International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Spatial Data Mining towards Geospatial Data Analysis for Discovery of Spatial Correlations

D. V. Lalitha Parameswari[1] , Ch. Mallikarjuna Rao[2] , Bh. Prashanthi[3], D.Ushasree[4],B.Indu Priya[5]

Submitted: 11/11/2022 **Revised:** 13/01/2023 **Accepted:** 10/02/2023

Abstract—Spatial data provides geographical correlations that can be discovered through Spatial Data Mining (SDM). Such discovery can have potential benefits as it bestows necessary knowledge to understand the patterns. There are many existing methods for correlation analysis. In this paper we focused on the discover of spatial correlations based on G statistic and ZG score computations. We proposed a framework for geospatial data analytics. We also proposed an algorithm known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD). This algorithm is meant for spatial correlation analysis and Principal Component Analysis (PCA) to discover trends pertaining to spatial correlations in the given Twitter data based on given words. The algorithm performs spatial correlation analysis based on given words and the location from which such tweet has originated. Experimental results revealed that our framework is useful for geospatial data analysis.

Keywords— *Spatial Data Mining, Geospatial Data Analysis, G Statistic, ZG Score Computations, Spatial Correlation Analysis*

1. Introduction

Spatial data analysis has become an important research in the contemporary era. It can be used indifferent applications such as traffic forecasting, weather updates and many such applications. It is observed that spatial data can also have non-spatial observations that play important role in discovery of knowledge. Different techniques are found in literature as explored in [4], [5], [6], [7] and [8]. Qinjun *et al*. [4] proposed a text mining approach that is coupled with spatial data processing towards generating spatial analysis results in terms of geoscience reports. Senzhang *et al*. [5] used deep learning for discovering spatial features from given dataset. Maria *et al*. [6] incorporated different techniques considering big spatial data towards emergency management for given business system. Wesley [7] focused on analysis of climate data and challenges in its processing. Fernandez *et al*. [8] explored SDM for finding situational analysis in maritime related analysis. From the literature, it is observed that there are many techniques used for spatial data analysis considering temporal domain as well. In this paper, we threw light on spatial correlation discovery based on geographical dataset and given analysis words. Our contributions are as follows.

1. We proposed a framework where we focused on the discover of spatial correlations based on G statistic and ZG score computations.
2. We also proposed an algorithm known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD).
3. We built an application that is used to realize the framework and underlying algorithm besides testing them for their intended functionality.

The remainder of the paper is structured as follows. Section 2 reviews literature on different existing SDM methods. Section 3 presents proposed methodology. Section 4 reveals the observations in the empirical study. Section 5 concludes our work.

1 Associate Professor, Department of CSE, GNITS, Hyderabad
2,3,4,5Department of Computer Science and Engineering, GRIET, Bachupally, Hyderabad
lplalitap97@gmail.com1, professorcmrao@gmail.com2, bhupathi.prashnthi@gmail.com3 , dupakuntlausha@gmail.com 4 , indu.balappagari@gmail.com5

2. Related Work

This section provides review of different existing works. Soltani *et al*. [1] proposed a method for spatial and temporal analysis to know house price variations in different regions. Jinchao *et al*. [2] focused on SDN towards finding traffic congestion in given regions. It also discovers factors pertaining to traffic congestion. Shashi *et al*. [3] explored computations involved in spatio-temporal mining. Qinjun *et al*. [4] proposed a text mining approach that is coupled with spatial data processing towards generating spatial analysis results in terms of geoscience reports. Senzhang *et al*. [5] used deep learning for discovering spatial features from given dataset. Maria *et al*. [6] incorporated different techniques considering big spatial data towards emergency management for given business system. Wesley [7] focused on analysis of climate data and challenges in its processing. Fernandez *et al*. [8] explored SDM for finding situational analysis in maritime related analysis. Yousuf *et al*. [9] proposed a clustering mechanism that is used for spatial data analysis. Hamdi *et al*. [10] focused on SDM dynamics and

the opportunities associated with besides problems in knowledge discovery.

Monidipa *et al*. [11] explored a deep learning model on the remote sensing data to discover patterns linked to geographical analysis. Ghislain *et al*. [12] proposed a forecasting model on spatio-temporal data using short term based deep learning model. Xiao-Li [13] followed data science approach to discover trends and patterns from spatial data. Berkay *et al*. [14] focused on the discovery of co-occurrence patterns based on SDM techniques from the given spatial data. Shaik *et al*. [15] investigated on SDM procedures based Recurrent Neural Network (RNN) method. From the literature, it is observed that there are many techniques used for spatial data analysis considering temporal domain as well. In this paper, we threw light on spatial correlation discovery based on geographical dataset and given analysis words.

3. Materials and Methods

Tweets dataset containing tweets that come from various regions of Canada is used for the empirical study. The dataset is collected by writing a Python program that exploits Twitter API that is publicly available. It is known as geospatial dataset which contains geography and spatial information.

3.1 The Framework

We designed and implemented a framework for automatic analysis of geospatial data based on given geospatial dataset and certain words for analysis. The framework is as shown in Figure 1. The given inputs are processed by the framework with the help of SDM phenomena. The framework is aimed at finding the regions from which given words are originated and provide a visualization so that it is easy to understand the patterns and word usage dynamics across given regions of Canada.

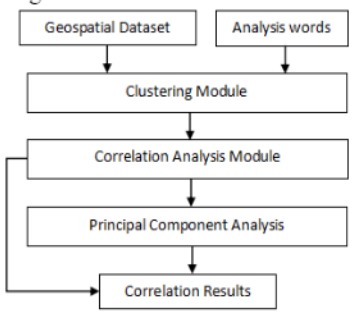


Fig 1. Proposed system for spatial data analysis

The given dataset is subjected to clustering using Fuzzy K Means algorithm. It is soft clustering which has determination of number of clusters based on the given number of analysis words. After performing clustering, the given clusters help in improving speed in correlation analysis. This correlation analysis module has underlying algorithm proposed. The proposed algorithm is known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD). This algorithm is meant for spatial correlation analysis and Principal Component Analysis (PCA) to discover trends pertaining to spatial correlations in the given Twitter data based on given words. The algorithm performs spatial correlation analysis based on given words and the location from which such tweet has originated. Then PCA analysis provides top 3 PCA components that reflect the word usage dynamics in the given geographical area.

3.2 Clustering

Fuzzy K-Means [16] is used for clustering in the proposed framework. It performs clustering based on given analysis words. The algorithm computes the degree of belongingness as expressed in Eq. 1.

$$\forall x \sum_{k=1}^{num.clusters} u(x) = 1 \quad (1)$$

The centroid computation is carried out as given in Eq. 2.

$$center_k = \frac{\sum_x uk(x) mx}{\sum_x uk(x) m} \quad (2)$$

The cluster and its inverse of distance dynamics with respect to degree of belonging is computed as in Eq. 3.

$$u_{k(x)} = \frac{1}{d(Centerk,x)} \quad (3)$$

Then the normalization of coefficients and fuzzification are carried out as in Eq. 4.

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(Centerk,x)}{d(Centerj,x)}\right)^{2/(m-1)}} \quad (4)$$

The normalization standardises the sum as 1. Closer the m value to 1 indicates a point has probability to belong to given cluster.

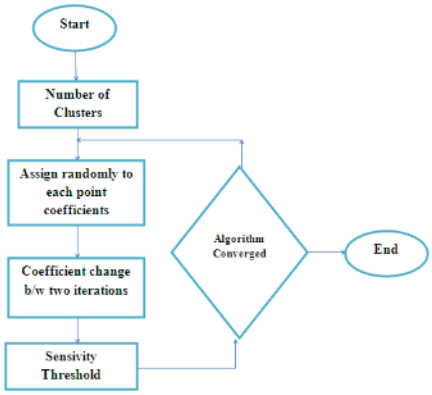


Fig 2. Shows Fuzzy C means algorithm's functionality

As presented in Figure 2, it is evident that there is an iterative process until convergence into final set of clusters.

3.2 Algorithm Design

Given area of the study has different concentrations of words. The concentration may be low or high. Let G be the spatial correlation; it is computed as in Eq. 5.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \forall j \neq i \quad (5)$$

Where the features are denoted as i and j. Between given features the spatial weight is represented as wi, j. The total number of features is represented as n. A rule denoted as ∀j≠i is followed which indicates that two features should not be equal. Then the G and ZG are computed as in Eq. 6, Eq. 7 and Eq. 8.

$$Z_G = \frac{G - E[g]}{\sqrt{V[G]}} \quad (6)$$

$$E[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)}, \forall j \neq I \quad (7)$$

$$V[G] = E[G^2]] - E[G]^2 \quad (8)$$

The G value in Eq. 1 can be evaluated a value in range 0 to 1. G statistic is based on null hypothesis which evaluates the presence of spatial correlations. In the process when z-score is computed, a positive value shows the possibility of bigger index for G and the cluster contains more concentrated values. This kind of analysis provides useful knowledge on the concentration of given words in different regions. And finally PCA analysis provides maximum variations that occurred in the given geospatial dataset.

Algorithm: Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD)
Inputs: Geospatial dataset D, analysis words W
Output: Spatial correlation discovery and visualization

Begin
C←FuzzyCMeans(D,W)
For each c in C
  Compute G statistic as in Eq. 1
  Compute E[G] as in Eq. 7
  Compute V[G] as in Eq. 8
  Compute ZG as in Eq. 6
  Use ZG to find correlations
  Visualize correlations
  Compute PCA
End For
 For each pca in top 3 PCAs
   Print pca
Visualize pca
End For
End

Algorithm 1: Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD)

As presented in Algorithm 1, it takes Geospatial Dataset D, analysis words W as inputs and performs spatial correlation mining in order to find correlations among regions based on given analysis words. It has iterative process to compute ZG score and also PCA besides visualization results.

4. Experimental Results

This section presents experimental results in terms of value of getisord for given words, PCAs and also visualization of correlations shown for each analysis word

Table 1: Shows Value Of Getisord For Given Words

CSDUID	Software	Hardware	English	Work	Job	Skill	Footbal l
5915022	1.140124	0.402140	0.516985	4.354657	0.588630	0.602145	0.740006
5919008	0.625988	0.065075	0.062320	4.096794	0.096330	0.293729	0.303626
3520005	4.931679	5.272617	5.054173	1.184907	2.577718	4.242798	5.199579
4611040	-0.504226	-0.385416	-0.701735	-0.303505	-0.683416	-0.322143	-0.195547
3539036	6.784597	6.990854	5.283371	1.369412	3.647614	5.406260	7.218184

As presented in Table 1, the CSDUID based analysis is provided to reflect getisord value computed for each analysis word.

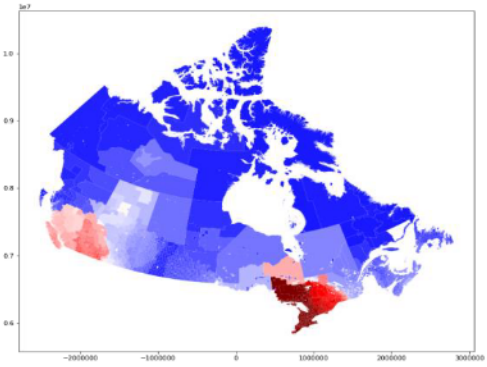
Table 2: Shows The Top 3 Principal Components And Values For Corresponding Csduid

	Pc1	Pc2	Pc3	CSDUID
0	96.651481	99.256246	2.913660	5915022
1	51.111416	73.827625	-0.228917	5919008
2	447.509678	-14.257330	2.031535	3520005
3	-36.079143	-7.418042	-1.113544	4611040
4	595.960360	-20.103161	2.954482	3539036

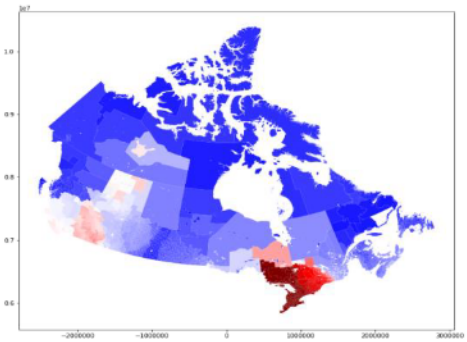
As presented in Table 2, it shows the top 3 principal components and values for corresponding CSDUID.

Word Visualization of discovered spatial correlation

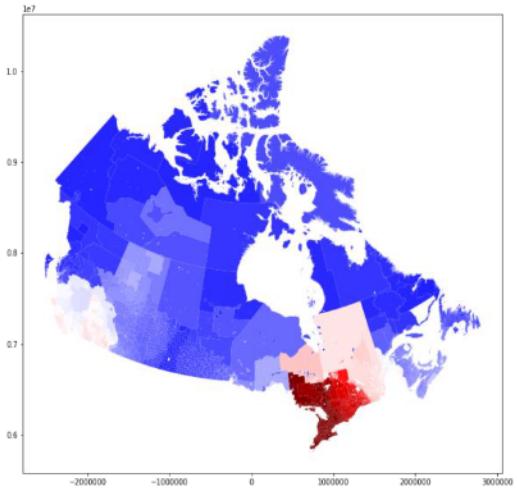
Software



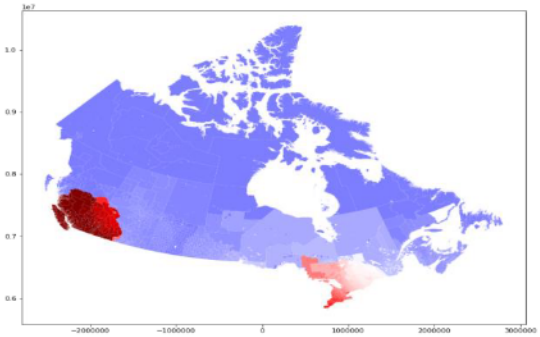
Hardware



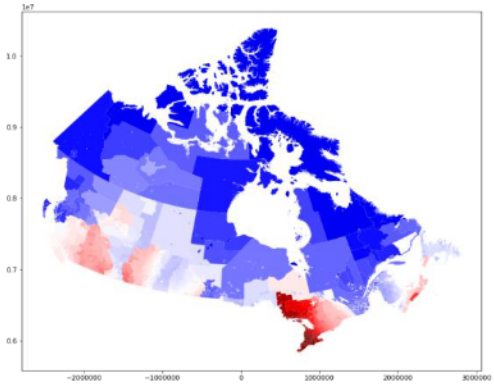
English



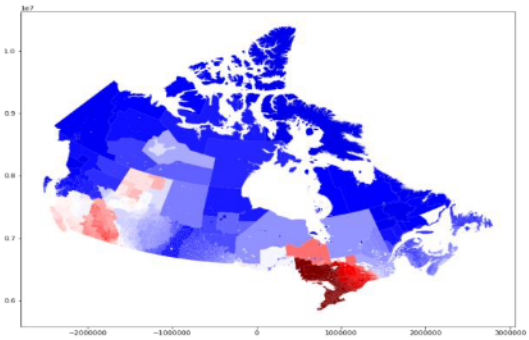
Work



job



Skill



Football

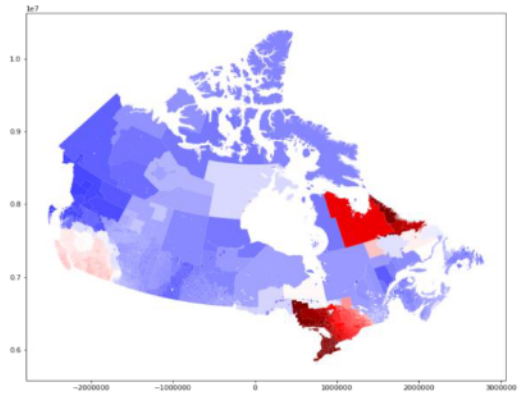
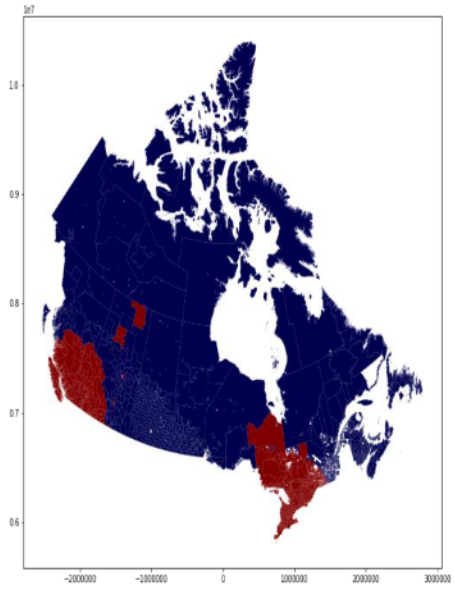


Fig 3. Shows analysis word and corresponding visualization of geographic correlation

As presented in Figure 3, there is spatial correlation analysis visualized for each analysis word given. For each word, the geographical analysis result is visualized.

pc1



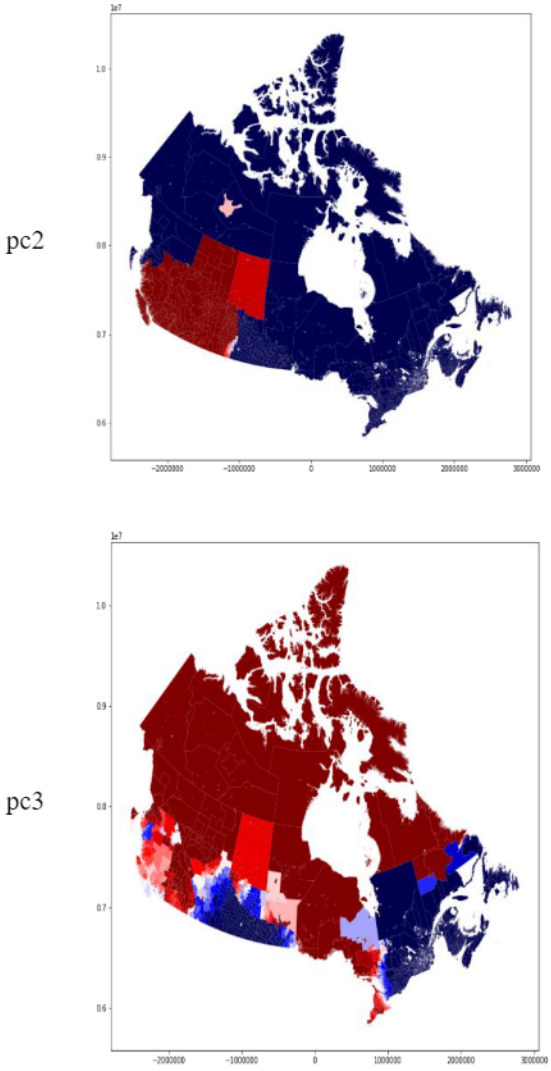


Fig 4. Top 3 principal components visualization

As presented in Figure 4, top 3 PCAs are discovered and visualized. These components provide the concentrations of words in different regions. Thus spatial correlation analysis using SDM provides useful insights that can be integrated with any real world applications.

V. Conclusion And Future Work

In this paper, we proposed a framework for geospatial data analytics. We also proposed an algorithm known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD). This algorithm is meant for spatial correlation analysis and Principal Component Analysis (PCA) to discover trends pertaining to spatial correlations in the given Twitter data based on given words. The algorithm performs spatial correlation analysis based on given words and the location from which such tweet has originated. Experimental results revealed that our framework is useful for geospatial data analysis. In future, we intend to improve our framework to have automatic discovery of trending words and perform correlation analysis.

References

[1] Ali Soltani;Christopher James Pettit;Mohammad Heydari;Fatemeh Aghaei; (2021). Housing price variations using spatio-temporal data mining techniques . Journal of Housing and the Built Environment, p1-29.

[2] Song, Jinchao; Zhao, Chunli; Zhong, Shaopeng; Nielsen, Thomas Alexander Sick; Prishchepov, Alexander V. (2019). Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques. Computers, Environment and Urban Systems, 77, p1-12.

[3] Shekhar, Shashi; Jiang, Zhe; Ali, Reem; Eftelioglu, Emre; Tang, Xun; Gunturi, Venkata; Zhou, Xun (2015). Spatiotemporal Data Mining: A Computational Perspective. ISPRS International Journal of Geo-Information, 4(4), p2306–2338.

[4] Qiu, Qinjun; Xie, Zhong; Wu, Liang; Tao, Liufeng (2020). Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. Earth Science Informatics, p1-18.

[5] Wang, Senzhang; Cao, Jiannong; Yu, Philip (2020). Deep Learning for Spatio-Temporal Data Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering, p1–20.

[6] Dagaeva, Maria; Garaeva, Alina; Anikin, Igor; Makhmutova, Alisa; Minnikhanov, Rifkat (2019). Big spatio-temporal data mining for emergency management information systems. IET Intelligent Transport Systems, 13(11), p1649–1657.

[7] Chu, Wesley W. (2014). [Studies in Big Data] Data Mining and Knowledge Discovery for Big Data Volume 1 || Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities. , p83–116.

[8] Arguedas, Virginia Fernandez; Mazzarella, Fabio; Vespe, Michele (2015). [IEEE OCEANS 2015 - Genova - Spatio-temporal data mining for maritime situational awareness. , p1–8.

[9] Ansari, Mohd Yousuf; Ahmad, Amir; Khan, Shehroz S.; Bhushan, Gopal; Mainuddin, (2019). Spatiotemporal clustering: a review. Artificial Intelligence Review, p1-43.

[10] Ali Hamdi;Khaled Shaban;Abdelkarim Erradi;Amr Mohamed;Shakila Khan Rumi;Flora D. Salim; (2021). Spatiotemporal data mining: a survey on challenges and open problems . Artificial Intelligence Review, p1-48.

[11] Das, Monidipa; Ghosh, Soumya K. (2016). Deep-STEP: A Deep Learning Approach for Spatiotemporal Prediction of Remote Sensing Data. IEEE Geoscience and Remote Sensing Letters, p1–5.

[12] Agoua, Xwegnon Ghislain; Girard, Robin; Kariniotakis, Georges (2017). Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production. IEEE Transactions on Sustainable Energy, p1–9.

[13] Li, Xiao-Li; Cao, Tru; Lim, Ee-Peng; Zhou, Zhi-Hua; Ho, Tu-Bao; Cheung, David (2015). [Lecture Notes in Computer Science] Trends and Applications in Knowledge Discovery and Data Mining Volume 9441 || Mining Massive-Scale Spatiotemporal Trajectories in Parallel: A Survey. , p41–52.

[14] Aydin, Berkay; Kempton, Dustin; Akkineni, Vijay; Gopavaram, Shaktidhar Reddy; Pillai, Karthik Ganesan; Angryk, Rafal (2014). [IEEE 2014 IEEE International Conference on Big Data (Big Data) - Washington, DC, USA (2014.10.27-2014.10.30)] 2014 IEEE International Conference on Big Data (Big Data) - Spatiotemporal indexing techniques for efficiently mining spatiotemporal co-occurrence patterns. , p1–10.

[15] Mohammed Ali Shaik;Dhanraj Verma;P Praveen;K Ranganath;Bonthala Prabhanjan Yadav; (2020). RNN based prediction of spatiotemporal data mining . IOP Conference Series: Materials Science and Engineering, p1-12.

[16] Sheshasaayee, A., & Sridevi, D. (2016). *Fuzzy C-means algorithm with gravitational search algorithm in spatial data mining. 2016 International Conference on Inventive Computation Technologies (ICICT).* P1-5.