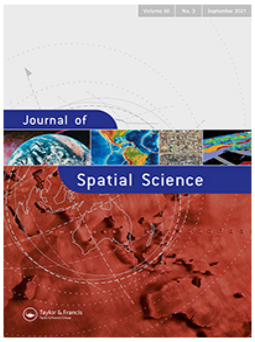


Taylor & Francis
Taylor & Francis Group

Journal of Spatial Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjss20

# Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity

Zeinab Ghaemi & Mahdi Farnaghi

**To cite this article:** Zeinab Ghaemi & Mahdi Farnaghi (2021): Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity, Journal of Spatial Science, DOI: 10.1080/14498596.2021.2002201

**To link to this article:** https://doi.org/10.1080/14498596.2021.2002201

JOURNAL OF SPATIAL SCIENCE
https://doi.org/10.1080/14498596.2021.2002201

# Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity

Zeinab Ghaemi[a] and Mahdi Farnaghi[b]

[a]Faculty of Geodesy and Geomatics, K.N.Toosi University of Technology, Tehran, Iran; [b]Faculty of Geo-Information Science and Earth Observation, Geo-Information Processing, University of Twente, Enschede, Netherlands

**ABSTRACT**
Twitter, as the most popular social media platform, has made a great revolution in producing real-time user-generated data. This research aims to propose a method to extract the latent spatial pattern from geotagged tweets. We take both spatial autocorrelation and spatial heterogeneity into account while revealing the underlying pattern from geotagged tweets. Moreover, the textual similarity is considered to extract spatial-textual clusters. The method was implemented and tested during hurricane Dorian on the east coast of the U.S. The results proved the superiority of the proposed method against Moran's Index and VDBSCAN algorithms in extracting clusters with various densities.

**KEYWORDS**
Spatial clustering; twitter; hurricane; coastal cities; spatial autocorrelation; spatial heterogeneity

## Introduction

Over the past decades, hurricanes have adversely affected coastal cities and imposed massive loss of lives and damages to the properties (Webster *et al*. 2005). Therefore, gathering real-time information about disasters is paramount for hazard mitigation, preparedness, response, and recovery in residential areas (Gall *et al*. 2011, Gonick and Errett 2018).

In recent years, social media has played an essential role in disaster management (Basu *et al*. 2016, Kankanamge *et al*. 2019). The increasing popularity of social media, along with equipping smart devices with the Global Positioning System, have resulted in exponential growth in user-generated spatial data (Kisilevich *et al*. 2009, Ghodousi *et al*. 2019). Twitter, as one of the most popular social networks, provides users with the possibility to share their opinions and emotions about the events they have witnessed (Lachlan *et al*. 2014, Qi and John E. 2014, Wei 2020). Analysis of such data, enriched with geo-location, has become an opportunity for managers to obtain appropriate information to determine efficient strategies for disaster preparedness and response (Nolasco and Oliveira 2019, Yabe and Ukkusuri 2019).

In this regard, the purpose of this research is to propose a method called Varied Density-based spatial Clustering for Autocorrelated Twitter data (VDCAT), to extract the latent spatial pattern from geotagged tweets. The significant aspect of VDCAT is its ability

**CONTACT** Zeinab Ghaemi zghaemi@mail.kntu.ac.ir
Supplemental data for this article can be accessed here

to consider both spatial autocorrelation and spatial heterogeneity while revealing the underlying pattern from geotagged tweets. Moreover, in order to extract spatial-textual clusters, VDCAT account for the textual similarity as another dimension in addition to spatial proximity. Having a method with the mentioned specification, this study intends to explore and analyze spatial-textual clusters during hurricane Dorian on the eastern coastal cities of the U.S. Analyzing such clusters assist in investigating areas affected by the hurricane and places where residents need help.

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3 introduces the study area and dataset, the theoretical background, and the proposed method for event detection from geotagged tweets. The results are presented in section 4 and discussed in section 5. Finally, section 6 concludes the study and suggests future works.

## Background and related works

Geographical Information Systems (GIS) and spatial analysis are great tools for analyzing, managing, and extracting latent information from spatial data (Saeidian *et al*. 2018), and in our case, geotagged tweets. Spatial clustering, in particular, is an efficient method for discovering the spatial pattern of geographical phenomena from geotagged data (Kisilevich *et al*. 2009, Stojanova *et al*. 2013, Ghodousi *et al*. 2016). Clustering methods have been used in different applications (Aggarwal, 2014), for example, disaster management (Röthlisberger *et al*. 2017, Wu *et al*. 2017), sentiment analysis during and after a disaster (Kankanamge et al., 2019; Karmegam and Mappillairaju 2020), health analysis (Smiley *et al*. 2010, Roth *et al*. 2016, Wu *et al*. 2017), and event detection, to name a few. Regarding the event detection from geotagged tweets, Farnaghi *et al*. (2020) proposed a spatio-temporal tweet mining method for dynamic event extraction from geotagged tweets. Souza *et al*. (2019) extracted regions prone to dengue disease, taking the location and content of tweets into account. Temporal hashtag clustering methods were used for spatio-temporal event detection from the Twitter stream (Feng *et al*. 2015). Local events were explored in real-time from a Twitter stream by clustering tweets based on spatial proximity and keyword similarity (Abdelhaq *et al*. 2013).

Meanwhile, two spatial effects of autocorrelation and heterogeneity lead the analysis of spatial data, including the analysis of geotagged tweets, to be more complicated. Spatial autocorrelation occurs when the proximate instances of a phenomenon are more similar to each other than the further ones. (Stojanova *et al*. 2011). Considering spatial autocorrelation in spatial analysis is crucial as it violates the common assumption of observation independence in traditional statistical methods (Balducci and Ferrara 2018). Ignoring spatial autocorrelation in a dataset with spatial dependency may lead to obscure results or even inverted observation patterns (Kühn 2007, Stojanova *et al*. 2013). In the case of geotagged tweets, previous studies have proved that geotagged tweets are affected by spatial autocorrelation, meaning that tweets are usually surrounded by similar ones, and nearby tweets are dealing with the same topics (Steiger *et al*. 2015). Consequently, to perform a more robust and accurate clustering of geotagged tweets, spatial autocorrelation should be taken into account (Wang and Yue 2013).

Although several studies have employed spatial clustering to identify events from geotagged tweets, considering spatial autocorrelation is rarely conducted. Yang and Mu (2015) extracted depressed users from Twitter geotagged data using local Morans's Index and hot spot analysis. Exploring urban structures through spatiotemporal and semantic pattern analysis of Twitter users' social activities was conducted by Steiger *et al*. (2015). Spatial and social clustering of cholera-related tweets using Moran's Index has been conducted by Giebultowicz *et al*. (2011).

Despite its strength in taking spatial autocorrelation into account, local Moran's I is unable to deal with spatial heterogeneity in the input dataset (Besag and Newell 1991, Zhang and Lin 2016). Moreover, variation in the distance band of local Moran's I leads to various results and output clusters. The other limitation of using local Moran's I in extracting clusters from geotagged tweets is its inability to consider various attribute values concurrently. It means that only clusters related to one topic can be extracted in each execution.

Spatial heterogeneity, which refers to the non-stationarity of the underlying spatial processes (Brunsdon *et al*. 1998), is another critical issue that should be considered in the spatial clustering of geotagged tweets (Ghaemi and Farnaghi 2019). This phenomenon should be considered especially for analysis of Twitter data in large areas, where the number of Twitter users varies over the study area, leading to variation in the density of published tweets (Blank 2017). Different density-based clustering methods have been proposed for event detection based on the density of tweets. However, only a few studies considered spatial heterogeneity for event extraction from geotagged tweets (Ghaemi and Farnaghi 2019). Fast-greedy optimization of modularity (FGM) clustering algorithm was enhanced using VDBSCAN (varied density-based spatial clustering of applications with noise) to extract communities during typhoon Haiyan (Bakillah *et al*. 2015). K-dist plot was utilized by VDBSCAN to extract density levels. In another study, VDCT (Varied Density-based spatial Clustering for Twitter data) was proposed to extract clusters from geotagged tweets. Exponential Spline Interpolation was employed in this study to extract various density levels (Ghaemi and Farnaghi 2019). Despite their ability to handle spatial heterogeneity, VDBSCAN and VDCT still encounter some challenges in event detection from geotagged tweets. Using k-dist plot and exponential spline interpolation for calculating distances does not guarantee that spatial autocorrelation is considered. The second challenge is that the algorithms, such as DBSCAN and VDBSCAN work with 2-dimensional data (x,y). However, working with the Twitter dataset, it is required to consider the similarity between content of tweets to distinguish various events.

Although spatial heterogeneity and spatial autocorrelation are two momentous issues in clustering geotagged tweets, methods for dealing with both simultaneously are still missing in the previous studies. To address this requirement, the current study proposes a method for event detection from geotagged tweets using density-based spatial clustering, which takes spatial autocorrelation and heterogeneity into account to extract accurate clusters with varied densities and combines textual similarity with spatial proximity to extract spatial-textual clusters. The method provides the disaster response organizations with information to help the affected areas more efficiently, assist victims, and mitigate damages.

## Materials and methods

### *Study area and input dataset*

Hurricane Dorian was a severe tropical cyclone that struck the Bahamas in August and September 2019 and also affected Florida and Georgia. The area affected by hurricane Dorian including Bahamas, Florida, and Georgia, was chosen as the study area. According to the U.S. Census Bureau estimates, Florida and Georgia's population was almost 21 and 10 million, where 51.1 and 51.4% were females, respectively. Also, 19.7% and 23.6% of the population were under 18 years old, and 20.9% and 14.3% were over 65 years old in Florida and Georgia, respectively. The population of Bahamas was almost 395,000, where 22.4% were under 18, and 9.3% were over 60 years old.

Utilizing the Twitter streaming API, the geotagged tweets were collected from August 31 to September 10 for the study area. Based on the collected tweets, the number of shared tweets increased from September 1, when the hurricane made landfall in the Bahamas, till September 8, when its intensity dwindled (Figure S1). The collected geotagged tweets are depicted in Figure 1. A few examples of the collected tweets during the hurricane are listed below.

- *'Please don't let your guard down. Be prepared and safe. we are tracking #hurricanedorian and will keep you updated.', (Sat August 31 11:34:59, 2019)*
- *'Calm before the storm. Everything is going to alright #hurricanedorian #allgoodintentions', (Sat August 31 13:12:22, 2019).*
- *'Hurricane #dorian will hit the coast at 140 mph but we're ready and prepared for the worst . . . ', (Sat August 31 14:35:29, 2019)*
- *'Pray for all my friends and fam in the path of Dorian . . . ', (Sun September 01 19:13:37, 2019).*
- *"Good Morning! The sun is shining and looks like #Dorian is giving us a break, (Sun September 08 16:08:02, 2019).*

### *Spatial autocorrelation*

In order to extract the latent pattern from geotagged tweets, it is required to analyze the dependency of similar or dissimilar tweets over the study area. One fundamental principle of spatial analysis is that the values of a variable in approximate locations are more similar than the distinct ones which are quantified by Tobler's first law of geography (Tobler 1970). In positive spatial autocorrelation, the values of a variable in nearby locations are surrounded by other similar values (Stojanova *et al*. 2013). Considering spatial autocorrelation in spatial analysis, and specifically clustering, is important as it results in extracting more stable clusters (Glotsos *et al*. 2004, Jahani and Bagherpour 2011). Global spatial autocorrelation evaluates whether the existing pattern over the study area is clustered, random, or dispersed. Global Moran's Index (Moran 1950) is utilized in this study to reveal the latent pattern of geotagged tweets. Local spatial autocorrelation, on the other hand, identifies the location of spatial clusters (Anselin 1995).

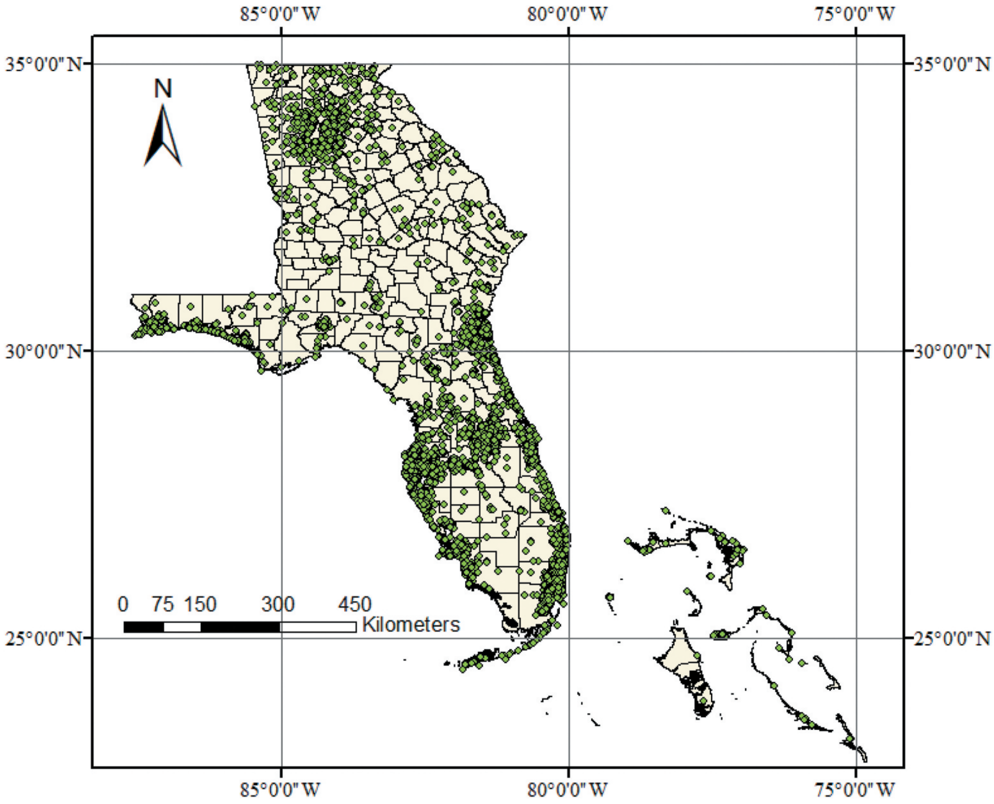


**Figure 1.** Study area and collected tweets.

In order to have robust results, it is efficient to set the input distance band when applying Moran's I (Global or local), and the tweets which are located outside the distance are ignored. In order to calculate an appropriate distance value for Moran's Index, Variogram is used in this study (Cressie and Hawkins 1980). In a variogram graph, the range indicates the distance beyond which the samples are not correlated because the similarity between close points is greater than the distant ones (Bohling 2005). This value is a suitable candidate for the distance band in Moran's I. Two indicators of z-score (deviation from the mean) and p-value (probability that the spatial pattern is randomly created), calculated by spatial autocorrelation, are utilized to indicate whether the null hypothesis (Complete Spatial Randomness) is rejected (Mohammadinia *et al*. 2017).

Along with the above-mentioned indexes, incremental spatial autocorrelation (ISA) was proposed and has been employed to test spatial autocorrelation within a sequence of distances (Choudhary *et al*. 2015). In ISA, an iterative process is conducted, and spatial autocorrelation is determined at various distances using global Moran's I. At each iteration, z-score is obtained as the result of global Moran's I and indicates the significance level of spatial autocorrelation (Lu *et al*. 2019). The output of ISA is a graph that presents the association between z-score and various distances. The peaks of the graph are appropriate candidates for distance-based algorithms as they present the distances where spatial clustering is significant.

### *Spatial heterogeneity*

Spatial heterogeneity, in the case of the Twitter dataset, refers to the uneven distribution of tweets (Brunsdon *et al*. 1998). The number of Twitter users varies over the study area based on different factors, for example, population density, education, age, and economic condition (Mislove *et al*. 2011, Sloan *et al*. 2015, Blank 2017). Such variations result in heterogeneity in the distribution and density of shared tweets. Among density-based clustering algorithms, VDBSCAN (varied density-based spatial clustering of applications with noise) was proposed by Liu *et al*. (2007) to extract clusters with varied densities from heterogeneous input datasets. It is an extension of DBSCAN (Density-based spatial clustering of applications with noise) (Ester *et al*. 1996) and utilizes k-dist plot to determine different levels of densities over the input dataset. Sharp changes of the plot are considered as the candidate distances for clustering. The distance values are sorted in ascending order, and clustering is iteratively performed based on the DBSCAN mechanism utilizing the lowest to the highest values (Liu *et al*. 2007). DBSCAN receives two parameters of *epsilon* and minimum points (*minPnts)* as inputs and extract clusters based on the density of points. Randomly selecting a point, DBSCAN investigated if there exist at least *minPnts* within the distance of *epsilon* around the point. If yes, the point is considered a *core point*, and the cluster will grow up by checking the same procedure for neighboring points.

### *Event detection considering spatial autocorrelation and heterogeneity from twitter*

This study proposes an algorithm for event detection from geotagged tweets, called Varied Density-based spatial Clustering for Autocorrelated Twitter data (VDCAT) that can overcome the shortcomings of existing methods in dealing with both spatial autocorrelation and heterogeneity simultaneously. The algorithm has been built as an extension to previous methods, mainly VDCT (Ghaemi and Farnaghi 2019) and VDBSCAN (Liu *et al*. 2007).

VDCAT method consists of two main steps of *text pre-processing* and *event detection*. In general, geotagged tweets are collected and pre-processed in the first step. Latent Dirichlet Allocation (LDA) is then used to extract topics and corresponding probability distribution over tweets. The output of this step is fed to *event detection* for extracting spatial-textual clusters. In the second step, ISA is initially utilized to extract values of *epsilon*, the distance threshold for searching neighbouring tweets, and then the algorithm iteratively runs through different values of *epsilon* to extract clusters with various densities using DBSCAN algorithm. Considering different values for distance based on density and using ISA to calculate these distance values lead to considering both spatial heterogeneity and autocorrelation in extracting clusters. The overall analysis framework of each step is demonstrated in Figure 2.

As an initial step, geotagged tweets (only tweets with latitude and longitude) are retrieved in real-time using the official Twitter streaming API. Then, in a pre-processing procedure, the tweet texts are converted to lowercase, unwanted symbols, URLs, and numbers are removed. The punctuation signs are deleted, and hashtags are replaced by their text. Tokenization is used to split each tweet into an array of single words, and short

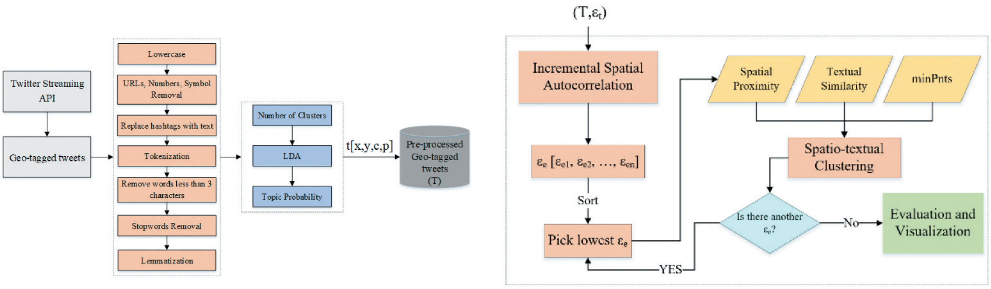

Twitter Streaming API
Geo-tagged tweets
Lowercase
URLs, Numbers, Symbol Removal
Replace hashtags with text
Tokenization
Remove words less than 3 characters
Stopwords Removal
Lemmatization
Number of Clusters
LDA
Topic Probability
t[x,y,c,p]
Pre-processed Geo-tagged tweets (T)
(T,ε_t)
Incremental Spatial Autocorrelation
ε_e [ε_e1, ε_e2, …, ε_en]
Sort
Pick lowest ε_e
Spatial Proximity
Textual Similarity
minPnts
Spatio-textual Clustering
Is there another ε_e?
No
YES
Evaluation and Visualization

**Figure 2.** Text pre-processing (left) and event detection (right) steps.

words with less than three characters were eliminated. Stop words are then removed as they do not contain valuable information. The remaining words are then converted to their roots through a lemmatization process.

Having the pre-processed texts, the next step is to uncover meaningful topics from geotagged tweets. Currently, there is no labelled dataset of tweets that can be used to train a supervised classifier for event detection. Also, preparing such a dataset is time-consuming and would not be helpful in the long run because of the changes in the content of the tweets in response to a different event. Therefore, unsupervised methods were used in this study to classify tweets into different topics. LDA proposed by Blei *et al.* (2002) is a widely used unsupervised classification method with a particular use case in natural language processing. It is a standard topic modeling method that receives documents (in our case, geotagged tweets) as input and extracts possible topics for the received documents as output. Since LDA has proven its feasibility in several Twitter event detection studies (Cheng *et al.* 2014, Morchid *et al.* 2015, Steiger *et al.* 2015), we used it in this research to classify tweets based on their textual contents.

LDA uses a 'bag of words' approach and considers each document (in our case, each tweet) as a probability distribution over topics and regards each topic as a probability distribution over words. LDA model calculates the probability that a tweet may belong to a topic. This probability value is used to conduct ISA in this study. LDA extracts the topics and calculates a probability value for each topic-tweet pair that estimates how each particular tweet may be assigned to each specific topic. At this stage, which is the final step of text pre-processing, each tweet is saved in a database as a tuple t[x, y, c, p], where *x* and *y* are the geographical coordinates, *c* is the cleaned text of the tweet and *p* is the probability of the tweet being assigned to the topic of interest in this study, which is the hurricane Dorian.

As shown in Figure 2, (T, $\varepsilon_t$) are fed to the event detection step to extract clusters from geotagged tweets. *T* is the collection of tweets that are saved in the database as tuple [x, y, c, p], and $\varepsilon_t$ is the threshold of text similarity between tweets. First, the value of *minPnts* (the minimum number of tweets that should exist in a cluster) is calculated based on the total number of geotagged tweets (*n*) using the heuristic approach proposed by Suthar *et al.* (2013). To extract clusters over the heterogeneous Twitter dataset, various levels of densities should be determined. Extracted values are used as radius distance ($\varepsilon_e$) for searching neighbouring tweets. Given the topic probability of each tweet (*p*), Incremental Spatial Autocorrelation (ISA) is utilized to extract the density levels while considering spatial autocorrelation between tweets. ISA measures spatial autocorrelation based on Global Moran's I for a series of distances between tweets and calculates z-score for each distance. The Global Moran's I is calculated through the following equations:

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{ij}(p_i - \hat{p})(p_j - \hat{p})}{\sum_{i=1}^{N} (p_i - \hat{p})^2} \tag{1}$$

$$S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{2}$$

In Equation 1, $p_i$ and $p_j$ are the dependencies of tweet $i$ and tweet $j$ on hurricane topic, respectively; $\hat{p}$ is the mean of probability values; $W_{ij}$ is the spatial weight between tweet $i$ and $j$ which varies based on the variation in distance of ISA; $N$ is the total number of tweets; and $S_0$ is the sum of all spatial weights. The output graph, which shows the relationship between z-scores and various distances, is drawn, and its peaks are extracted as appropriate values of $\varepsilon_e$. In order to extract clusters, the values of $\varepsilon_e$ are sorted in ascending order, and the lowest value of $\varepsilon_e$ is chosen as the starting step of spatial clustering. A tweet is randomly selected, and its neighbours are extracted. In order to find neighbor tweets, both spatial closeness and text similarity between tweets are considered. To add text similarity as a new dimension, the conditions presented in Equations 3 and 4 are used in this study to find the neighboring tweets.

$$\mathrm{EuclDist}(\mathrm{t}, \mathrm{t}') \leq \varepsilon_{\mathrm{e}} \text{ , } TextSim(\mathrm{t}, \mathrm{t}') \geq \varepsilon_t \tag{3}$$

$$N_\varepsilon(t \text{ , } t') \equiv \frac{\varepsilon_t}{\varepsilon_e} \times \frac{EuclDist(t,\ t')}{TextSim(t,\ t')} \leq 1 \tag{4}$$

In Equations 3 and 4, $\varepsilon_{\mathrm{t}}$ is the threshold of text similarity between geotagged tweets, and $\varepsilon_{\mathrm{e}}$ is the distance threshold for searching neighbors. Also, Euclidean distance is used to calculate the closeness in spatial dimension.

Cosine similarity is employed to measure the similarity of the contents of the tweets. For a tweet $t \in T$, a vector of $[n_{w_1}, n_{w_2}, n_{w_3}, \ldots, n_{w_k}]$, where $n_{w_i}$ is the number of times word $w$ occurs in tweet $t$, is utilized to present the content of the tweet. Cosine similarity between two tweets of $t_1 = [x_1, y_1, c_1, p_1]$ and $t_2 = [x_2, y_2, c_2, p_2]$ measures the cosine angle between them using Equation 5.

$$TextSim(t,\ t') = \cos(\theta) = \frac{c_1 \cdot c_2}{||c_1||||c_2||} = \frac{\sum_{i=1}^{n} c_1^i c_2^i}{\sqrt{\sum_{i=1}^{n} {c_1^i}^2}\sqrt{\sum_{i=1}^{n} {c_2^i}^2}} \tag{5}$$

Geotagged tweets with Euclidian distance less than $\varepsilon_{\mathrm{e}}$ and text similarity higher than $\varepsilon_{\mathrm{t}}$ are considered as neighbours. If the number of neighbours for a tweet reaches to minPnts, a cluster label is assigned to the tweet and its neighbours; otherwise, the tweet is considered as noise. After extracting the clusters related to the first $\varepsilon_{\mathrm{e}}$, the next value of $\varepsilon_{\mathrm{e}}$ is opted to extract clusters from tweets with undefined labels or the ones defined as a noise in previews iteration. The procedure is repeated until all values of $\varepsilon_{\mathrm{e}}$ are participated in extracting clusters. By iterating through different values of $\varepsilon_{\mathrm{e}}$, clusters with various densities are extracted in decreasing order of density.

At the final step, the results of the proposed method are compared to those of Moran's I and VDBSCAN as standard methods for considering spatial autocorrelation and heterogeneity, respectively.

### *Quality measures*

Three internal evaluation criteria of the Davies–Bouldin index, Dunn index and Silhouette coefficient are used in this study to measure the quality of the clustering. Davies–Bouldin index calculates the ratio of inter-cluster distances to intra-cluster distances using Equation 6. The lower value of this index indicates better clustering (Davies and Bouldin 1979).

$$DB = \frac{1}{n} \sum\nolimits_{i=1}^{n} \max\left(\frac{\sigma_i + \sigma_j}{d\left(c_i, c_j\right)}\right) \tag{6}$$

where $c_i$ is the i$^{\text{th}}$ cluster center, $\sigma_i$ is the average distance of objects in cluster $i$ to the cluster center, $d\left(c_i, c_j\right)$ presents the distance between cluster $c_i$ and $c_j$. Moreover, $n$ denotes the number of clusters.

The ratio between minimum inter-cluster distances to maximum intra-cluster distances is calculated using the Dunn index based on Equation 7 (Dunn 1974). The algorithm which obtains a higher Dunn index has better performance.

$$D = \frac{\min d(i,j)}{\max d'(k)} \tag{7}$$

In Equation 7, *d(i,j)* indicates the inter-cluster distance of clusters *i* and *j* and *d′ (k)* is the distance between objects in cluster *k*.

The internal measure of the Silhouette coefficient (Rousseeuw 1987) shows how well an object is matched to its cluster. It ranges from − 1 to 1 where higher values prove appropriate clustering while low or negative values indicates a poor clustering.

$$S(i) = \frac{b(i) - a(i)}{Max\{a(i), b(i)\}} \tag{8}$$

Where *a*(*i*) is the mean intra-cluster distance of an object and *b*(*i*) is the distance between the object and the nearest cluster that the object does not belong to.

## Results

### *Spatial autocorrelation result*

To assess whether the clustering pattern exists in geotagged tweets and nearby tweets cover similar topics, spatial autocorrelation measurements were applied in this study. In order to choose an appropriate distance band for Global Moran's I, a variogram was used in this study, and the value of the range was set to 0.2 (Figure S2) as a cut-off distance for calculating Global Moran's I.

Applying the Moran's I Index resulted in four values of Moran's Index, expected value, z-score, and p-value. The positive value of Moran's Index and z-score of 289.39 rejected the null hypothesis that the tweets related to the hurricane were randomly distributed (Table 1).

**Table 1.** The output of Global Moran's Index for Dorian Tweets.

| Moran's Index | Expected Value | Z-Score | P-Value |
|---|---|---|---|
| 0.085912 | −0.000054 | 289.389523 | 0.000000 |

The global spatial autocorrelation tests proved that the pattern expressed by tweets related to hurricane Dorian was clustered. The next step was to extract the existing clusters from geotagged tweets.

### LDA results

In order to apply LDA, the number of topics should be determined in advance. Zhao *et al.* (2015) and Huang *et al.* (2017) demonstrated that perplexity is a promising method to choose the optimum number of topics for LDA by testing it on different datasets. According to the output results of perplexity (Figure S3), 40 was chosen as the best number of topics for LDA.

Having calculated the number of topics, LDA was applied to Dorian tweets to identify the existing topics and the probability of each tweet being related to each topic. This value was fed to the next step to calculate the density levels. Table 2 presents topics related to the hurricane and their frequent terms. As presented in this table, there were four extracted topics related to hurricane Dorian. The maximum probability value of hurricane-related topics was assigned to each tweet as *p* and fed to the next step for event detection from geotagged tweets.

### Incremental spatial autocorrelation

The probability of topics related to the hurricane extracted from LDA was utilized as the value of interest in assessing spatial autocorrelation. Z-score values and distances of peaks are presented in Table 3. The graph has three peaks associated with distances of 0.2, 0.7, and 1.2 (Figure S4). These peaks indicate the distances where the pronounced spatial pattern was clustering. These distance values were used as input parameters for the proposed algorithm. Passing the last peak, the value of z-score decreased, which means less significant clustering patterns.

### Clustering results and comparison

The extracted peak values of ISA (0.2, 0.7, and 1.2) were used as *epsilon* values ($\varepsilon_e$) in the proposed method. Also, the k-dist plot was used to calculate the values of *epsilon* in VDBSCAN. Based on the best practices from the literature, 0.5 was selected as $\varepsilon_t$ for text similarity threshold (Ozdikis *et al.* 2014, Chellal *et al.* 2017, Ghaemi and Farnaghi 2019). The

**Table 2.** Extracted topics related to the Dorian Hurricane.

| Topic Number | First 10 terms in each topic | Preferred topic heading |
|---|---|---|
| Topic 19 | Storm, tropic, condit, walk, report, saint, mile, finish, port, water | Storm |
| Topic 22 | hurrican, dorian, wait, stay, keep, safe, close, move, everyon, | Hurricane Dorian |
| Topic 31 | bahama, help, need, Island, people, support, pray, donate, prayer, pleas | Bahamas |
| Topic 34 | forecast, tstorm, sunni, cloud, cloudi, clear, part, chanc, sunday, tuesday | Forecast |

**Table 3.** The peaks of ISA and relative distances and Z-score.

| Peaks | Distance | Value |
|---|---|---|
| First Peak | 0.2 | 288.851342 |
| Second (Max) Peak | 0.7 | 362.438802 |
| Third Peak | 1.2 | 345.472537 |

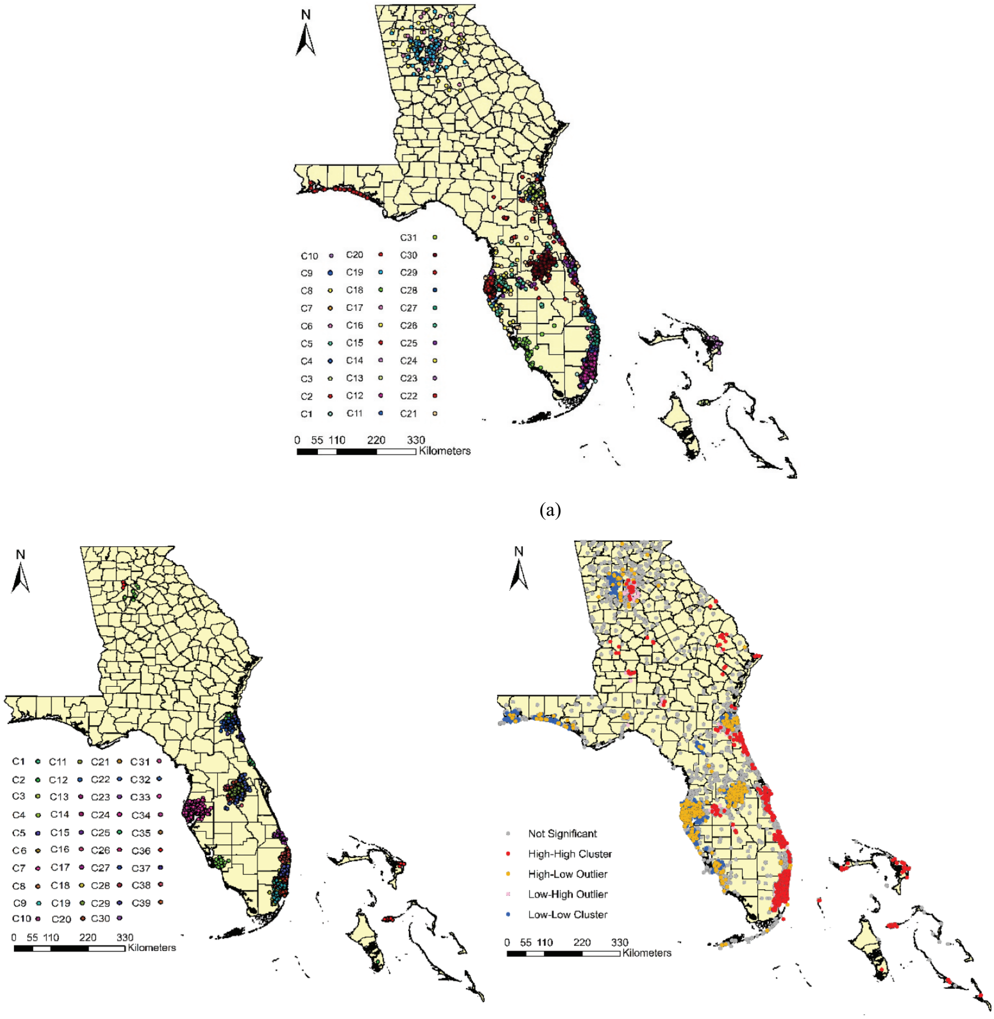
output clusters extracted by VDCAT and VDBSCAN are illustrated in Figure 3a and Figure 3b, respectively, where tweets of each cluster are represented by a particular colour.

Also, local Moran's Index was utilized as the measure of spatial autocorrelation to extract clusters related to hurricane Dorian. The output map is illustrated in Figure 3c. In this map, H.H. (High-High) displayed the hot spots, which means that tweets related to the hurricane were surrounded by other tweets discussing the cyclone. Therefore, H. H. indicated clusters related to hurricane Dorian. H.L. (High-Low) and L.H. (Low-High) means that tweets were dissimilar to those of their neighbors. In other words, H.L. refers to a tweet about the hurricane. Still, its neighbors are not discussing the storm, whereas L. H. refers to a tweet that is not about the hurricane, but it is surrounded by tweets discussing the hurricane. L.L. (Low-Low) cluster displayed the cold spots, which means that users did not share tweets about the hurricane. In this study, spatial clusters characterized by H.H. (indicated by red dots) were the most momentous ones as they revealed the spatial clusters related to hurricane Dorian.

Moreover, Figure 4a and Figure 4b illustrate all clusters which were related to the hurricane extracted by VDCAT and VDBSCAN, respectively. At the same time, Figure 4c depicts H.H. clusters extracted by Local Moran's I. As demonstrated in these figures, the most popular location of geotagged tweets related to the hurricane was in the Bahamas and around the east coast of Florida which was the path of hurricane Dorian. Visual comparison of the figures revealed that hurricane clusters extracted by VDCAT has a similar pattern with the local Moran's I results along with the east coast of Florida. Although the patterns are still different from each other for other areas, VDCAT clusters were more similar to those extracted by local Moran's I in comparison with hurricane clusters detected by VDBSCAN. Considering that Moran's I is the most popular method to tackle spatial autocorrelation, the similarity between VDCAT and Moran's results proved that the proposed method could deal with spatial autocorrelation in extracting clusters. Also, as illustrated in Figure 3a, VDCAT could extract other clusters that are not related to the hurricane. In contrast, Moran's I can only detect clusters of a particular topic (marked with H.H.), which is not desirable for event detection procedures. An example is demonstrated in Figure S5, where VDCAT was able to extract clusters related to a fire event. At the same time, Moran's I just indicated these points as L.L. (tweets which are not related to the hurricane).

### Quality measures results

In order to calculate the internal evaluation criteria of the Davies–Bouldin index, Dunn index, and Silhouette coefficient, some information such as distance to cluster centre and number of clusters were required, which were not available in Moran's I. Therefore, the


N
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10
C11 C12 C13 C14 C15 C16 C17 C18 C19 C20
C21 C22 C23 C24 C25 C26 C27 C28 C29 C30
C31
0 55 110 220 330 Kilometers
N
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10
C11 C12 C13 C14 C15 C16 C17 C18 C19 C20
C21 C22 C23 C24 C25 C26 C27 C28 C29 C30
C31 C32 C33 C34 C35 C36 C37 C38 C39
0 55 110 220 330 Kilometers
N
Not Significant
High-High Cluster
High-Low Outlier
Low-High Outlier
Low-Low Cluster
0 55 110 220 330 Kilometers

(a)

(b)

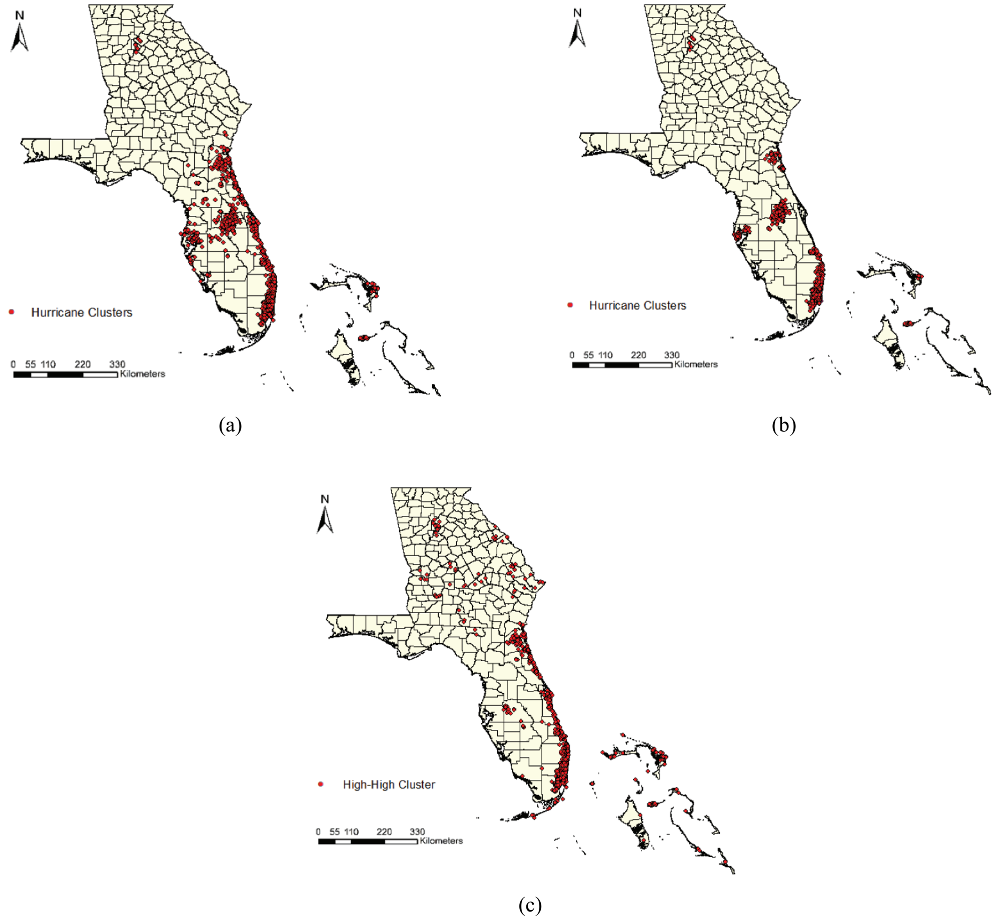
(c)

**Figure 3.** Extracted clusters by a) VDCAT and b) VDBSCAN and (c) local Moran's Index.

evaluation criteria were just calculated for VDCAT and VDBSCAN (Table 4). Lower Davies–Bouldin Index and higher Dunn index and Silhouette coefficient obtained from VDCAT indicated the superiority of the model over VDBSCAN in extracting clusters.

## Discussion

This section discusses extracted clusters by VDCAT and VDBSCAN and analyses the output clusters. Although there exist common clusters between the two algorithms, some extracted clusters are different in some areas. Some clusters are extracted by only one algorithm (VDCAT or VDBSCAN), and some clusters of VDBSCAN are merged into one by VDCAT. For instance, cluster C1 extracted by VDCAT is the combination of three clusters of C4, C7, and C34 of VDBSCAN. Investigating the word cloud of these clusters (Figure S6)



(a)

(b)

(c)

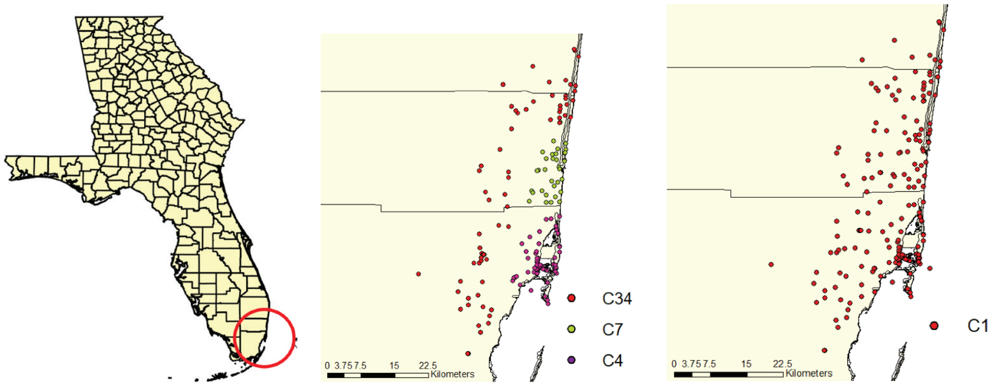
**Figure 4.** Extracted clusters related to the hurricane by a) VDCAT, b) VDBSCAN and C) local Moran's I.

**Table 4.** The Davies–Bouldin, the Dunn index and Silhouette coefficient obtained from VDCAT and VDBSCAN.

| Method | VDCAT | VDBSCAN |
|---|---|---|
| Davies–Bouldin Index | 5.1214296 | 5.4311547 |
| Dunn index | 0.03425498 | 0.02310275 |
| Silhouette Coefficient | 0.513243 | 0.452896 |

indicates that all three clusters by VDBSCAN contain almost the same set of words. It proves that they all refer to the same event, but unsuitable distance leads to the generation of separate clusters by VDBSCAN. The locations of these clusters are illustrated in Figure 5. Spatial proximity of points, in addition to a similar set of words, indicates that these points belong to the same group. As illustrated in Figure 5, these clusters were located east of Florida and the hurricane Dorian path, where there were dense tweets in the area.

Also, cluster C9 by VDCAT is separated into four clusters of C12, C17, C36, and C37 by VDBSCAN, while all clusters refer to one event, 'donation and help to Bahamas' with common frequent words. It seems that these clusters should be merged into one, but



(a) (b) (c)

**Figure 5.** A) the location of clusters in study area, b) C4, C7, and C34 by VDBSCAN and c) C1 by VDCAT.

they are separated into different clusters (Figures S7 and S8). The other clusters extracted by VDCAT and VDBSCAN, along with their frequent words, are presented in Tables S1 and S2.

Considering the extracted clusters, it can be concluded that in addition to the event, the name of the place where the event occurred can be extracted using frequent words in the word cloud. For example, in cluster C3 (Figure S9), the frequent words of this cluster depict that a tropical storm happened near Augustine Street and Flagler beach.

In addition to clusters related to hurricane Dorian, some other clusters have been extracted by the algorithms indicating an event or a location. In cluster C2, Endomondo and Orlando are the most frequent words. Endomondo refers to a cycling team from Florida's west coast, and Orlando is the place where the cycling event happened. Likewise, there are some other words including 'Cycling', 'Mile' and 'Finish' which help to better recognize the event (Figure S10 (a)). Cluster C10 (Figure S10 (b)) shows a musical event by 'Knotfest Roadshow' in 'Tampa' and cluster C19 (Figure S10 (c)) refers to a football game in Atlanta. Cluster C29 (Figure S10 (d)) depicts a fire cluster perhaps due to 'vehicle collision' in 'Ulmerton'. According to the frequent word cloud, it is clear that two clusters of C16 and C17 relate to topics of 'home' and 'work' respectively. Extracting such clusters assist decision-makers to keep pace with challenges in urban areas.

Also, in some areas, there are some overlapped clusters. It happens because users in the same place may talk about various topics, and therefore their tweets would be grouped into different clusters. Consider clusters C2, C5, and C7 as an example (Figure S11). Cluster C5 includes tweets related to 'hurricane Dorian', and it is placed in 'Orlando'. Cluster C7 consists of words related to 'Bahamas', 'relief' and 'help' which indicates that a group of users is trying to help victims in the Bahamas.

Moreover, clusters C6, C9, and C12 placed in the same location but refer to different events (Figure S12). C6 includes tweets related to 'hurricane', C9 consists of tweets related to 'donation' and 'help' to 'Bahamas' and C12 includes tweets related to 'weather' and 'forecast'. It proves that the algorithm is able to efficiently separate the events which occur in the same place but refer to various topics. It confirms the effective role of taking text similarity into account in the extraction of various events.

As mentioned before, the name of places in tweets can be helpful in recognizing the name of the place where users share tweets. However, this is not always the case. Sometimes users in a location share posts related to an event that happened in another place. For example, in clusters C25, and C27 (Figure S13), users talk about Bahamians who need help, but the shared tweets are in another place: Florida. Therefore, it is not possible to detect the name of the place where tweets are shared by only considering the frequent name of the place in tweets. Therefore, the name of places in shared tweets can be confusing, and they should be used with caution.

This study also suffers from some limitations. In this study, our focus was dealing with spatial autocorrelation and heterogeneity in event detection procedure and we only considered geolocated tweets and disregarded tweets that were not geo-tagged to reduce the complexity. Using geoparsers to extract the location of tweets could increase the number of tweets and enriches the dataset. Our study used LDA to extract topics from tweets, which requires the number of topics in advance. However, determining the proper number of topics, considering the real-time nature of the analysis, is challenging. Finally, considering the time component could provide the possibility of monitoring the current state of an event (Farnaghi *et al*. 2020).

## Conclusions

A new method, named VDCAT, was proposed in this study for event detection from geotagged tweets during hurricane Dorian. It was designed to consider both spatial autocorrelation and heterogeneity in extracting clusters. Moreover, text similarity was considered as another dimension in addition to spatial proximity to extract spatial-textual clusters. The output clusters were compared to those of local Moran's I and VDBSCAN. Quantitative and visual comparison between methods proved the superiority of VDCAT in revealing clusters from geotagged tweets. Comparing the extracted hurricane clusters of VDCAT with local Moran's I confirmed that the proposed method was able to consider spatial autocorrelation in cluster extraction as well as local Moran's I. Additionally, the proposed method was capable of extracting clusters with various densities. Meanwhile, taking spatial autocorrelation into account led to better clustering results in comparison with VDBSCAN. Analysing the output clusters clarified that VDCAT was able to efficiently extract clusters related to hurricane Dorian in the Bahamas and east coast of the U.S. Besides, the method extracted clusters related to other events, for example, fire, football game, and cycling. Furthermore, geotagged tweets in approximate locations, which contained various sets of words, were appropriately divided into separate clusters. Last but not least, the proposed method can provide the authorities and decision-makers with proper information to set efficient measurements for the resilience of urban systems. The governments can detect damage-prone areas and recognize the demand for public services to allocate required resources for recovery and hazard mitigation.

Future work should focus on applying new methods such as the Hierarchical Dirichlet Process (HDP) instead of LDA (Teh *et al*. 2006, Srijith *et al*. 2017) since it does not require prior knowledge about the number of topics. Additionally, improving the proposed solution to consider the semantic similarity among the tweets (Khatua *et al*. 2019) will be a field of future investigation. Utilizing geoparser (Alex *et al*. 2016) to extract the location of tweets with no geographical location in the event detection procedure can increase the number of georeferenced tweets and result in more accurate outputs.

Extending the method to consider both space and time dynamically to detect events, as proposed by (Farnaghi *et al*. 2020) is also another future work. In this context, considering both spatial and temporal autocorrelation in the existence of heterogeneity will be a challenging issue that needs thorough investigation.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Zeinab Ghaemi http://orcid.org/0000-0002-3071-5563

## References

Abdelhaq, H., Sengstock, C., and Gertz, M., 2013. Eventweet: online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6 (12), 1326–1329. doi:10.14778/2536274.2536307

Aggarwal, C.C., 2014. *Data classification: algorithms and applications*. Boca Raton : CRC press, Taylor & Francis.

Alex, B., *et al*. 2016, May. Homing in on twitter users: evaluating an enhanced geoparser for user profile locations. *In*: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), 23-28 May 2016 Portorož, Slovenia, 3936–3944.

Anselin, L., 1995. Local indicators of spatial association—LISA. *Geographical Analysis*, 27 (2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Bakillah, M., Li, R.-Y., and Liang, S.H., 2015. Geo-located community detection in twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon haiyan. *International Journal of Geographical Information Science*, 29 (2), 258–279. doi:10.1080/13658816.2014.964247

Balducci, F. and Ferrara, A., 2018. Using urban environmental policy data to understand the domains of smartness: an analysis of spatial autocorrelation for all the Italian chief towns. *Ecological Indicators*, 89, 386–396. doi:10.1016/j.ecolind.2017.12.064

Basu, M., Bandyopadhyay, S., and Ghosh, S., 2016. Post disaster situation awareness and decision support through interactive crowdsourcing. *Procedia Engineering*, 159, 167–173. doi:10.1016/j.proeng.2016.08.151

Besag, J. and Newell, J., 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154 (1), 143–155. doi:10.2307/2982708

Blank, G., 2017. The digital divide among twitter users and its implications for social research. *Social Science Computer Review*, 35 (6), 679–697. doi:10.1177/0894439316671698

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent dirichlet allocation. The Journal of machine Learning research, 3, 993-1022.,

Bohling, G., 2005. Introduction to geostatistics and variogram analysis. *Kansas Geological Survey*, 1, 1–20.

Brunsdon, C., Fotheringham, S., and Charlton, M., 1998. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47 (3), 431–443. doi:10.1111/1467-9884.00145

Chellal, A., Boughanem, M., and Dousset, B., 2017. Word similarity based model for tweet stream prospective notification. *In*: European conference on information retrieval. Springer, 8-13 April Aberdeen, Scotland, UK, 655–661.

Cheng, T., Wicks, T., and Bejon, P., 2014. Event detection using twitter: a spatio-temporal approach. *PloS One*, 9 (6), e97807. doi:10.1371/journal.pone.0097807

Choudhary, J., Ohri, A., and Kumar, B., 2015. Spatial and statistical analysis of road accidents hot spots using GIS. *In*: Proceedings of the 3rd Conference of transportation research group of India (3rd CTRG), Kolkata, India, 17–20.

Cressie, N. and Hawkins, D.M., 1980. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, 12 (2), 115–125. doi:10.1007/BF01035243

Davies, D.L. and Bouldin, D.W., 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1 (2), 224–227. doi:10.1109/TPAMI.1979.4766909

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4 (1), 95–104. doi:10.1080/01969727408546059

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. , 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: *Kdd,* Portland, Oregon, USA, (Vol. 96, pp. 226–231, Vol. 34).

Farnaghi, M., Ghaemi, Z., and Mansourian, A., 2020. Dynamic spatio-temporal tweet mining for event detection: a case study of hurricane florence. *International Journal of Disaster Risk Science*, 11 (3), 378–393. doi:10.1007/s13753-020-00280-z

Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C., and Huang, J., 2015. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. *In*: 2015 IEEE 31st international conference on data engineering. IEEE, 13-17 April 2015 Seoul, Korea (South), 1561–1572.

Gall, M., *et al*., 2011. The unsustainable trend of natural hazard losses in the United States. *Sustainability*, 3 (11), 2157–2181. doi:10.3390/su3112157

Ghaemi, Z. and Farnaghi, M., 2019. A varied density-based clustering approach for event detection from heterogeneous twitter data. *ISPRS International Journal of Geo-Information*, 8 (2), 82. doi:10.3390/ijgi8020082

Ghodousi, M., Alesheikh, A.A., and Saeidian, B., 2016. Analyzing public participant data to evaluate citizen satisfaction and to prioritize their needs via K-means, FCM and ICA. *Cities*, 55, 70–81. doi:10.1016/j.cities.2016.03.015

Ghodousi, M., *et al*., 2019. Evaluating citizen satisfaction and prioritizing their needs based on citizens' complaint data. *Sustainability*, 11 (17), 4595. doi:10.3390/su11174595

Giebultowicz, S., *et al*., 2011. A comparison of spatial and social clustering of cholera in Matlab, Bangladesh. *Health & Place*, 17 (2), 490–497. doi:10.1016/j.healthplace.2010.12.004

Glotsos, D., Tohka, J., Soukka, J., and Ruotsalainen, U., 2004. A new approach to robust clustering by density estimation in an autocorrelation derived feature space. *In*: Proceedings of the 6th nordic signal processing symposium, 9-11 June 2004 Espoo, Finland. NORSIG 2004. IEEE, 296–299.

Gonick, S. and Errett, N., 2018. Integrating climate change into hazard mitigation planning: a survey of state hazard mitigation officers. *Sustainability*, 10 (11), 4150. doi:10.3390/su10114150

Huang, L., Jinyu, M., and Chen, C., 2017. Topic detection from microblogs using T-LDA and perplexity." *In*: 2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW). IEEE, 4 Dec. 2017, Nanjing, Jiangsu, China, 71–77.

Jahani, S. and Bagherpour, M., 2011. A clustering algorithm for mobile ad hoc networks based on spatial auto-correlation. *In*: 2011 International symposium on computer networks and distributed systems (CNDS). IEEE, 23-24 Feb. 2011 Tehran, Iran, 136–141.

Kankanamge, N., *et al*., 2019. Determining disaster severity through social media analysis: testing the methodology with South East Queensland Flood tweets. *International Journal of Disaster Risk Reduction*, 42, 101360.

Karmegam, D. and Mappillairaju, B., 2020. Spatio-temporal distribution of negative emotions on twitter during floods in Chennai, India, in 2015: a post hoc analysis. *International Journal of Health Geographics*, 19 (1), 1–13. doi:10.1186/s12942-020-00214-4

Khatua, A., Khatua, A., and Cambria, E., 2019. A tale of two epidemics: contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management*, 56 (1), 247–257. doi:10.1016/j.ipm.2018.10.010

Kisilevich, S., Mansmann, F., Nanni, M., and Rinzivillo, S., 2009. Spatio-temporal clustering. *In*: *Data mining and knowledge discovery handbook*. Springer, Boston, MA., 855–874.

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. *Diversity & Distributions*, 13 (1), 66–69.

Lachlan, K.A., Spence, P.R., and Lin, X., 2014. Expressions of risk awareness and concern through twitter: on the utility of using the medium as an indication of audience needs. *Computers in Human Behavior*, 35, 554–559. doi:10.1016/j.chb.2014.02.029

Liu, P., Zhou, D., and Wu, N., 2007. VDBSCAN: varied density based spatial clustering of applications with noise. *In*: 2007 international conference on service systems and service management. IEEE, 9-11 June 2007 Chengdu, China,1–4.

Lu, P., *et al.*, 2019. Landslides detection through optimized hot spot analysis on persistent scatterers and distributed scatterers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156, 147–159. doi:10.1016/j.isprsjprs.2019.08.004

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., and Rosenquist, J., 2011. Understanding the demographics of twitter users. *In*: Fifth international AAAI conference on weblogs and social media, 17-21 July 2011 Barcelona, Catalonia, Spain.

Mohammadinia, A., Alimohammadi, A., and Saeidian, B., 2017. Efficiency of geographically weighted regression in modeling human leptospirosis based on environmental factors in Gilan Province, Iran. *Geosciences*, 7 (4), 136. doi:10.3390/geosciences7040136

Moran, P.A., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1–2), 17–23. doi:10.1093/biomet/37.1-2.17

Morchid, M., *et al.*, 2015. An author-topic based approach to cluster tweets and mine their location. *Procedia Environmental Sciences*, 27, 26–29. doi:10.1016/j.proenv.2015.07.109

Nolasco, D. and Oliveira, J., 2019. Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, 93, 290–303. doi:10.1016/j.future.2018.09.008

Ozdikis, O., Senkul, P., and Oguztuzun, H., 2014. Context based semantic relations in tweets. State of the art applications of social network analysis. Cham: Springer, 35–52.

Qi, W. and John E., T., 2014. Quantifying, comparing human mobility perturbation during hurricane sandy, typhoon wipha, typhoon haiyan. *Procedia Economics and Finance*, 18, 33–38. doi:10.1016/S2212-5671(14)00910-1

Roth, D., *et al.*, 2016. Identification of spatial and cohort clustering of tuberculosis using surveillance data from British Columbia, Canada, 1990–2013. *Social Science & Medicine*, 168, 214–222. doi:10.1016/j.socscimed.2016.06.047

Röthlisberger, V., Zischg, A.P., and Keiler, M., 2017. Identifying spatial clusters of flood exposure to support decision making in risk management. *Science of the Total Environment*, 598, 593–603. doi:10.1016/j.scitotenv.2017.03.216

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Saeidian, B., *et al.*, 2018. Optimized location-allocation of earthquake relief centers using PSO and ACO, complemented by GIS, clustering, and TOPSIS. *ISPRS International Journal of Geo-Information*, 7 (8), 292. doi:10.3390/ijgi7080292

Sloan, L., *et al.*, 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS One*, 10 (3), e0115545. doi:10.1371/journal.pone.0115545

Smiley, M.J., *et al.*, 2010. A spatial analysis of health-related resources in three diverse metropolitan areas. *Health & Place*, 16 (5), 885–892. doi:10.1016/j.healthplace.2010.04.014

Souza, R.C., *et al.*, 2019. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. *Spatial and Spatio-temporal Epidemiology*, 29, 163–175. doi:10.1016/j.sste.2018.11.005

Srijith, P.K., *et al.*, 2017. Sub-story detection in twitter with hierarchical dirichlet processes. *Information Processing & Management*, 53 (4), 989–1003. doi:10.1016/j.ipm.2016.10.004

Steiger, E., *et al.*, 2015. Twitter as an indicator for whereabouts of people? Correlating twitter with U. K. census data. *Computers, Environment and Urban Systems*, 54, 255–265. doi:10.1016/j.compenvurbsys.2015.09.007

Stojanova, D., Ceci, M., Appice, A., Malerba, D., and Džeroski, S., 2011. Global and local spatial autocorrelation in predictive clustering trees. *In*: *International conference on discovery science*. Springer,  5-7 October 2011 Espoo, Finland, 307–322.

Stojanova, D., *et al*., 2013. Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecological Informatics*, 13, 22–39. doi:10.1016/j.ecoinf.2012.10.006

Suthar, N., Indr, P., and Vinit, P., 2013. A technical survey on DBSCAN clustering algorithm. *International Journal of Scientific and Engineering Research*, 4, 1775–1781.

Teh, Y.W., *et al*., 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101 (476), 1566–1581. doi:10.1198/016214506000000302

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46 (sup1), 234–240. doi:10.2307/143141

Wang, T.-C. and Yue, C.-S.J., 2013. Spatial clusters in a global-dependence model. *Spatial and Spatio-temporal Epidemiology*, 5, 39–50. doi:10.1016/j.sste.2013.03.003

Webster, P.J., *et al*., 2005. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309 (5742), 1844–1846. doi:10.1126/science.1116448

Wei, H. (2020). Local News and Event Detection in Twitter (Doctoral dissertation).

Wu, Y., *et al*., 2017. A dynamic spatial clustering for emergency response based on hierarchical-partition model. *Procedia Computer Science*, 111, 485–492. doi:10.1016/j.procs.2017.06.051

Yabe, T. and Ukkusuri, S.V., 2019. Integrating information from heterogeneous networks on social media to predict post-disaster returning behavior. *Journal of Computational Science*, 32, 12–20. doi:10.1016/j.jocs.2019.02.002

Yang, W. and Mu, L., 2015. GIS analysis of depression among twitter users. *Applied Geography*, 60, 217–223. doi:10.1016/j.apgeog.2014.10.016

Zhang, T. and Lin, G., 2016. On Moran's I coefficient under heterogeneity. *Computational Statistics & Data Analysis*, 95, 83–94. doi:10.1016/j.csda.2015.09.010

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W., 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC bioinformatics. BioMed Central, (Vol. 16, pp. S8, Vol. 13), https://doi.org/10.1186/1471-2105-16-S13-S8.