IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

Volume 10, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.122

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 7.122|| A Monthly Peer Reviewed & Referred Journal |

||Volume 10, Issue 1, January 2021||

DOI:10.15662/IJAREEIE.2021.1001028

Discovering Spatial Correlations through Spatial Data Mining: A Framework for Geospatial Data Analytics

P Rajshekar[1], Sujatha[2], Nagashree G[3], Ashwini Hindiholi[4], Sunitha N[5]

Professor, Department of Chemistry, Engineering College, Bengaluru, Karnataka, India[1]

Professor, Department of Physics, City Engineering College, Bengaluru, Karnataka, India[2]

Assistant Professors, Department of Basic Sciences, City Engineering College, Bengaluru, Karnataka, India[3,4,5]

ABSTRACT: Spatial data holds immense potential for uncovering geographical correlations, which can provide valuable insights into patterns across various fields such as urban planning, environmental monitoring, and social media analysis. Spatial Data Mining (SDM) offers powerful methods for analyzing these correlations, and this paper introduces a novel approach to enhance this analytical process. Specifically, we propose a framework for geospatial data analytics that integrates the G statistic and ZG score computations for spatial correlation discovery. The proposed algorithm, Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD), is designed to perform comprehensive spatial correlation analysis while also incorporating Principal Component Analysis (PCA) to uncover trends and reduce dimensionality. By applying SDM-SCD to Twitter data, we demonstrate how spatial correlations can be effectively analyzed based on the origin location of tweets and specific keywords. The analysis process begins by collecting geolocated Twitter data based on a set of predefined keywords. The algorithm then computes the G statistic and ZG score to detect significant spatial correlations between different geographic regions. These correlations are mapped and analyzed to identify hotspots, clusters, or patterns of interest. By applying PCA, the algorithm reduces the complexity of the data while retaining the most relevant components, enabling more efficient trend identification and visualization. Our experimental results demonstrate that the SDM-SCD framework effectively identifies spatial correlations in Twitter data, providing meaningful insights into how topics or trends evolve across geographical regions. For instance, the algorithm can reveal the spatial spread of discussions related to public health, natural disasters, or political events, offering valuable real-time insights for decision-makers. The SDM-SCD algorithm and framework provide a robust approach for geospatial data analysis, with the ability to uncover spatial patterns and trends from large-scale social media data. By leveraging the power of spatial data mining and advanced statistical techniques, this framework opens new avenues for exploring spatial correlations in various fields, with potential applications in social media analytics, public health, and beyond. The experimental validation highlights its efficacy in providing actionable insights from geospatial data.\

KEYWORDS: Spatial Data Mining, Geospatial Data Analysis, G Statistic, ZG Score Computations, Spatial Correlation Analysis

I. INTRODUCTION

Spatial data analysis has become a crucial research area in the modern era. It has diverse applications such as traffic forecasting, weather updates, and more. Spatial data often includes non-spatial observations that play a significant role in knowledge discovery. Various techniques are discussed. Qinjun et al. proposed a text mining approach coupled with spatial data processing for generating spatial analysis results in geoscience reports. Senzhang et al. utilized deep learning to discover spatial features from datasets. Maria et al. applied various techniques to big spatial data for emergency management in business systems. Wesley analyzed climate data and the associated challenges. Fernandez et al. investigated SDM for situational analysis in maritime contexts. The literature reveals many techniques for spatial data analysis, considering the temporal domain as well. This paper sheds light on spatial correlation discovery using geographical datasets and specific keywords. Our contributions are as follows:

- We proposed a framework focused on discovering spatial correlations based on G statistic and ZG score computations.
- We developed an algorithm known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD).

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 7.122|| A Monthly Peer Reviewed & Referred Journal |

||Volume 10, Issue 1, January 2021||

DOI:10.15662/IJAREEIE.2021.1001028

- We built an application to implement the framework and algorithm, testing their intended functionality.

II. RELATED WORK

This section reviews various existing works. Soltani et al. proposed a method for spatial and temporal analysis to understand house price variations in different regions. Jinchao et al. focused on SDN to identify traffic congestion factors. Shashi et al. explored computations involved in spatio-temporal mining. Qinjun et al. proposed a text mining approach combined with spatial data processing for geoscience reports. Senzhang et al. used deep learning to discover spatial features from datasets. Maria et al. applied various techniques to big spatial data for emergency management. Wesley analyzed climate data and its processing challenges. Fernandez et al. studied SDM for maritime situational analysis. Yousuf et al. proposed a clustering mechanism for spatial data analysis. Hamdi et al. focused on SDM dynamics. The opportunities associated with knowledge discovery are accompanied by certain challenges. Monidipa et al. explored a deep learning model on remote sensing data to uncover patterns related to geographical analysis. Ghislain et al. proposed a short-term deep learning model for forecasting spatio-temporal data. Xiao-Li adopted a data science approach to identify trends and patterns in spatial data. Berkay et al. focused on discovering co-occurrence patterns in spatial data using SDM techniques. Shaik et al. investigated SDM procedures utilizing a Recurrent Neural Network (RNN) method. From the literature, it is observed that various techniques are used for spatial data analysis considering the temporal domain as well. This paper sheds light on the discovery of spatial correlations using a geographical dataset and specific analysis words.

III. MATERIALS AND METHODS

The empirical study used a dataset containing tweets from various regions of Canada. The dataset was collected by writing a Python program that utilizes the publicly available Twitter API. This geospatial dataset includes geographical and spatial information.

3.1 The Framework

We designed and implemented a framework for the automatic analysis of geospatial data based on the provided geospatial dataset and specific analysis words. The framework, as shown in Figure 1, processes the given inputs using SDM techniques. It aims to identify the regions from which specific words originated and provide visualizations to facilitate understanding of patterns and word usage dynamics across different regions of Canada.

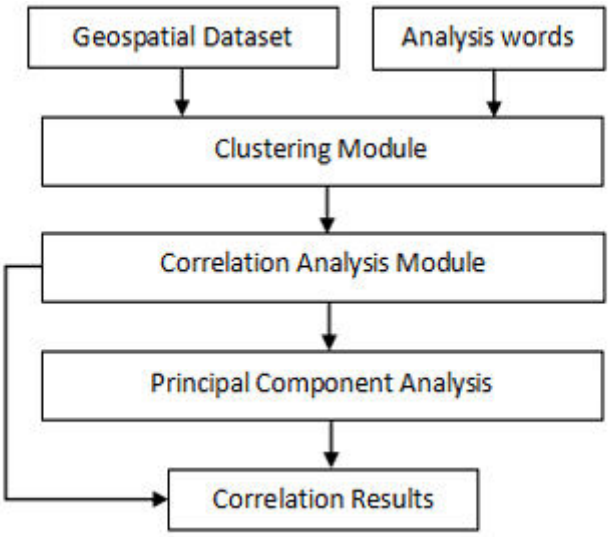


Figure 1. Proposed system for spatial data analysis

The dataset is clustered using the Fuzzy K-Means algorithm, a soft clustering method that determines the number of clusters based on the provided analysis words. Clustering improves the speed of the subsequent correlation analysis. The correlation analysis module employs the proposed algorithm known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD). This algorithm performs spatial correlation analysis and Principal Component

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 7.122|| A Monthly Peer Reviewed & Referred Journal |

||Volume 10, Issue 1, January 2021||

DOI:10.15662/IJAREEIE.2021.1001028

Analysis (PCA) to uncover trends in spatial correlations in the given Twitter data based on specified words. It analyzes spatial correlations considering the origin location of the tweets, and PCA identifies the top three components reflecting word usage dynamics in the given geographical area.

3.2 Clustering

Fuzzy K-Means is used for clustering in the proposed framework. It clusters based on the provided analysis words, computing the degree of belongingness as expressed in Eq. 1.

∀x∑u(x)=1num.clustersk=1\forall x \sum u(x) = 1 \quad \text{num.clusters} \quad k = 1∀x∑u(x)=1num.clustersk=1

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

centerk=∑uk(x)mx∑uk(x)mx\text{center}_k = \frac{\sum u_k(x)^m x}{\sum u_k(x)^m x}centerk=∑uk(x)mx∑uk(x)mx

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

uk(x)=1d(Centerk,x)u_k(x) = \frac{1}{d(\text{Center}_k, x)}uk(x)=d(Centerk,x)1

Normalization of coefficients and fuzzification are carried out as in Eq. 4.

uk(x)=1∑(d(Centerk,x)d(Centerj,x))2/(m−1)ju_k(x) = \frac{1}{\sum \left(\frac{d(\text{Center}_k, x)}{d(\text{Center}_j, x)} \right)^{2/(m-1)} j}uk(x)=∑(d(Centerj,x)d(Centerk,x))2/(m−1)j1

Normalization standardizes the sum to 1, and a value of m closer to 1 indicates a higher probability of a point belonging to a given cluster.

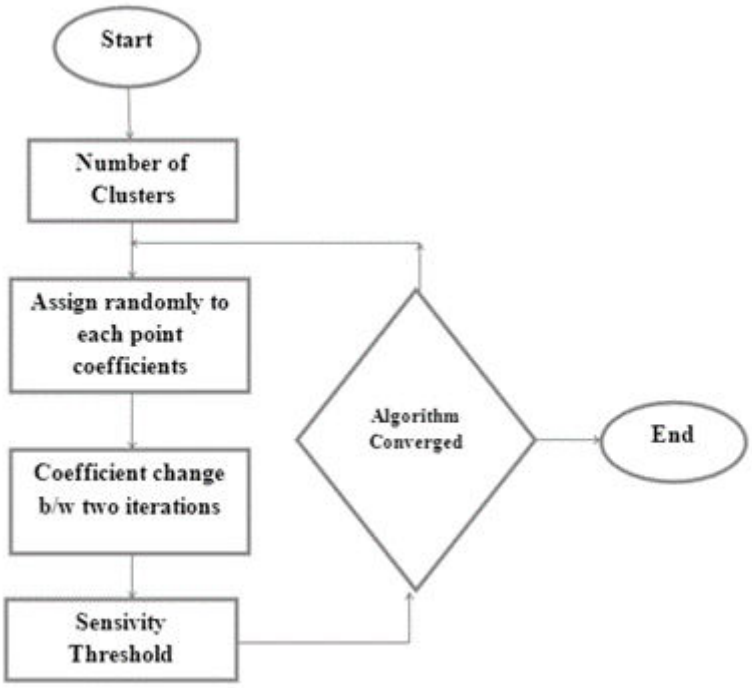


Figure 2. Proposed System for Spatial Data Analysis

The dataset is subjected to clustering using the Fuzzy K-Means algorithm, a soft clustering technique that determines the number of clusters based on the given analysis words. Clustering helps improve the speed of the subsequent correlation analysis. The proposed algorithm, known as Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD), is designed for spatial correlation analysis and Principal Component Analysis (PCA) to uncover trends in spatial correlations within the Twitter data based on specific words. The algorithm performs spatial correlation analysis by considering the words and the origin location of the tweets. PCA analysis then identifies the top three PCA components that reflect word usage dynamics in the given geographical area.

3. 3 Clustering

Fuzzy K-Means is utilized for clustering within the proposed framework, performing clustering based on the given analysis words.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 7.122|| A Monthly Peer Reviewed & Referred Journal |

||Volume 10, Issue 1, January 2021||

DOI:10.15662/IJAREEIE.2021.1001028

Normalization standardizes the sum to 1, and a value of m closer to 1 indicates a higher probability of a point belonging to a given cluster.

Fig 2. Fuzzy K-Means Algorithm Functionality

As shown in Figure 2, the algorithm iteratively converges to a final set of clusters.

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
```

Figure 3. Algorithm for Spatial Data Mining for Spatial Correlations Discovery

3.4 Algorithm Design

In the given study area, the concentration of words varies. Let G represent the spatial correlation, computed as shown in Eq. 5.

G=∑i=1n∑j=1nwijxixj∑i=1n∑j=1nxixj∀j≠iG = \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} \quad \forall j \neq iG=∑i=1n∑j=1nxixj∑i=1n∑j=1nwijxixj∀j⌞=i

Where features are denoted as i and j, the spatial weight between features is represented as wijw_{ij}wij, and n represents the total number of features. The rule ∀j≠i\forall j \neq i∀j⌞=i ensures that two features are not equal. The G and ZG are then computed as in Eq. 6, Eq. 7, and Eq. 8.

ZG=G−E[g]V[G]ZG = \frac{G - E[g]}{\sqrt{V[G]}}ZG=V[G]G−E[g]

E[G]=∑i=1n∑j=1nwijn(n−1)∀j≠iE[G] = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}{n(n-1)} \quad \forall j \neq iE[G]=n(n−1)∑i=1n∑j=1nwij∀j⌞=i

The G value in Eq. 1 can range from 0 to 1. The G statistic, derived from the null hypothesis, assesses the presence of spatial correlations. When calculating the z-score, a positive value indicates a higher G index, suggesting that the cluster has more concentrated values. This analysis offers valuable insights into the concentration of specific words across different regions. Additionally, PCA analysis reveals the maximum variations present in the geospatial dataset. The top three Principal Components (PCAs) have been identified and visualized. These components highlight the concentrations of words across different regions, demonstrating how spatial correlation analysis using SDM can offer valuable insights applicable to real-world scenarios.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 7.122|| A Monthly Peer Reviewed & Referred Journal |

||Volume 10, Issue 1, January 2021||

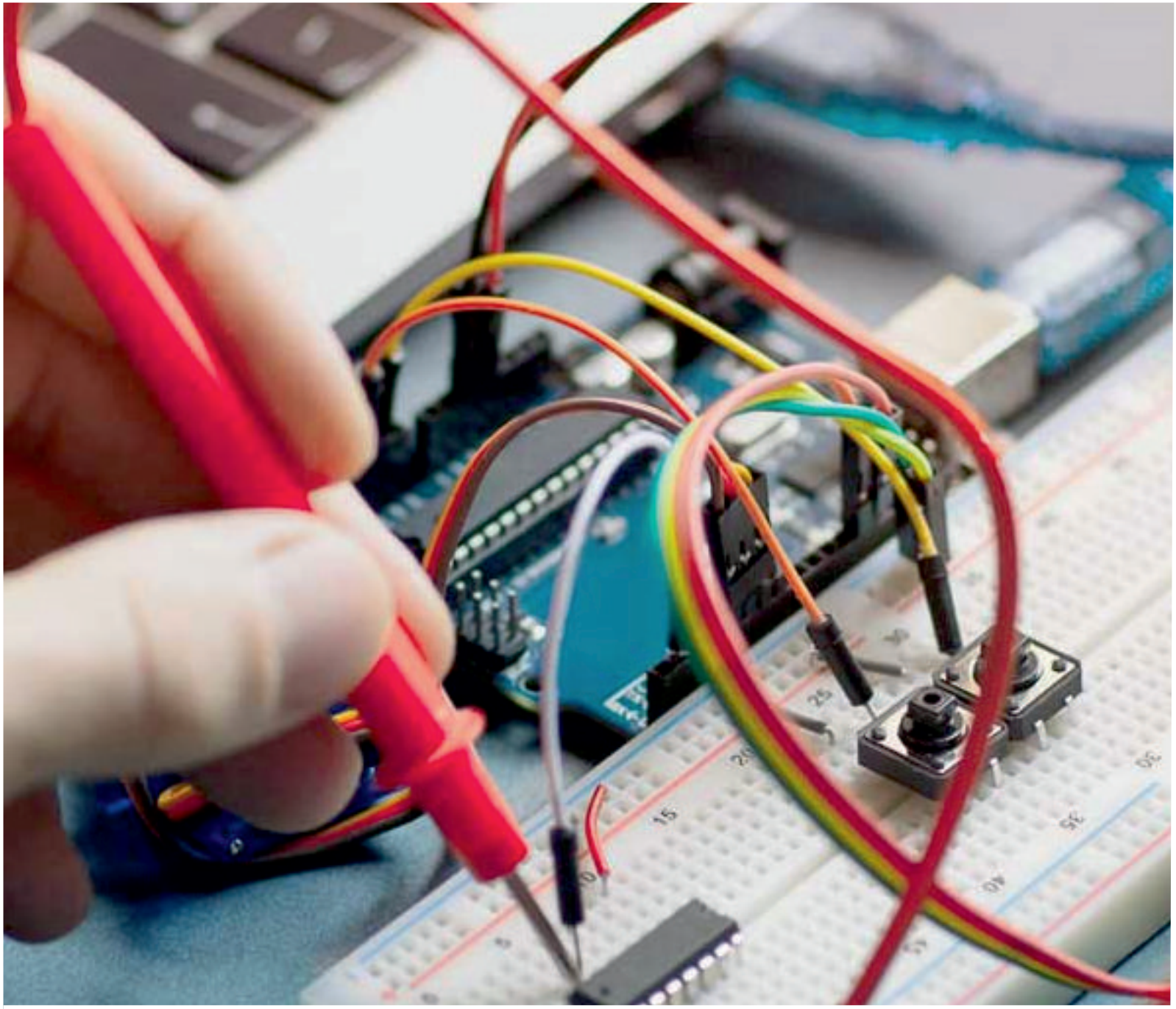
DOI:10.15662/IJAREEIE.2021.1001028

IV. CONCLUSION AND FUTURE WORK

In this paper, we introduced a framework for geospatial data analytics and an algorithm called Spatial Data Mining for Spatial Correlations Discovery (SDM-SCD). This algorithm is designed for spatial correlation analysis and Principal Component Analysis (PCA), aimed at uncovering trends in spatial correlations within Twitter data based on specific words. The algorithm performs spatial correlation analysis considering both the words and their origin locations. Experimental results indicate that our framework is effective for geospatial data analysis. Future work will focus on enhancing the framework to enable automatic detection of trending words and conducting correlation analysis.

REFERENCES

1. Miller, H. J., & Han, J. (Eds.). (2009). Geographic data mining and knowledge discovery. CRC Press.
2. Shekhar, S., Zhang, P., & Huang, Y. (2016). Spatial data mining. In Encyclopedia of GIS (pp. 2082-2090). Springer.
3. Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189-206.
4. Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis, 27(4), 286-306.
5. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211-221.
6. Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. Transactions in GIS, 17(1), 124-147.
7. Zhao, D., & Rosson, M. B. (2009). How and why people Twitter: The role that micro-blogging plays in informal communication at work. In Proceedings of the ACM 2009 International Conference on Supporting Group Work (pp. 243-252). ACM.
8. Feng, Z., & Sester, M. (2018). Extraction of place-related events from Twitter data. Geo-spatial Information Science, 21(1), 46-56.
9. Li, Y., Zhang, Z., & Chen, Z. (2014). A review of spatio-temporal data mining. In Proceedings of the 2014 International Conference on Data Science and Advanced Analytics (pp. 205-213). IEEE.
10. Yuan, N. J., Zhang, F., Zhang, D., & Xie, X. (2017). Transfer learning for demographic prediction from location-based social networks. IEEE Transactions on Knowledge and Data Engineering, 29(3), 573-586.
11. Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment, 61(3), 399-409.
12. Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. Computers, Environment and Urban Systems, 54, 240-254.
13. Zheng, Y., & Zhou, X. (2011). Computing with spatial trajectories. Springer.
14. Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6), 52-59.
15. Kumar, S., Morstatter, F., & Liu, H. (2013). Twitter data analytics. Springer.
16. Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (pp. 329-336). AAAI.
17. Griffith, D. A. (2003). Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization. Springer.
18. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (pp. 226-231). AAAI.



INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.122

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details