IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 6, Issue 11, November 2019

A Linear Regression Model for Spatial Data Mining: An Experimental Approach

Arvind Sharma[1], R K Gupta[2]

RJIT, Tekanpur[1]

MITS, Gwalior[2]

Abstract: A GIS data may be collection of spatial and non spatial data types. Feature extraction of spatial data types from a huge data is called spatial data mining. Spatial data mining has accepted as very new and emerging technology for development of system which are applicable directly or indirectly in various field of human needs as e-marketing, cluster analysis of population density , cost estimation of land with forest clustering, geographic trend detection etc. This paper is based on spatial analysis of a city locations which are linearly auto correlated with each other. A real data set has used in this paper and a linear auto correlation is shown between them. Two major attributes as longitude and latitude are selected for experimental purpose. A model is also designed for experimental setup. Experimental setup is completed in PYTHON with graphical touch. Results are very specific and supporting with author's key objectives.

Keywords: Linear regression, Auto correlation, PYTHON, Spatial Data mining

I.INTRODUCTION

In recent years data mining has become the very interesting research area for researchers and other agencies who are working continuously for getting new trends and extraction of new and innovative applications for society. The vast amount of data is analyzed for trend setting and decision making. In this research article we will focus on unique features of spatial data mining and its applications for geospatial analysis, land cost estimation, geo marketing, forest land bifurcation etc. In this research article the author has decided to reveal working of linear regression model for spatial data .The author has selected this type of data because it is more complex than traditional data. The data inputs of spatial data mining have two distinct types of attributes: non spatial attribute and spatial attribute. Non spatial attributes are used to characterize non spatial features of objects such as name, population, and unemployment rate of a city. Spatial attributes are used to define the spatial location and extent of spatial objects[1]. The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape. The linear regression model plays a very important role in data analysis and decision making[3]. In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. It is a method for predicting the value of dependent variable Y, based on the value of an independent variable X.

II.MEANING AND ROLE OF SPATIAL DATA AND SPATIAL DATA MINING

Spatial data is special in many terms with comparison of traditional data[4]. As we know better that spatial data is a special data which contains attributes related to space. Some of the special characteristics are –

1. A spatial pattern has the spatial outliers, location prediction models, spatial clusters and location as attributes (longitude and latitude)and so on.
2. Spatial data mining is used in GIS, geo marketing, crime mapping, land costing, census data, transportation, public safety, natural resource availability and many more. With the help of analysis of spatial data and spatial data mining we can answer the following type of questions –
How is the global earth system changing? What is the cost of a particular land area? How does the earth system respond to natural and human included changes? How well we can predict the local and global trends of human likes?
3. Data has different meaning for simple data mining and spatial data mining. So the associated terms are also have different meaning as information, knowledge for different data mining.

As to predict, we can give a sample for both continuous and disccerte types
Continuous: trend e.g., regression
Location aware: spatial auto – regression model
Discrete: classification e.g., Baysian classifiers
Location aware: Random fields
In this research article, we will through some light on trend detection i.e. for linear regression.

DOI 10.17148/IARJSET.2019.61110

III. RELATED WORK

The actual definition and concept of linear regression was first explained by Sir Francis Galton in 1894. Linear regression [5]is a statistical test applied to a data set to define and quantify the relation between the considered variables. Chang (2003,2004) applied univariate tests such as t-test, Chi-Square, Fisher's Exact test to firm this concept that the effect of other covariates/confounders during analysis is meaningless. Chang (2003) stated that partial correlation and regression are the tests that allow the researcher to control the effect of confounders in the understanding of the relation between two variables[6].

In any application of various fields such as medical, engineering, finance or environmental, the researchers often try to understand or relate two or more independent (predictor) variables to predict an outcome or dependent variable[2]. This may be understood as how the risk factors or the predictor variables or independent varaibles account for the prediction of the chance of a disease occurance, i.e., dependent variable. Risk factors (or dependent variables) associate with biological (such as age or gender) physical (such as body mass index or Blood Pressure[BP]), or lifestyle (such as smoking and alcohol consumption) variables with the disease. It was well defined and explained by Gaddis and Gaddis in 1990. Regression analysis allows predicting the value of a dependent variable based on the value of at least one independent variable[7].

Significance of Linear Regression

The linear regression model[2] is used in various applications due to following reasons-

1. Descriptive- Strength of bond between the dependent and independent variables
2. Adjustment- To adjust effect of cofounders or covariates
3. Predictors- To estimate the risk factors
4. Extent of prediction – Change in independent variable by one unit , how much change towards dependent variables
5. Prediction – It helps in quantifying the new cases

IV. DATA COLLECTION AND ANALYSIS

Data sets were collected as real data sets by the author from different agencies and internet sources. Application of results and data sets depends on the need of society and users. In first stage the collected spatial data was applied for generation of results as a linear regression model. This concept was implemented with the help of Python language.
Following data set is applied for checking linear regression concept between spatial attributes and found result accordingly[10]. This data is imported from 044218.csv file.
A sample of data set is shown below-

Table 1

id	Location	Opening Hours	Dates	Seniors/children	Latitude	Longitude
1	Church of Nazarene, 6B Meadowlands Road, Carina	9am - 11am	Tuesday weekly	children/seniors	-27.491294	153.112764
2	Chermside Library- meeting room, 375 Hamilton ...	1pm - 3pm	Tuesday weekly	children/seniors	-27.385810	153.035109
3	Corinda Community Health Centre, 2 Clara St, C...	1.30pm - 3pm	Thursday weekly	children/seniors	-27.537139	152.984941
4	Forest Lake Community House, corner Forest Lak...	9am - 10.30am	Monday fortnightly	children/seniors	-27.611204	152.961724
5	Holy Family Church Hall (lower level), ward St...	9am - 11am	Wednesday weekly	children/seniors	-27.499859	152.981808

Only 5 samples are shown from thousands of records.
Code of piece of python for reading this data from file-

```
df = pd.read_csv('E:\\phdwork\\044218.csv')
df.head()
```

Data type is identified as for the given data set-

```
id                int64
Location          object
Opening Hours     object
Dates             object
Seniors/children  object
```

```
Latitude          float64
Longitude          float64
dtype: object
df[['Latitude','Longitude']].corr()
```

Output for above line of code[8] produced in the given form below for getting relationship between spatial attributes–

Table 2

	Latitude	**Longitude**
Latitude	1.000000	0.138717
Longitude	0.138717	1.000000

Intercepting status is identified as
Longitude = 0.1022 + 155.844*Latitude

V. RESULT AND DISCUSSION

The data is mentioned in Table 1[12] is analyzed and imported by using a standard platform of Python language. With the application of linear regression concept , the program generates a graphical view for assertion of relation ship between two attributes as Longitude and Latitude[9]
Code for plotting of graph with result is given below-

```
import matplotlib.pyplot as plt
sns.regplot(x=x,y=yhat,data=df)
plt.xlabel('Latitude')
plt.ylabel('Longitude')
plt.title('Longitude vs Latitude')
plt.show()
```

Graphical representation of linear regression is given below-

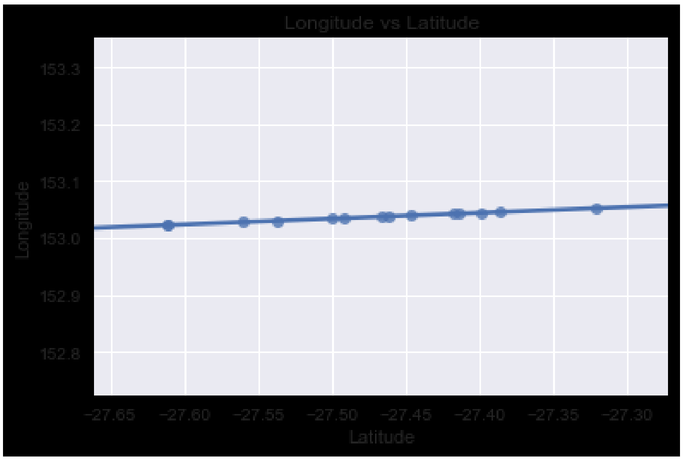

Figure 1 : A graphical plotting of spatial attributes longitude vs latitude for given set of data in table 1 as linear regression model.

Above figure favors the phenomena that in spatial data mining nearby locations have autocorrelation between them. This trend was also tested and verified for other different data sets and found that nearby locations set certain trends and after a certain time and distance these trends are shifted in another cluster of objects with similar properties[11. When these trends are linked together and analyzed, it has found that their formation is based on auto correlation and helpful in prediction of trends of any field which applicable for society and different organization.

VI. CONCLUSION

The method or technique for setting relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. In this research article, we have used simple example and calculations for illustration of linear regression analysis and pave the way of further research and study the concept of linear regression.

Conflicts of Interests

There are no conflicts of interest

REFERENCES

[1]. Arvind Sharma , R K Gupta . An improved IDBSCAN algorithm for spatial databases.
[2]. Chan YH. Biostatistics 201: Linear regression analysis. Age(Years). Singapore Med J 2004;45:55-61.
[3]. Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl int 2010;107:776-82
[4]. Arvind Sharma ;R K Gupta .Intelligent knowledge discovery in spatial data sets .IJRECE Vol.4,issue 1,Jan-March. 2016.pp- 67-73.
[5]. M Ester,Hans Peter Kriegel,J sander,X Xu.Density connected sets and their applications for trend detection in spatial databases.KDD-97.
[6]. Jin Xingxing,cai Yingkun.Ma XiuJun at el.Novel method to integrate spatial data mining and geographic information system. IEEE 2005.pp. 764-767.
[7]. M may ,and A. savinov .An integrated platform for spatial data mining and interactive visual analysis .Proceeding data mining 2002, Bologna,Italy.
[8]. Y xia ,X X Fu.An improved approach and application for spatial data mining.IEEE ,2007. Pp. 32-37.
[9]. Li D.R ,wang S.L. at el . Theories and technologies of spatial data knowledge discovery.Geomatics and information science of Wuhan university, vol. 27,N0.-3,Pp. 221-233,2002.
[10]. Jain ,Murty and Flynn.Geographic data mining and knowledge discovery.
[11]. Arvind Sharma , R K Gupta.Improved density based spatial clustering and applications with noise(IDBSCAN).Vol 2016,Article id 1564516.SCI Hindawi publication.
[12]. Arvind Sharma, R K Gupta. A survey of spatial data mining :Algorithms and architecture. Pp. 15-22,2012.
[13]. Arvind Sharma, R K Gupta. Spatial data mining with the application of spectral clustering. A trend detection approach.Vol 173, 2017, pp.11-18
[14]. Markus M. Breunig,hans Peter Kriegel at el.LOF:identifying density based local outliers.Proceedings ACM SIGMOD,Dalles,TX,2000.
[15]. Jing Gao,Suny Buffalo.Density based methopds.

DOI 10.17148/IARJSET.2019.61110