heat-map of the number of COVID-19 cases in different states and union territories of India over five months. In this example, the states are arranged in a column, the dates on which the case numbers are recorded is along the rows, and the observation values are colour-coded. Note that in this simplified heat-map an important information which may be of relevance is missing, that is, the information about the neighboring states/union territories. This could be realized by color coding each state in the actual map of India.

In many applications of spatial data, it is observed that observations from nearby locations are associated. This is a major departure from the predominant assumption in classical statistics that the observations are mutually independent. The observed association can be due to various reasons such as a spatial spillover effect (for example economic conditions in a big city may affect the local economies of the smaller cities near to it), distance decline effect (as with temperature where it is found that with increase in distance the degree of association between temperature of two places declines) etc. The lack of independence in the observations makes spatial data analysis challenging.

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

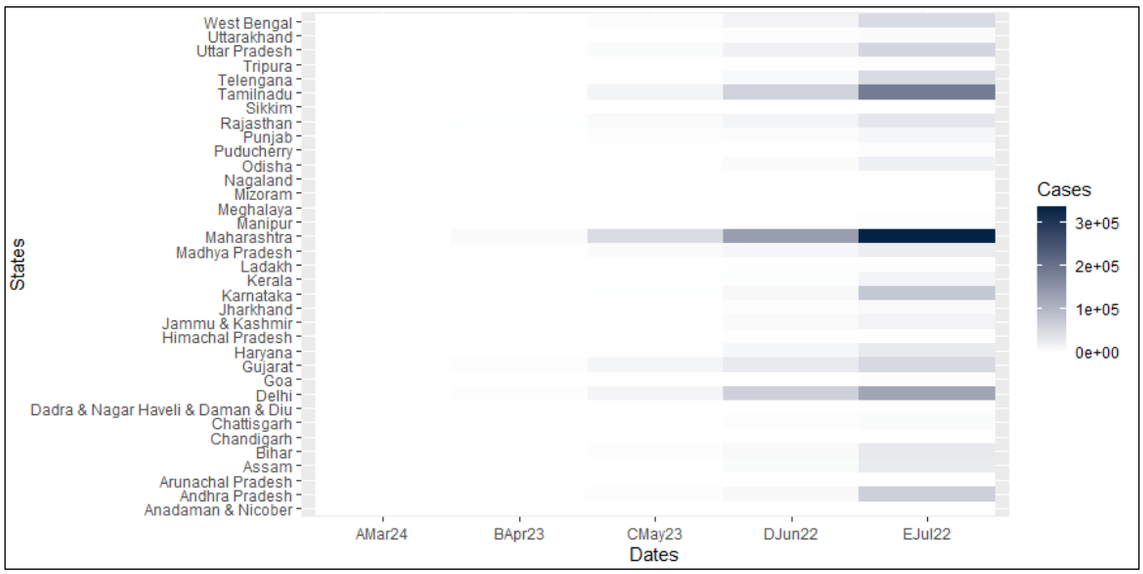


**Fig. 1: Heat Map Depicting the Number of COVID-19 Cases in Different States and Union Territories of India during March-July, 2020**

Several measures have been proposed to capture spatial association such as Moran's *I* and Geary's *c*. In this article we confine our discussion to the widely used measure Moran's I. It is defined as $I = \frac{n}{W_o} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i-\bar{z})(z_j-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2}$ where $W_o = \sum_{i=1}^{n}\sum_{j\neq i}^{n} W_{ij}$ where n is the number of locations, $z_i$'s are value of the variable of interest and $W_{ij}$ are weights that capture the "similarity" of the locations i and j. By convention $W_{ii} = 0$ for all i. Since $E(I) = -\frac{1}{n-1}$ values of I greater than $-\frac{1}{n-1}$ indicate positive spatial autocorrelation while a value of I lesser than $-\frac{1}{n-1}$ indicates negative spatial autocorrelation. Fig. 2 provides a visualization of positive and negative spatial autocorrelation on a grid structure. As in heatmap the darkness of the cell represent the values. We observe that positive spatial autocorrelation leads to clusters of high values whereas negative spatial autocorrelation leads to alternating high and low values preventing any cluster formation. Readers interested to learn more about Spatial data analysis may look at Fischer and Wang (2011).

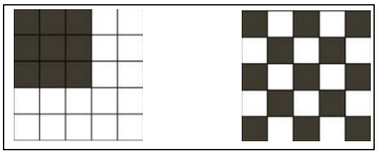

**Fig. 2: The Figure on the Left Hand Side Indicates Positive Spatial Autocorrelation While the Figure on the Right Indicates Negative Spatial Autocorrelation**

Next, we examine the spread of COVID-19 epidemic in India over the five-month period using Moran's I as the measure of spatial autocorrelation. We consider all the 35 States and Union Territories (UTs) that have reported at least one confirmed COVID-19 case in this five-month period. The data for the number of confirmed COVID-19 cases at different dates for these states/UTs are collected from *covid19india.org*. We define $W_{ij} = 1$ if the State/UT 'i' and State/UT 'j' are neighbouring i.e. share a land border else $W_{ij} = 0$. Since n = 35 we have E(I) = -0.029. Table 1 below gives the obtained value of the Moran's I at thirty day intervals between March 24, 2020, to July 22, 2020.

**Table 1: Variation in Moran's I in the Five Month Period March 24, 2020 – July 22, 2020**

| *Date* | *24-03-2020* | *23-04-2020* | *23-05-2020* | *22-06-2020* | *22-07-2020* |
|---|---|---|---|---|---|
| Moran's I | 0.178* | 0.195* | 0.032 | -0.045 | 0.025 |

(* indicates positive spatial autocorrelation at 5% level of significance)

We find from Table 1 that during the months of March and April, 2020 positive spatial autocorrelation was present. This means the states/UTs with higher number of COVID-19 cases were clustered together. This effect could be due to the national lockdown which severely

restricted movement of people across the country, thereby confining the growth of the epidemic to neighbouring states/UTs. However in the later three months i.e. May - July 2020 we do not find any significant spatial autocorrelation in the number of confirmed COVID-19 cases. A possible reason for this could be the gradual freeing of the movement of people across the country since May, 2020.

## Reference

Fischer, M. M., & Wang, J. (2011). *Spatial data analysis: Models, methods and techniques.* Springer.