Case Studies on Transport Policy 12 (2023) 100992

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

Analyzing changes in travel patterns due to Covid-19 using Twitter data in India

Swapnil Shende, Eeshan Bhaduri *, Arkopal Kishore Goswami

Ranbir and Chitra Gupta School of Infrastructure Design and Management, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

ARTICLE INFO

Keywords:
Location based social media (LBSM)
Travel pattern
Machine learning
Classification
Clustering
COVID-19

ABSTRACT

Advancements in technology have enabled researchers to gather large-scale mobility information cost-effectively. In fact, with millions of active users, location-based social media (LBSM) platforms such as Facebook, Twitter, Instagram, Flickr, etc., have become potential big data sources to measure individual behaviour. Despite such passive data collection techniques primarily not providing individual-level information, the sheer volume of such data facilitates a better understanding of aggregate patterns. Besides, conducting the conventional transportation survey during the initial waves of the COVID-19 pandemic was nearly impossible due to social- distancing and lockdown rules. In such context, the present research showcases a method for extracting the mobility traces and identifying the travel patterns of visitors and residents in Delhi from geo-labelled posts on Twitter.

Initially, a heuristic classification strategy has been developed based on a few spatiotemporal assumptions to identify and differentiate visitors and residents based on user coordinates. Also, three supervised machine learning techniques, i.e., support vector machine (SVM), k-nearest neighbours (kNN) and decision tree, were used to classify users based on their historical coordinates. Afterwards, the spatial variation of their destination preferences was studied using K-Means, DBSCAN, and Means-Shift clustering techniques, out of which the K-Means clustering method performed best. Lastly, the travel patterns from tweets during pre and during pandemic (COVID-19) were compared using respective clusters.

We observed that the performance of the proposed heuristic classifier is comparable with the supervised machine learning (ML) technique used for classification. Furthermore, the results indicate that the proposed model can successfully identify the cluster coordinates for visitors' spots as well as the locations of residents. During the pandemic situation, the mean distance travelled by users is significantly reduced. The study also shows that the number of long-distance trips has also decreased. Also, during COVID, tweets were done from very few unique tourist spots. This suggests lower tendencies of people to travel for tourism purposes. The proposed methods for classification and clustering in the present study will be crucial to obtain individual travel patterns from LBSM data.

1. Introduction

Recognition of travel patterns is essential for policymakers and transport planners to develop, assess, and select longer-term travel demand for a city where household surveys act as an established data source. Variations of these surveys, such as complementing them with phone, mail or the internet, have become prevalent in transportation research (Arentze et al., 2005). However, implementing these methods is expensive and consumes a lot of time, especially for cities with a large population. On top of it, most of these cities experience a high number of floating populations or visitors whose activities largely remain uncaptured in traditional travel surveys (Hasnat and Hasan, 2018). Nevertheless, these extra operations are superimposed on the city's service network, which may not have been originally built to support them. So, it becomes essential to classify users into visitors and residents and attempt to understand the heterogeneity in individual travel patterns.

India has scaled the second-highest number of social media users in the world (UNDP), with the majority being young. Interestingly, the ever-growing number of engaged users on location-based social media (LBSM) sites such as Instagram, Twitter, and Facebook provide an

* Corresponding author.
E-mail addresses: swapnilshende31@gmail.com (S. Shende), eeshanbhaduri@iitkgp.ac.in (E. Bhaduri), akgoswami@infra.iitkgp.ac.in (A.K. Goswami).

https://doi.org/10.1016/j.cstp.2023.100992
Received 25 April 2022; Received in revised form 9 February 2023; Accepted 13 March 2023
Available online 16 March 2023

excellent opportunity to gain insights into individual travel behaviour. However, the underlying issue is that data from most social media is not public and does not include relevant information for transport research. Twitter has been an exception as tweets are available through simple APIs provided by the organisation and include different types of anonymous information for each Tweet.

The present study has two broad objectives: (1) to develop a methodological framework to extract and analyze geolocated Twitter data; and (2) to illustrate the utility of Twitter data in determining travel behaviour change. However, this comes with the caveat that a very small percentage (approximately 1%) of the available Twitter feeds can be freely downloaded via the API, and an even smaller portion of it is geotagged. This is because this geotag service is optional rather than default. Overall, we attempt to stress the methodological novelty while indicating the immense possibility of utilising the Twitter dataset, provided it reaches the requisite representativeness for the whole population. In an attempt to do that, the current study segregates the travel patterns of visitors and residents using Twitter data. Also, it establishes a framework for individual-level travel pattern analyzes. The first step, i.e., *classification* of users into residents and visitors within the investigative area, is done using heuristic as well as ML-supervised classifiers whereas the latter part, i.e., *clustering* of travel destinations, was identified with open spatial clustering methods to identify the most commonly visited sites. Subsequently, we also analyzed the change in their travel behaviour pre and during COVID to assess the performance of the proposed classification and clustering techniques. From the results of this research, a fair amount of insights has been obtained regarding visitors' and residents' travel destination preferences.

2. Literature review

The primary motivation for the current research is to find the methods and implications used in the field of LBSM data for the analysis and planning in the area of transportation research. This section presents insights into the usability of social media datasets and relevant analysis methods.

2.1. Social media for travel detection and analysis

Gao et al. (2013) used location-based social media to analyze to identify users' social behaviour using check-in data. In another approach, web-based social media was used to detect traffic incidents, whereas Fu et al. (2015) used tweets containing predefined keywords related to traffic incidents. The research showed that tweets are useful for detecting an incident as early as possible and can be used as an additional data source for the administrative body. Mai & Hranac (2013) used a similar approach by comparing reported events of the California Highway Patrol with associated tweets by envisioning the frequency of tweets within the location under consideration. Steur (2015) used a comparative method of dealing with traffic incidents in the Netherlands. Hasan et al. (2013) used check-in and location-based social media data to extract activity and travel patterns. They combined the geotagged tweets data and check-in data to extract activity patterns. Hasan & Ukkusuri (2014) utilised a similar dataset to predict individual travel patterns.

Analysing visitor behaviour is more recent, and the research is dominated mainly by the Flickr dataset. Girardin et al. (2008) combined data from mobile networks and geotagged posts from Flickr to understand visitors' travel behaviour. Xiang & Gretzel (2010) examined the travel planning related keywords searched on major search engines to predict visitor activity. Pozdnoukhov & Kaiser (2011) combined the check-in service, Foursquare, and the social media service, Twitter, to study spatial and temporal variations of topics. The method enforced both LDA-based linguistic and geotagged spatial variations. Ichimura & Kamada (2012) set up an Android framework to gather visitors' spot information whenever a Tweet is made through the app. Popescu & Grefenstette (2011) utilised historical Flickr geotagged users' photos to create a framework for a recommendation system for visitors based on historical data. Majid et al. (2013) proposed a method to predict visitors visiting locations preference in a new city based on historical movement images posted via web-based media by sharing geo-labelled photographs on Flickr. Sun et al. (2013) used geo-labelled Flickr images to study the travel pattern of visitors in Vienna, Austria, using spatial scan statistics and kernel density function. In this research, posts containing 'Vienna' keyword were scanned, and almost 245,000 geo-labelled images were collected and analyzed. De Choudhury et al. (2010) established a technique to build travel itineraries using geo-labelled Flicker images. A two-step approach created an intra-city itinerary for Barcelona, London, New York, Paris, and San Francisco.

2.2. User classification

It is important to classify social media users to understand their travel behaviour better. For this research, users were categorised into visitors and residents based on predefined criteria. There is literature that has classified user profiles according to the research needs. A simple heuristic technique was used to differentiate home and work areas using geo-labelled tweets of a person (McNeill et al., 2017). Geo-labelled tweets were used (Abbasi et al., 2015) to distinguish the most dynamic visitors in the city of Sydney who visited various sites during the data assortment period. Users with at least nine new geo-labelled tweets, and if they are in a minimum of one (or two) cycles of the data assortment period, were considered dynamic visitors. To differentiate local citizens and sightseers in Barcelona (Manca et al., 2017) proposed a heuristic classifier that defined the duration of a user within the city boundary based on the 'user location' as provided by Twitter. A timeframe of twenty days was used for the classification of users. Andrienko et al. (2013) developed a method where residents were defined as users who spent at least ten days inside and no more than eight days outside the greater Seattle region during a two-month data collection period. The time and place stamps of tweets have been commonly used, as seen from the literature, even though the central aspect of these inquiries used only a single feature to separate travellers and people.

2.3. Spatial clustering

Even though clustering techniques are widely adopted in geospatial research, few have used them to detect travel patterns. Hasnat and Hasan (2018) utilised clustering techniques to find out the point of interest of visitors and residents in Florida. They found K-Means to be performing best among other clustering techniques. Abbasi et al. (2015) classified visitors as people who flew to and from Sydney within around a month. The local inhabitants and vacationers may remember the most visited locations by examining geo-labelled tweets. Using geo-labelled tweets, (Lee et al., 2016) showed changes in the movement spaces of 116 Twitter users and determined their major areas of movement. In literary works, various kinds of spatial clustering procedures have been utilised to discover the destinations that are closely related and points of interest. The most frequently used clustering techniques are, K-Means (Kanungo et al., 2002), Ward's method (Ward, 1963), and DBSCAN (Ester et al., 1996). Majid et al. (2013) utilised DBSCAN on geo-labelled photographs to recognise traveller areas of attractions.

2.4. Twitter dataset for COVID travel pattern analysis

The COVID-19 pandemic has disrupted mobility in numerous ways, primarily by reducing human interaction (De Vos, 2020). Globally, public authorities imposed various restrictions, including lockdowns, nighttime curfews, and school closures, among others, to curb the spread of COVID-19 (Das et al., 2021; Zannat et al., 2021). Understandably, the daily activity patterns have been significantly affected, resulting in, among other things, a decrease in commuting trips (Tirachini and Cats,

2020; Guzman et al., 2021) as well as changes in modal preferences (Bhaduri et al., 2020; Bhattacharyya et al., 2021). In general, it has been shown that pre-COVID and during-COVID periods differ significantly in terms of travel behaviour (Bucsky, 2020; Barbieri et al., 2021). For example, public transportation is observed to be one of the most impacted modes (Coppola and Fabiis, 2020; Coppola and De Fabiis, 2021). At the same time, other studies highlighted an increase in work-from-home (WFH) employment (Beck et al., 2020; Bhaduri et al., 2020) during the pandemic.

However, collecting mobility datasets in times of COVID-19 using conventional methods is nearly impossible due to difficulties in reaching out to people. Due to persistent lockdowns implemented, travel patterns have significantly changed worldwide. Buckee et al. (2020) showed how the aggregation of large datasets could help refine interventions by providing near real-time information about changes in human movement patterns. Huang et al. (2020) proposed the single-day distance and the cross-day distance which emphasises changes in distance between two consecutive days to understand the shift in mobility patterns. The research also proposed a mobility-based responsive index (MRI) that captures the overall degree of mobility changes within a time window. The study suggested that movement patterns derived from Twitter data may be used to quantitatively depict COVID-19 using the mobility-based responsive index pandemic mobility dynamics by comparing patterns of increased mobility and reduced mobility in different geographic regions.

3. Data collection and filtering

Twitter is used as the primary source of data in this study. The benefits of using Twitter data are that it is simple and free to use the platform and provides a substantially large dataset consisting of tweets from millions of users. However, this data source often comes with inherent pitfalls due to the tweet's massive volume of needless content, making the compilation and cleaning of the tweet a rather crucial phase. To capture real-time Twitter streams and past tweets, Twitter offers free APIs like Streaming API and REST API[1]. Twitter Streaming API allows the developer real-time access to public Tweets from the platform. This is ideally useful to programmatically retrieve and analyze the Twitter dataset. In the present study, we employed Streaming API to identify the active Twitter users during the data collection period within the boundary of Delhi and subsequently filter the users having active geolocation.

On the other hand, Twitter REST API aids in searching terms or obtaining tweets based on specific parameters instead of delivering real-time data. Expectedly, such API is more beneficial for analytics on historical data. In our case, REST API was used to get historical tweets for all the users who were earlier identified through Streaming API. The major steps of the data collection process (using Twitter API v2) have been mentioned below[2]:

Step 1: Defining the objective of the study- It includes defining the topic of research (say, sentiment analysis of election propaganda) and kind of data (say, live feed).

Step 2: Processing of the required data- This involves a set of operators and building a rule to obtain a specific type of data.

Step 3: Connecting and authenticating the appropriate endpoint- It requires connecting to the streaming endpoint to obtain data utilising bearer token and developer portal.

Step 4: Mitigating errors and disconnections- This handles the repercussions of voluntary or involuntary disconnections.

Step 5: Obtaining final results (including basic metrics)- The final step involves developing a basic display of the obtained dataset.

The real-time data collection effort (*using Twitter Streaming API*) included four distinct phases (See Table 1) from end-September to mid-October 2020: (1) Phase 1 for seven days (21 September to 27 September); (2) Phase 2 for eight days (28 September to 05 October); (3) Phase 3 for seven days (06 October to 12 October); and (4) Phase 4 for eight days (13 October to 20 October). In total, we collected public Tweets emanating from the pre-specified boundary of Delhi for thirty days (one month). At the same time, we collected 3200 historical tweets (*using Twitter REST API*) for each of the identified users. This corresponds to the fact that both the pre-COVID and during-COVID Tweets are historical Tweets. The pre-COVID part includes all the tweets done before 25 March 2020 (lockdown started in India), while the during-COVID Tweets are obtained during lockdown phases until unlock phase 1.0 was initiated in India, i.e., 01 June 2020. For all the phases, the geographical location had been the Delhi city area, identified by coordinates 77.15 E, 28.45 N (lower-left corner), and 77.30 E, 29 N (upper-right corner).

We plotted the activity taking into account all of the tweet posts for the data obtained via RESTAPI (See Table 2). Both user groups (residents and tourists) exhibit a similar pattern in their daily activity, which peaks at the end of the day. Figures (See Figs. 1-2) show that the overall number of postings for both residents and tourists is relatively similar. This part aids in recognizing the pattern of activities.

It demonstrates that tourists are less active from 12 am to 9 am, and that activity gradually increases until 12 pm. For residents, the least active times are from 12 am to 8 am, and activity picks up a little earlier in the morning, starting at 8 am. The continual activity updates in Figure make this shift very plain to see. Residents are busiest about 9p. m., whereas tourists are busiest in the late afternoon.

However, the tweets are not geotagged were subsequently removed as they were not significant for this research. We could observe substantial variation in the number of Tweets per user (average = 1170) as the minimum number is four, whereas the maximum value reaches 3250. Also, careful investigation regarding time-gap between consecutive Tweets by a particular user (average = 5.23 days) suggests a minimum and maximum value of 1.36 min and 81 days respectively. Moreover, we have included Figs. 3-4 to describe the temporal heterogeneity (within a 24-hour window) of Twitter users (separately for residents and tourists), which has also been used as part of the heuristics analysis.

4. Classifications

4.1. Heuristic classifier

For classifying whether or not the user is from Delhi, a basic heuristic solution is developed based on the premise that people are expected to tweet from their residing place/home during the night. In this process, users are referred to as *residents* with a higher number of their tweets having geolocation within the Delhi bounding box (77.15 E, 28.45 N (lower-left corner), and 77.30 E, 29 N (upper right corner)) and all the others as *visitors*. The study makes a heuristic assumption that a user usually posts more tweets from their home location during the night. Therefore, we chose a six-hour window, i.e., from 12 a.m. to 6 a.m.,

Table 1
Tweet Volume for Delhi over a period of one month.

Phase	1	2	3	4
Total Tweets (in thousand)	102.555	199.047	70.172	128.617
Tweetswithprecise coordinates	2817	4860	2219	4250
Number of unique users	964	1558	733	1294
Number of users with more than one tweet	338	582	235	467

[1] For details please refer to the Twitter tutorials: https://developer.Twitter.com/en/docs/tutorials/consuming-streaming-data and https://developer.Twitter.com/en/docs.

[2] For more details, please refer: https://developer.Twitter.com/en/docs/tutorials/stream-tweets-in-real-time.

Table 2
Tweet Volume for REST API.

	Total Residents	Tourists	Pre-COVID Residents	Tourists	During-COVID Residents	Tourists
Total (in thousand)	725.470	706.306	546.325	542.942	179.145	163.364
Geotagged (in thousand)	198.746	226.199	145.721	157.733	53.025	68.466

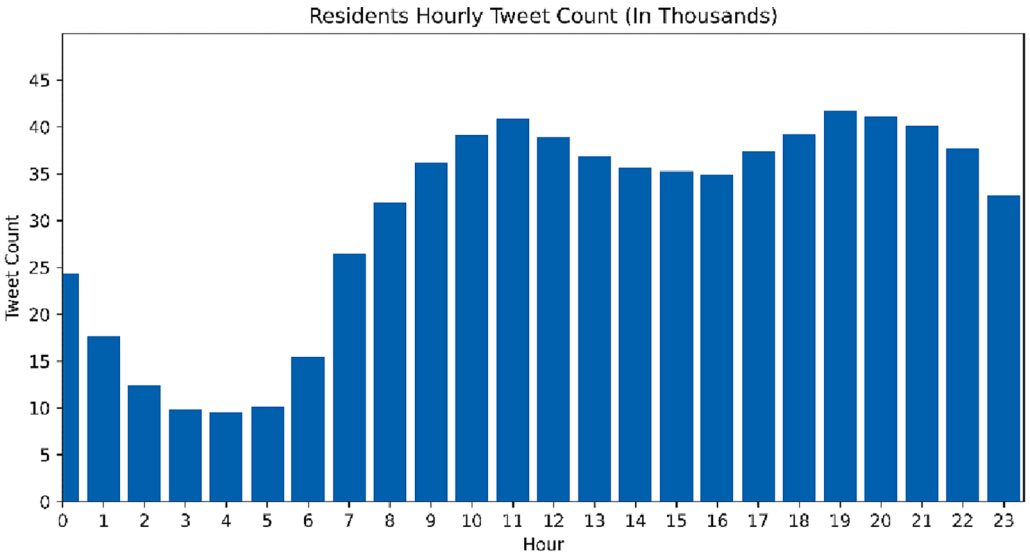

Fig. 1. Description of the REST API Tweet corpus (for residents).

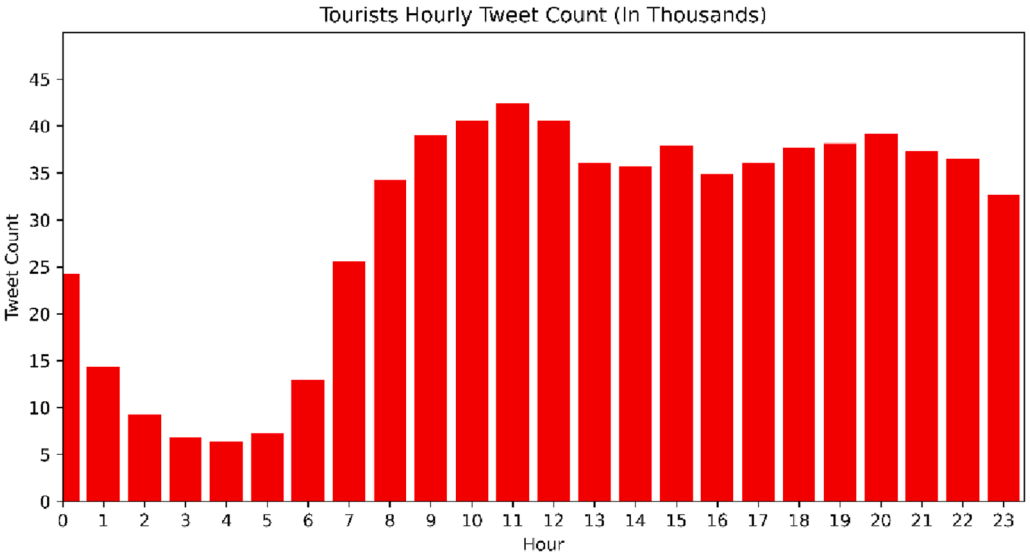

Fig. 2. Description of the REST API Tweet corpus (for tourists).

measured the number of tweet activities carried out during that phase, and compared it to the other phase (from 6 a.m. to midnight). The findings of this heuristic are checked using the ground reality information obtained in their Twitter profiles from users' posted paces. The heuristic classifier gave a precision of 72% when compared with ground truth values, i.e., user-defined location on users' profiles.

4.2. Supervised Machine learning

The primary reason for conducting this analysis is to compare how the proposed heuristic classifier is proposed. M. M. Hasnat & Hasan (2018) suggested supervised machine learning techniques to classify users into visitors and residents using some parameters included in this analysis. The study uses three supervised classification techniques to assess the accuracy of the proposed heuristic classifier: K-Nearest Neighbors (KNN) (Vechtomova, 2009), Support Vector Machine (SVM) (Cristianini & Shawe-Taylor, 2000), and Decision Tree (Safavian & Landgrebe, 1991)).

Three performance parameters are derived for each user:

(F1) mean distance between successive coordinates;

(F2) standard deviation of the distance between consecutive coordinates;

(F3) ratio of the number of tweets originating inside to the outside of the study area boundary.

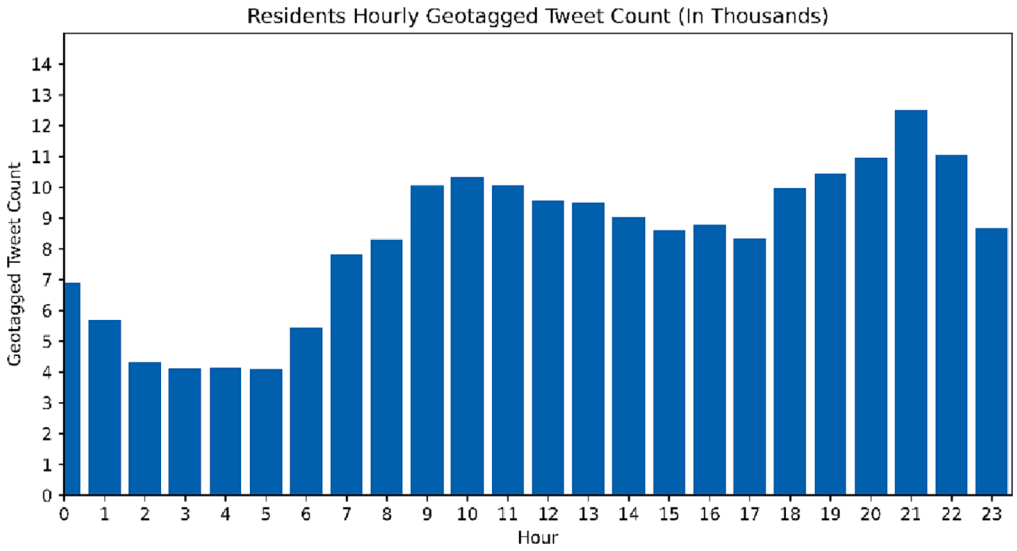

Fig. 3. Description of the REST API Tweet corpus (geo-tagged for residents).

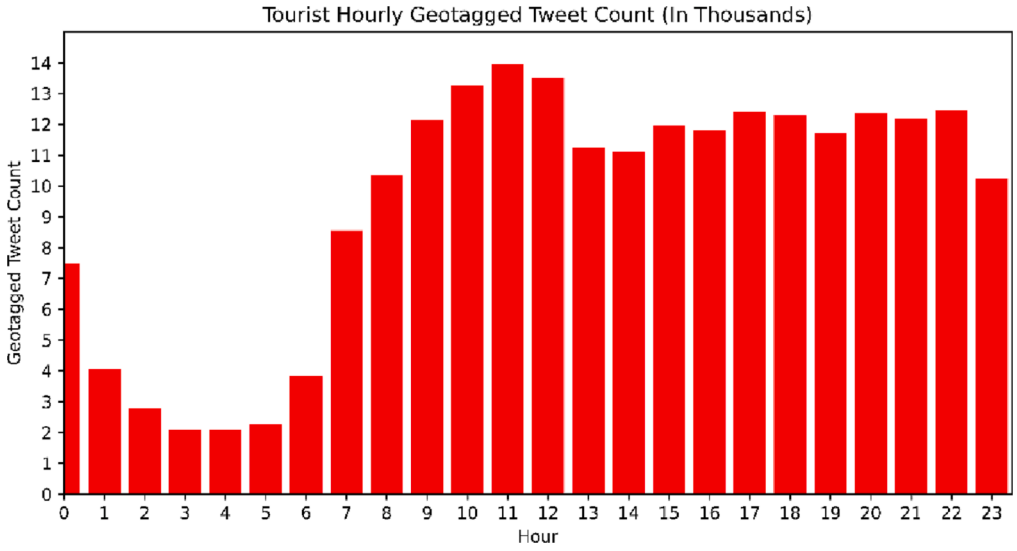

Fig. 4. Description of the REST API Tweet corpus (geo-tagged for tourists).

Along with the traditional classification techniques, we also used the ensemble classification technique. By mixing many models, ensemble learning improves machine learning outcomes. Compared to a single model, this technique provides higher predictive performance. The basic concept is to train a group of classifiers (experts) and then let them vote best classifiers. AdaBoost, Bagging, Random Forest, and Majority Vote are the ensemble approaches used in this study.

The ensemble classifiers could not enhance the performance of initial classification results since the sample size was smaller, so it became challenging to train the models. The other supervised machine learning models performed at par with the proposed heuristic classification technique, with the decision tree performing best.

5. Spatial clustering

We discover the spatial patterns of tourist and resident visitation patterns after identifying and validating the tourist accounts. We recognise the most popular tourist and residential areas using cutting-edge clustering techniques and compare them to changes in the most popular location throughout COVID. The primary objective of clustering is to find *visitors* and *residents* most visited places/regions within the city's boundary. The spots nearer to the cluster's centre would represent the popular spots which are visited by classified residents and visitors. Once the centres of each cluster were obtained, we used Google Streets API to obtain the street-level location of the K-means cluster centroids (latitude and longitude). Besides, the nearest landmark information was also extracted which aids in the identification of the most popular locations for visitors and residents. This study builds on the approach used by Hasnat (2018) to employ Google Maps API for drawing out the details of the locations connected with the coordinate's cluster centres from the road level. In addition, the number of outstanding consumers and the number of samples enable the framing of each cluster. The rationale behind using three distinct methodologies is to discover the best approach to our purpose.

The K-Means clustering technique splits a series of n observations into k sets (k_n) to decrease the within-cluster sum of squares or mean squared distance. With the input parameter k (number of predicted clusters), the method utilises Euclidean Distance as a measure, while variance is used to determine cluster dispersion.

Mean-Shift clustering, commonly known as the mode searching

method, is used to determine the maximum density function. The iterative approach starts with an estimate and then uses the Gaussian Kernel Density function to re-estimate the mean based on the weights of surrounding points. It necessitates using a parameter bandwidth, which defines the form and structure of the kernel density distribution.

DBSCAN is a density-based clustering method that generates a sequence of points. It groups points close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions. DBSCAN has two parameters: epsilon, the maximum distance between two samples which may be taken into account in the same area, and the minimum number of points needed to produce a thick field (Ester., 1996).

5.1. Parameter selection

Clustering algorithm performance may be impacted by the value of $k_{selected}$. Thus, a collection of values might be used instead of a single predetermined K value. It is necessary to reflect on the particular properties of the data sets, as the number of values considered is quite extensive. Simultaneously, the values chosen must be substantially less than the number of items in the data sets so proper representation in clusters is achieved.

Every observation is allocated to one of the k numbers of clusters in K-Means clusters. To choose k, we employed an elbow plot that is two-dimensional distortion compared to the number of clusters (k). The elbow plot visualises the standard deviation of each PC. Where the elbow appears is usually the threshold for identifying most variation. From Fig. 5 (a) and (b), we concluded the number of clusters for visitors and residents to be 11 and 8 respectively.

DBSCAN reduces the number of clusters with a better epsilon value, as each iteration reaches more neighbours. MeanShift would, however, select the densest region with a bandwidth-equivalent radius. DBSCAN chooses the distance in kilometres between two locations instead of the Euclidian Distance in Means and Meanshift. For DBSCAN, we computed the epsilon parameter, which is the maximum distance (1.5 km in our case) that points can be far from each other to be considered a cluster. The *min_samples* parameter is the minimum cluster size (everything else is classified as noise). We set the *min_samples* to 1 so that every data point gets assigned to either a cluster or forms its cluster.

5.2. Clustering performance measure

An external or internal validation technique may be used to assess clustering efficiency. Internal validation techniques were used in conjunction with unsupervised clustering methods. The indices used for internal performance measurement are the Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette validation index. We choose these metrics because of their precision and reputation in the literature and their ease and reliability of implementation. Calinski-Harabasz evaluates cluster validity using the cumulative inter-and within-cluster sum of squares. For better clusters, higher Calinski-Harabasz values are predicted. The Silhouette Index reflects the closeness and dissimilarity of a sample to samples in other clusters within the same cluster (cohesion and separation). It has values ranging from -1 to $+1$, with higher values showing a more notable resemblance within the cluster and lower values showing a lesser similarity within the cluster. For a well-separated cluster, the Davies-Bouldin Index should be lower. The highest value of semblance in Davies-Bouldin is established (i.e., the C1s) between a single cluster, and the index is then multiplied across all clusters (i.e. the C1s to the Cns).

Calinski-Harabasz, and Silhouette Score showed K-Means clustering worked better for both visitors and resident location clustering, as seen in Table 3. Silhouette Index Change is the better of the three approaches for visitor and resident position clustering. The clustering outputs show that K-Means clustering produces acceptable results when the input parameter (number of clusters) is carefully chosen. Based on the optimal values of these indices, the study has confidence in the selected clustering strategy (i.e., choosing proper *k* for K Means).

5.3. Clustering results

The main objective of clustering is to identify the most popular places for visitors and residents in Delhi. The only characteristic similar to the points within a cluster is that they are closer to each other than points outside clusters (outliers in DBSCAN). We determined the centres of each cluster from the output clusters in all three approaches. With the aid of Google Maps, the street level addresses the specifics of the place specified by the coordinates of the cluster centres. The aim of

Table 3
Clustering Performance.

Visitors	Silhouette Score	Calinski Harabas Score	Davies Bouldin Score
K-means	0.70	63361.46	0.57
DBSCAN	0.45	699.23	0.24
Mean shift	0.49	19879.72	0.65
Residents	Silhouette Score	Calinski Harabas Score	Davies Bouldin Score
K-means	0.70	176823.67	0.58
DBSCAN	0.59	4310.63	0.20
Mean shift	0.66	71076.29	0.47

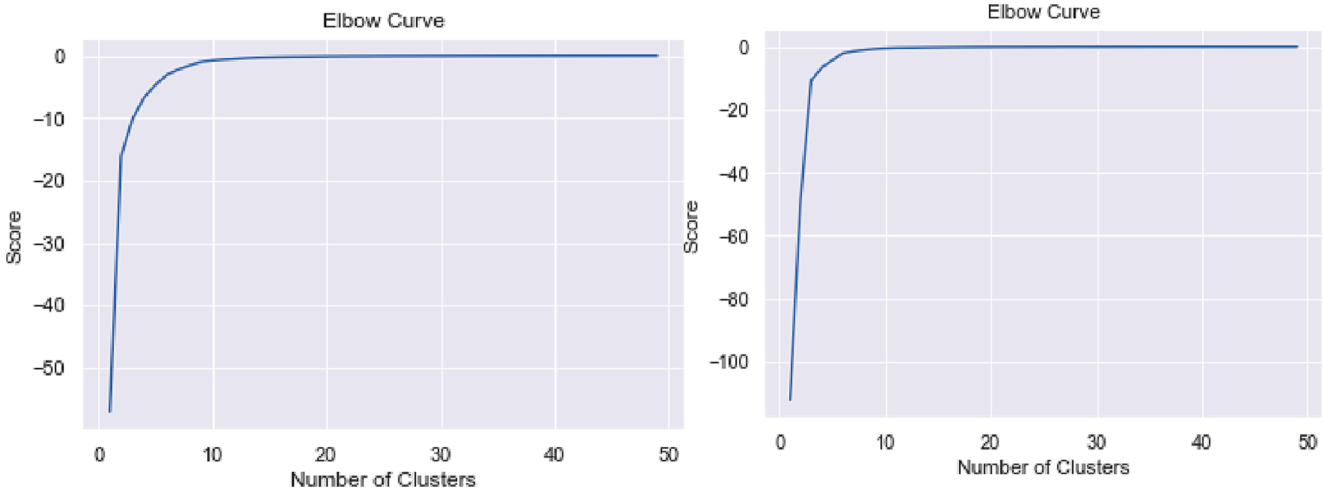

Fig. 5. Elbow Curves for (a) Visitors and (b) Residents.

experimenting with three different methods is to find the one that fits our goal, i.e., to group the coordinates into distinct clusters in real-time. The centres of clusters are depicted in the figures (Figs. 6-7) below. As internal performance measures showed K-Means to perform better than the other two methods, the street-level address of K-Means clustering is included.

The dots represent the coordinates within specific clusters separated by different colors and the comparatively larger dot with black colors represents the centers of the clusters for users classified as tourists (See Fig. 6). The dots within same clusters are represented having same colors. Table 4 below describes the Street Level Address and Points of attraction for the visitor clusters.

Connaught Place, Palika Bazar, Hauz Khas and Rajpath Area are common clustering centres in all three clustering algorithms. In the Table 4, the local landmarks column lists popular visiting sites and visitor attractions within 1.5 Kilometer of the centres.

The dots represent the coordinates within specific clusters separated by different colors and the comparatively larger dot with black colors represents the centers of the clusters for users classified as residents (See Fig. 7). The dots within same clusters are represented having same colors. Table 5 describes the Street Level Address and Points of attraction for the resident clusters.

In this section, the geographical patterns of destinations of visitors and residents were shown using the clustering algorithms K-means, Mean-Shift and DBSCAN. We found that most visitors cluster around renowned visitor sights based on local attractions in the top cluster centres. Most resident geotagged posts are situated in dense residential neighbourhoods with retail malls and the neighbourhood. All of us have Resident clusters of near-famous visitor spots identified in Delhi. The parameters of all three clustering techniques were discovered, and their performance was finally assessed based on typical internal validation indicators used in early studies. K-means was more effective than clustering algorithms DBSCAN and Mean-Shift.

6. Change in geolocation based parameters – Pre and during covid-19

The same three geo-location-based parameters, i.e., (F1) mean distance between successive coordinates; (F2) standard deviation of the distance between consecutive coordinates; and (F3) ratio of the number of tweets originating inside to outside of the study area boundary have been utilised to understand change in travel pattern due to the pandemic (See Fig. 8). It must be noted that both the pre-COVID and during-COVID Tweets are historical Tweets. Amongst that, the pre-COVID dataset includes Tweets before the lockdown started in India (25 March 2020) while the during-COVID dataset includes ones done during the lockdown phases from 25 March 2020 till unlock phase 1.0 was initiated in India, i.e., 01 June 2020. The number of Tweets used for the analysis of pre and during COVID behaviour has been mentioned in Table 2. It is worth mentioning that due to the fact that the number of geotagged Tweets is limited especially for during-COVID period, the interpretations might have certain biases and therefore to be understood with caution.

The first feature helps us understand the distance travelled between two activity events on Twitter based on respective geolocations. Therefore, a low value of mean distance travel describes a lesser tendency of a person to travel frequently to far-distant places. The second feature represents heterogeneity in travel destinations for the individual. A lower standard deviation in the distance travelled suggests trips primarily to workplaces or recreational spaces of choice as it can be assumed that most places will frequently be. In contrast, a higher standard deviation indicates a higher variety in trip nature for the user. The

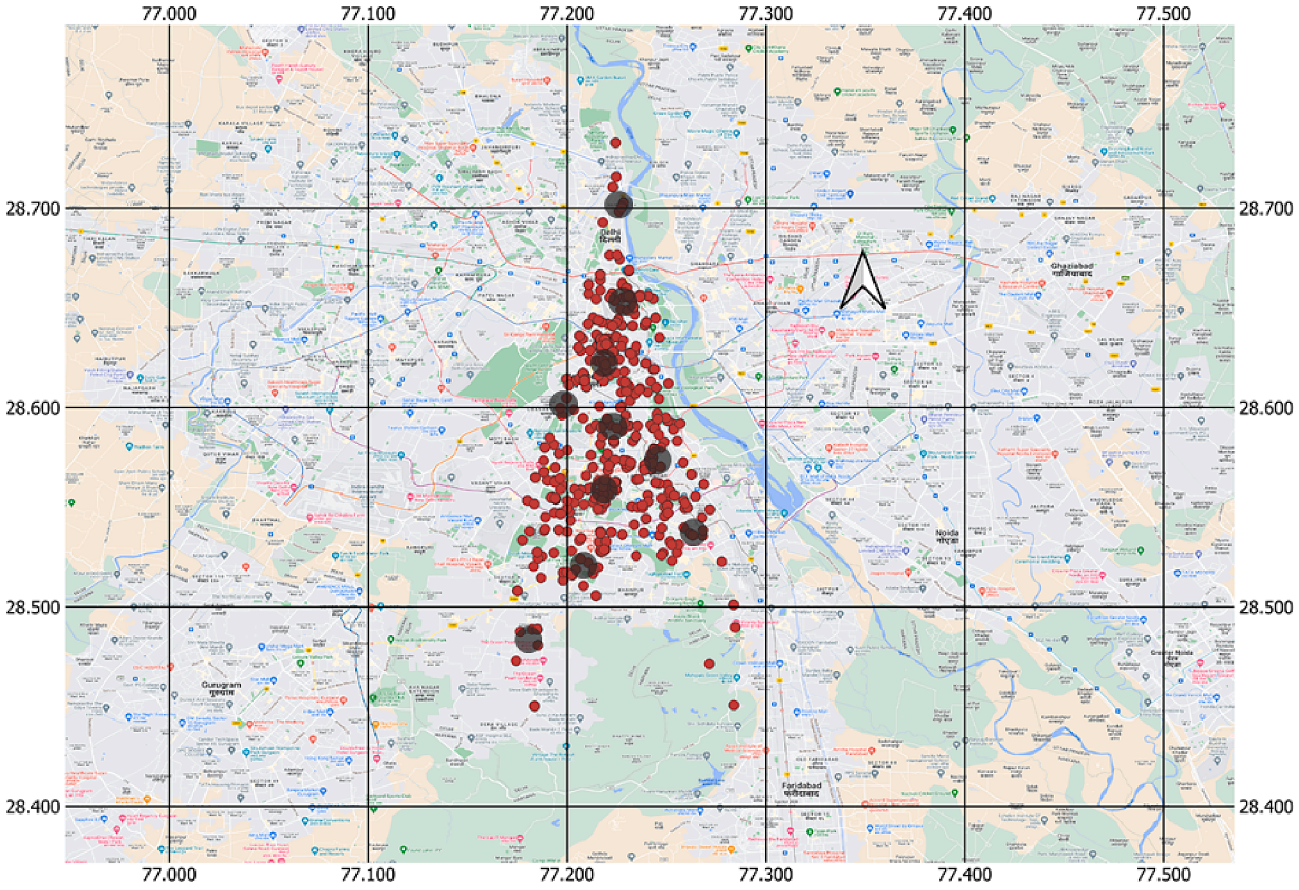

Fig. 6. K-Means Clustering for Visitors.

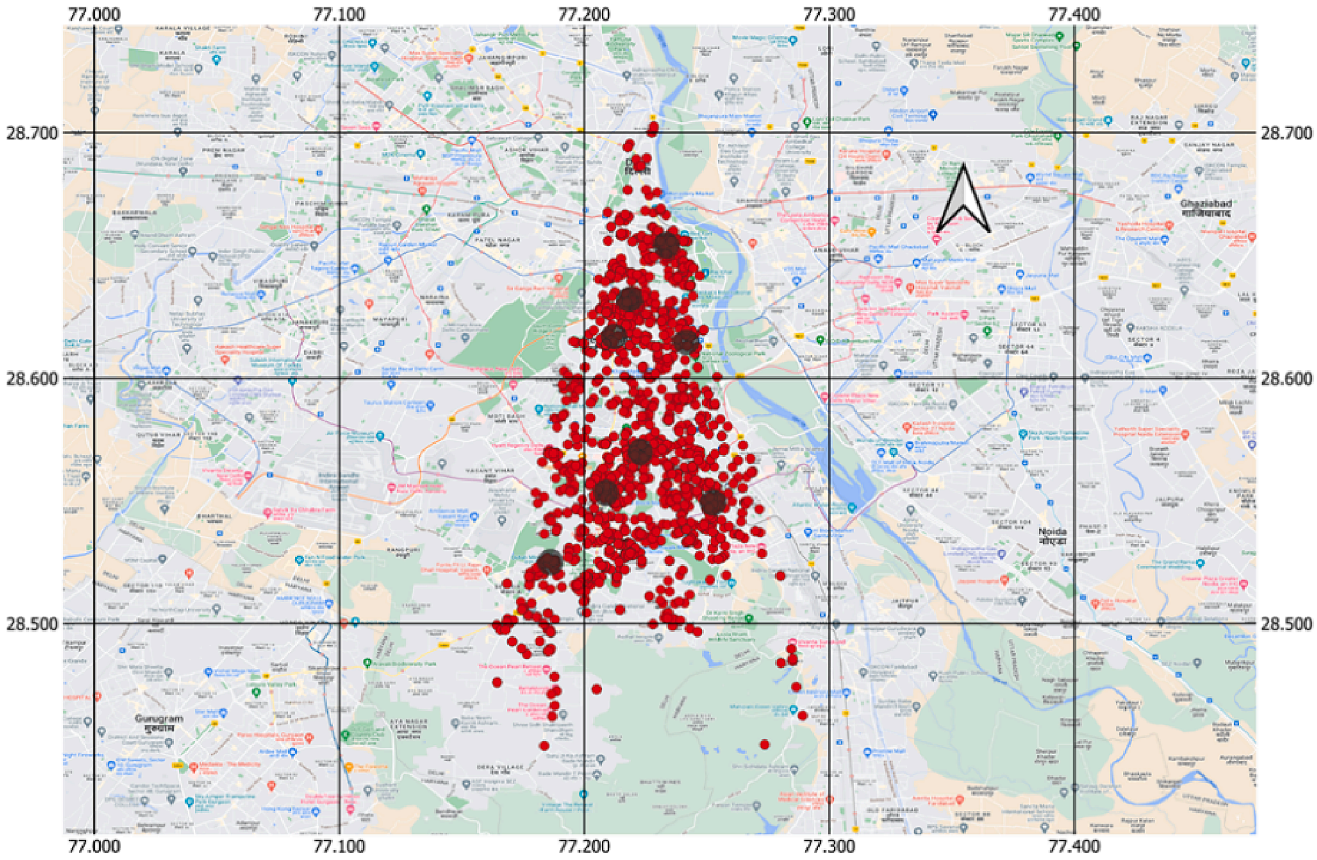

Fig. 7. K-Means Cluster for Residents.

Table 4
K-Means Clusters for Visitors.

Latitude	Longitude	Street Level Address	Point of Attracti ons
28.53	77.25	60, IIIrd Floor, Krishna Market, Kalkaji, New Delhi, Delhi 110019	Kali Temple, Lotus Temple
28.61	77.20	Rajpath Area, Central Secretariat, New Delhi, Delhi 110011	Rashtrapati Bhawan, Sansad
28.55	77.20	Block X, Green Park Extension, Green Park, New Delhi, Delhi 110016	HauzKhasDeer Park, Coaching Centres
28.52	77.21	Ashok Vihar, Saket, New Delhi, Delhi 110017	Qila Rai Pithora Park, Baba Deep Singh Gurudwara
28.66	77.22	3819/3, Mori Gate, Kucha Mohatter Khan, Kashmere Gate, New Delhi, Delhi 110006	MehandipurBalaji temple, RedFort
28.94	77.22	MeerutRd,NewColony, Baghpat,Uttar Pradesh 250609	Residential Areas
28.60	77.23	Golf Links, New Delhi, Delhi 110003	National Zoological Park
28.48	77.18	Khera no. 605, Sultanpur, New Delhi, Delhi 110030	Gummat Mandir
28.72	77.22	104, Gali Number 5, Wazirabad, Delhi, 110084	PrachinMaharshi Valmiki Mandir
28.57	77.23	C-49, Block B, Lajpat Nagar II, Lajpat Nagar, New Delhi, Delhi 110024	Residential Areas
28.63	77.21	Rajeev Chowk, Connaught Place, New Delhi, Delhi 110001	CentralPark, Connaught Palace, Madame Tussauds Delhi

Table 5
K-Means Clusters of Residents.

Latitude	Longitude	Street Level Address	Major Place
28.63	77.21	RadialRoadNumber1, PalikaBazar, Connaught Place, New Delhi, Delhi 110001	ConnaughtPlace, Palika Bazar
28.54	77.20	B/24, Mayfair Gardens, Hauz Khas, New Delhi,Delhi 110016	Hauz Khas Village, Coaching Center
28.94	77.22	New Colony, Ahera, Delhi, Uttar Pradesh 250609	Residential Areas
28.54	77.24	Block B, Chittaranjan Park, New Delhi, Delhi 110048	Residential Complexes and Market
28.61	77.20	Rajpath Area, Central Secretariat, New Delhi, Delhi	Parliament, Rashtrapati Bhavan
28.61	77.22	251,Maliwara,KatraLehswan, ChandniChowk, New Delhi, Delhi 110006	ChandaniChowk, Gandhi Park
28.51	77.18	Mittal Garden, Sainik Farm, New Delhi, Delhi 110030	Qutub Minar
28.58	77.23	Jawaharlal Nehru Stadium, Pragati Vihar, New Delhi, Delhi 110003	JawaharlalNehru Stadium, IndiaGate

third feature represents the amount of activity the user does inside the study area boundary. If the ratio is more, there is a higher chance for that individual to be a city resident since it describes higher activity is done within the boundaries of the city under consideration.

Besides, number of tweets with more than 100 km distance between those as shown earlier was collected for both pre and during-COVID period. This aids in understanding the impact of COVID restrictions as the possibility of long-distance travel is minimized and thus indicates

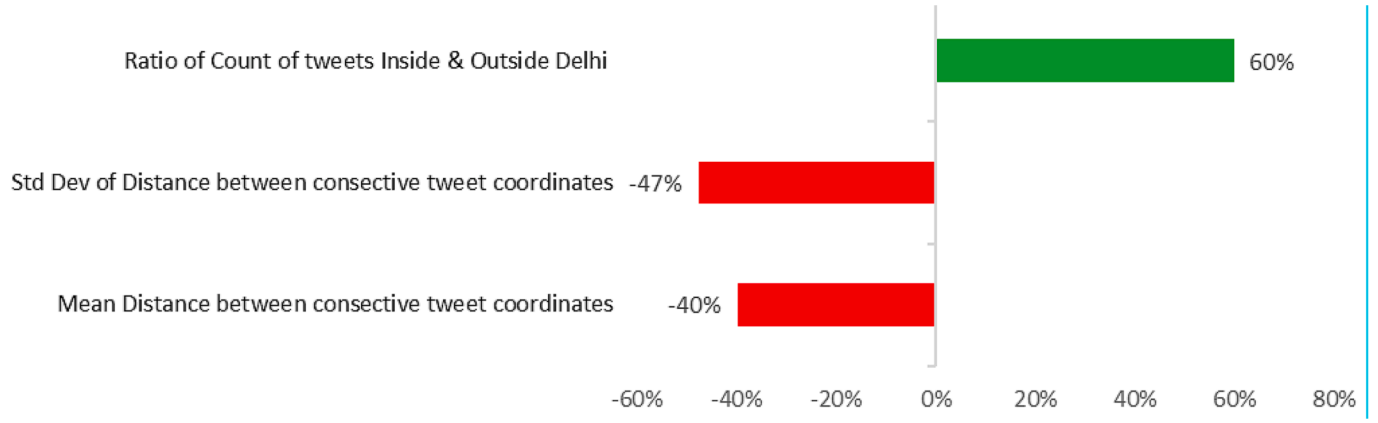

Fig. 8. Comparison of parameters pre and during COVID-19.

the effectiveness of the policy level changes.

The results suggest a drop in values of each parameter except F3 due to various restrictions being implemented all over the country. As the frequency of data for parameters is not uniformly distributed, we decided to compare the median values of each of the parameters. The median of the parameter mean distance travelled by each user during COVID has dropped by 40% compared to before COVID-19. There is a drop of 47% in the standard deviation of distance travelled during COVID-19. Similarly, the value of R, i.e., the ratio of the number of tweets inside the boundary of Delhi and Outside of Delhi, has increased by 59%. This is likely due to the fact that people had to stay for longer periods in their place of residence. Interestingly, the number of consecutive points having a jump of greater than 100 km has reduced by 50%. As a theoretical measure it indicates that the tendency of people traveling to longer distances is reduced significantly. This can be attributed to various limitations the state government employs to reduce the influx of people from outside the city and implies that the policy decisions taken by the government to curb movement of people during COVID was a success.

There are changes in overall travel patterns. But they are not as high as expected because of the decrease in restrictions after the first significant lockdown from March 2020 to May 2020.

7. Changes in clusters

The primary objective of clustering is to identify the most popular visitor and resident destinations. Expectedly, the number of clusters for the dataset before COVID (See Fig. 5) showed a higher number of cluster centres relative to the during COVID period (See Fig. 9) as the dataset for the earlier was more scattered.

The figures represent K-Means Clusters for residents and visitors respectively during the COVID-19 period (See Figs. 10-11). The number of clusters formed is lesser compared to our initial observations for clusters. The dots represent the coordinates within specific clusters separated by different colors and the comparatively larger dot with black colors represents the centers of the clusters. The dots within same clusters are represented having same colors. Table 6 and Table 7 describe the street-level address and points of attraction for the resident and vistor clusters respectively. The key observations are following:

(1) The number of clusters (based on Tweet geolocation) has significantly dropped during COVID. This helps us understand the spatial variation of tweets was reduced considerably as people could not travel to different places due to stringent lockdown restrictions. Furthermore, we validated our finding using ground truth data (GoogleEarth) since the cluster centres of the residents' tweets were primarily observed near the city's residential areas.
(2) As for the visitors' tweets, the number of during-COVID clusters also decreased relative to pre-COVID, more so from so-called tourist attraction points (say, heritage monuments) while places with specific characteristics emerge as new clusters. For example, we observe Connaught Place as one of the vital clusters attributed to its multi-purpose nature facilitating the supply of essential commodities during the pandemic. Also, another cluster centre was found near Hauz Khas which acts as a central hub for students preparing for job examinations. This is probably due to the

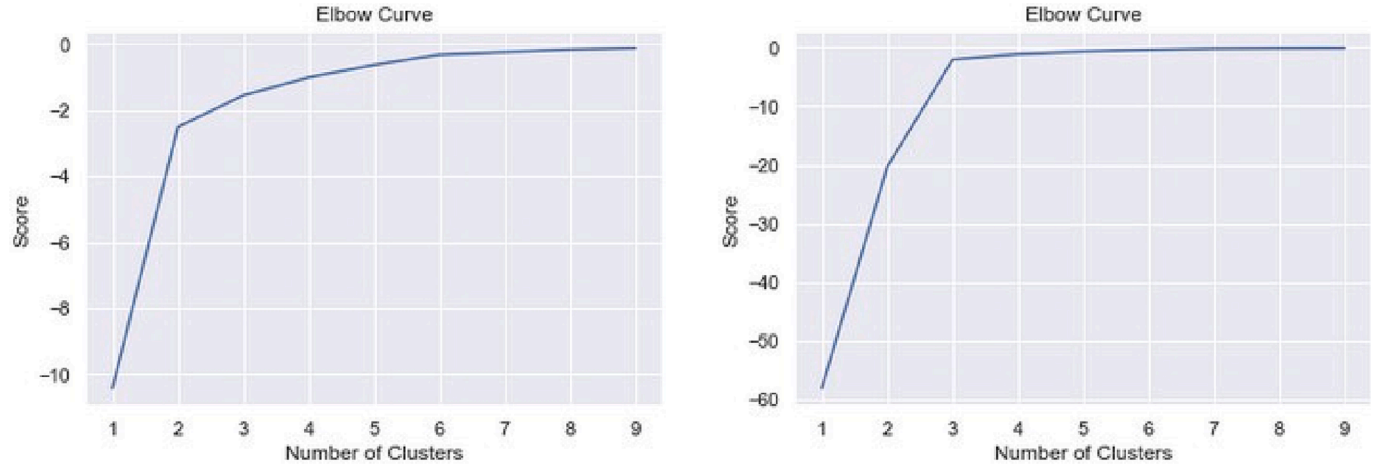

Fig. 9. Elbow Curves for (a) Visitors and (b) Residents.

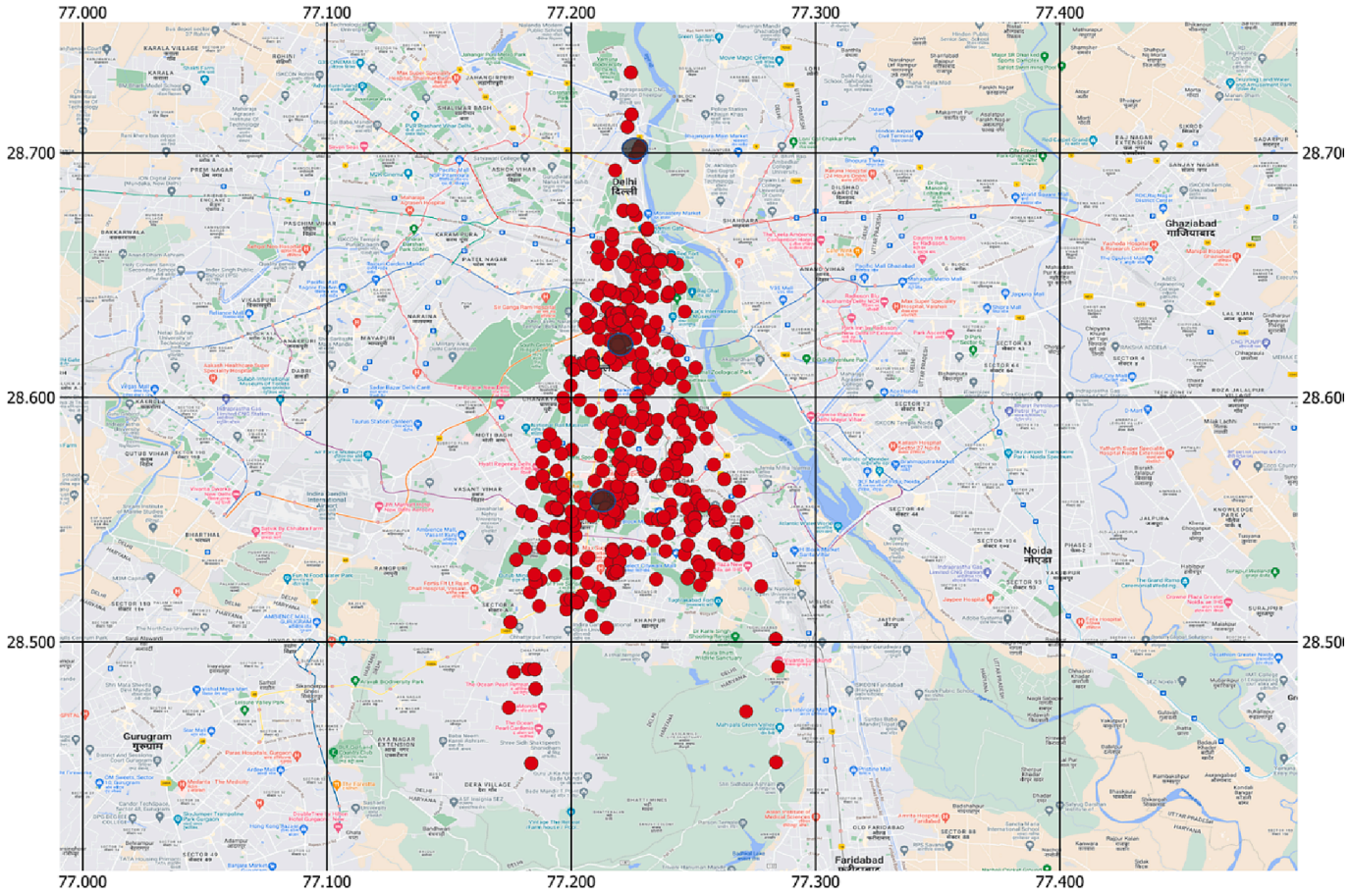

Fig. 10. Clustering Results for Residents during COVID.

fact that some of the students coming from other parts of the country could not move to their home locations due to nationwide travel restrictions.

The results of COVID clusters are very different to the clusters of the rest of the tweets. The reason being for visitor tweets, most of the clusters were spread among the whole city's visitors part, but here, they are only clustered around only one prominent visitor place in the city for both residents and visitors. This suggests lower tendencies of people to travel for tourism purposes.

8. Conclusions

The study proposed a technique for extracting and analysing big data from the Twitter platform of visitors and residents in Delhi. Filtering measures are used to ensure that the user has active geolocation. Using a basic heuristic classification approach, the study categorised visitors and residents from filtered Twitter info. The developed algorithm outperforms most of the commonly used supervised classification techniques. The Decision Tree classifier outperformed the suggested heuristic classification as opposed to other state-of-the-art ensemble classification techniques. Without extensive content-based studies, all characteristics for developing sophisticated classification algorithms are derived from geographical coordinates (from geotagged posts).

Three cluster methods, i.e., K-Means, Mean-Shift, and DBSCAN, have been used to find spatial trends and places of attractions in Delhi visited by visitors and locals. Many popular tourism spots, coaching areas and some urban residential areas are grouped from the visitor cluster centres. In addition, within a 1.5-kilometre radius, the residents' positions are divided into many schools, retail centres and parks around a residential apartment. The efficiency of the clustering methods is calculated based on the prevalent clustering validation tests. The K-Means clustering approach was superior to others.

There is a significant change in the parameters extracted from geographical coordinates for tweets done before COVID-19 and after its spread. All the changes are directed towards the lesser tendency of movement of Twitter users. Looking at clusters formed after the pandemic started, there are sparse geotagged posts with respect to visitor spots. The only places visited seem to be Connaught place and the Janpath area, the most prevalent visitor places in the city. A significant takeaway is the presence of a cluster centre in Hauz Khas village, a coaching hub. This suggests some students must have continued their studies even during the pandemic.

The findings of our analysis of visitor and resident habits have far-reaching consequences. First, it demonstrates how to use social media to gather and analyze accurate data on travel activity. Conventional surveys are generally costly and difficult to carry out; social media may be utilised in a fast-developing region as a cost-effective source of up-to-date travel data. Secondly, our analysis illustrates how the choice of tourism sites might create distinct patterns. The effects travellers may have on the traffic movement in an area can be determined by the combination of spatial clusters in temporal windows. Third, this research provides a method for understanding person-level travel activity for visitors and residents using robust data blending techniques.

This study, therefore, showed that using social media information to understand visitor behaviour, and the techniques and results of this study will be highly valuable in future investigations into the travel behaviour of visitors. An application was demonstrated where the change in travel behaviour between pre-COVID and post-COVID conditions was captured. The parameters selected for comparison show a

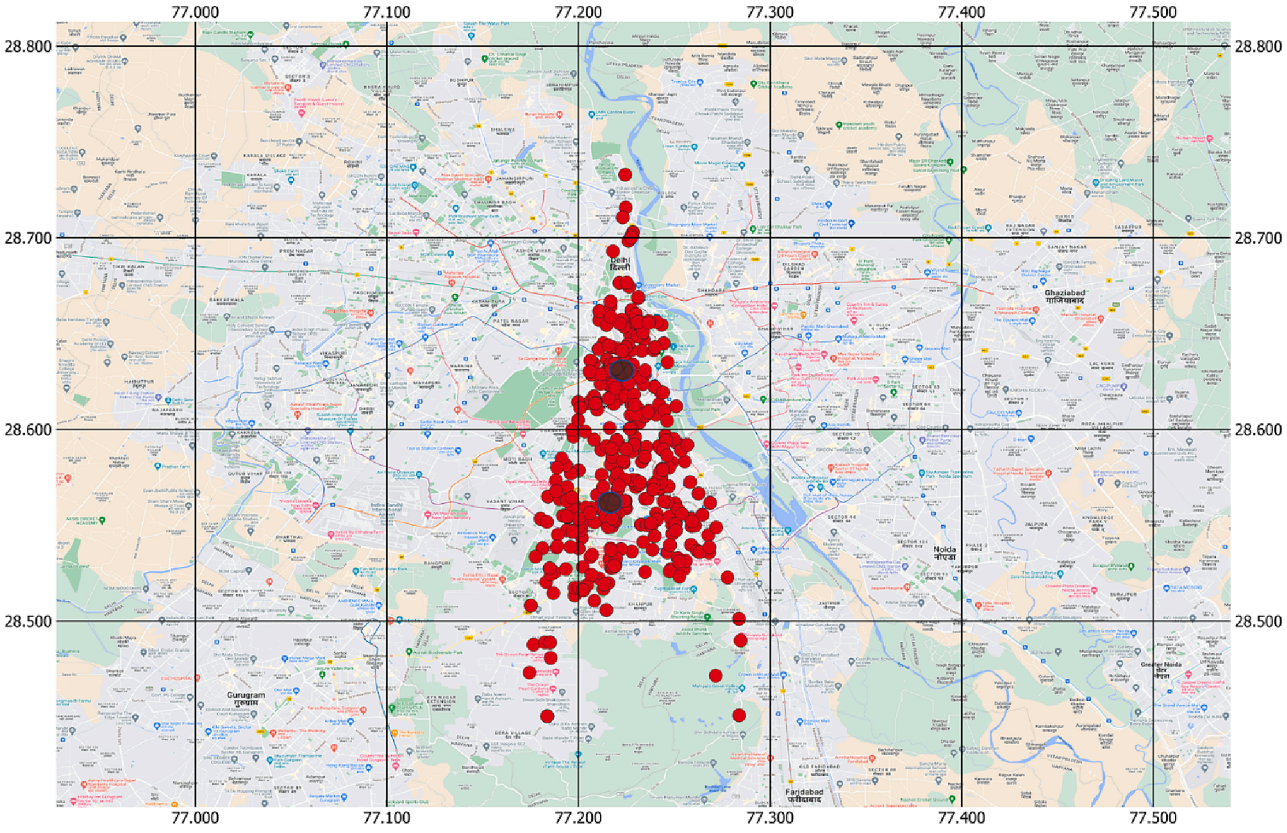

Fig. 11. Clustering Results for Visitors during COVID.

Table 6
During-COVID K-Means Clusters of Residents.

Latitude	Longitude	Street Level Address	Major Place
28.628	77.218	JanpathBhawan, Connaught Ln, Janpath, Connaught Place, New Delhi, Delhi 110001	Parliament, President House, Jantar Mantar
28.541	77.229	Block B, Panchsheel Enclave, New Delhi, Delhi 110049	Residential Complexes
28.945	77.22	Near, 1st Floor, Approva Tower, Nirojpur Road, Baghpat, Delhi	Residential Complexes

Table 7
During COVID K-Means Clusters of Visitors.

Latitude	Longitude	Street Level Address	Major Place
28.628	77.218	Palika Parking Way, Palika Bazar, Connaught Place, New Delhi, Delhi 110001	Connaught Place, Shopping Centres
28.544	77.223	Masjid Moth, Greater Kailash, New Delhi, Delhi	Hauz Khas,Forests, Coaching Centers

drop in values during COVID, while the clustering results show lesser spatial variation during the same period.

For future research, there is a possibility of extracting traffic impacts in a study area. In this case, however, researchers must be careful when using cluster tweets because people do not necessarily publish tweets at the beginning or the end. We did not use text mining in our analysis. Since text constitutes an essential part of Twitter results, future studies may include tweet text-extracted features and incorporate them.

CRediT authorship contribution statement

Swapnil Shende: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft. **Eeshan Bhaduri:** Conceptualization, Data curation, Methodology, Formal analysis, Visualization, Writing – review & editing. **Arkopal Kishore Goswami:** Formal analysis, Conceptualization, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors acknowledge the support from the Paramshakti supercomputing facility at IIT Kharagpur and Scheme for Promotion of Academic and Research Collaboration (SPARC), Ministry of Education, Government of India.

References

Abbasi, A., Rashidi, T. H., Maghrebi, M., & Waller, S. T. (2015). Utilising location based social media in travel survey methods: Bringing Twitter data into the play. Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location- Based Social Networks, LBSN 2015 - Held in Conjunction with ACM SIGSPATIAL 2015, 1–9. https://doi.org/10.1145/2830657.2830660.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., Thom, D., 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. Comput. Sci. Eng. 15 (3), 72–82. https://doi.org/10.1109/MCSE.2013.70.

Arentze, T., Bos, I., Molin, E., Timmermans, H., 2005. Internet-based travel surveys: Selected evidence on response rates, sampling bias and reliability. Transportmetrica 1, 193–207. https://doi.org/10.1080/18128600508685648.

Barbieri, D.M., Lou, B., Passavanti, M., Hui, C., Hoff, I., Lessa, D.A., Sikka, G., Chang, K., Gupta, A., Fang, K., Banerjee, A., Maharaj, B., Lam, L., Ghasemi, N., Naik, B., Wang,

F., Foroutan Mirhosseini, A., Naseri, S., Liu, Z., Qiao, Y., Tucker, A., Wijayaratna, K., Peprah, P., Adomako, S., Yu, L., Goswami, S., Chen, H., Shu, B., Hessami, A., Abbas, M., Agarwal, N., Rashidi, T.H., 2021. Impact of COVID-19 pandemic on mobility in ten countries and associated perceived risk for all transport modes. In: Pakpour, A.H. (Ed.), PLOS ONE 16. https://doi.org/10.1371/journal.pone.0245886.

Beck, M.J., Hensher, D.A., Wei, E., 2020. Slowly coming out of COVID-19 restrictions in Australia: Implications for working from home and commuting trips by car and public transport. J. Transp. Geogr. 88, 102846 https://doi.org/10.1016/j.jtrangeo.2020.102846.

Bhaduri, E., Manoj, B.S., Wadud, Z., Goswami, A.K., Choudhury, C.F., 2020. Modelling the effects of COVID-19 on travel mode choice behaviour in India. Transportation Research Interdisciplinary Perspectives 8, 100273. https://doi.org/10.1016/j.trip.2020.100273.

Bhattacharyya, K., Dandapat, S., Annam, S.K., Saysardar, K., Maitra, B., 2021. Exploring Public Perception towards Travel and COVID-19 Preventive Measures: Insights from the Early Stages of Lockdown in India. Transportation Research Board 100th Annual Meeting.

Bucsky, P., 2020. Modal share changes due to COVID-19: The case of Budapest. Transportation Research Interdisciplinary Perspectives 8, 100141.

Buckee, C.O., Balsari, S., Chan, J., Crosas, M., Dominici, F., Gasser, U., Grad, Y.H., Grenfell, B., Halloran, M.E., Kraemer, M.U.G., Lipsitch, M., Metcalf, C.J.E., Meyers, L.A., Perkins, T.A., Santillana, M., Scarpino, S.V., Viboud, C., Wesolowski, A., Schroeder, A., 2020. Aggregated mobility data could help fight COVID-19. Science 368 (6487), 145–146. https://doi.org/10.1126/SCIENCE.ABB8021.

Coppola, P., De Fabiis, F., 2021. Impacts of interpersonal distancing on-board trains during the COVID-19 emergency. Eur. Transp. Res. Rev. 13, 1–12. https://doi.org/10.1186/S12544-021-00474-6/TABLES/2.

Coppola, P., Fabiis, F.D., 2020. Evolution of mobility sector during and beyond COVID-19 emergency: a viewpoint of industry consultancies and public transport companies. TeMA - Journal of Land Use, Mobility and Environment 81–90. https://doi.org/10.6092/1970-9870/6900.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. In: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press. https://doi.org/10.1017/cbo9780511801389.

Das, S., Boruah, A., Banerjee, A., Raoniar, R., Nama, S., Maurya, A.K., 2021. Impact of COVID-19: A radical modal shift from public to private transport mode. Transp. Policy 109, 1–11. https://doi.org/10.1016/j.tranpol.2021.05.005.

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C., 2010. Automatic construction of travel itineraries using social breadcrumbs. In: HT'10-Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, pp. 35–44. https://doi.org/10.1145/1810617.1810626.

De Vos, J., 2020. The effect of COVID-19 and subsequent social distancing on travel behaviour. Transportation Research Interdisciplinary Perspectives 5, 100121. https://doi.org/10.1016/j.trip.2020.100121.

Ester. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise | Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Retrieved 16 September, 2020, from https://dl.acm.org/doi/10.5555/3001460.3001507.

K. Fu R. Nune J.X. Tao Social Media Data Analysis for Traffic Incident Detection and Management 2015.

Gao, H., Tang, J., Hu, X., Liu, H., 2013. Exploring temporal effects for location recommendation on location-based social networks. In: RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems, pp. 93–100. https://doi.org/10.1145/2507157.2507182.

Girardin, F., Blat, J., Calabrese, F., Dal Fiore, F., Ratti, C., 2008. Digital footprinting: Uncovering visitors with user-generated content. IEEE Pervasive Comput. 7 (4), 36–44. https://doi.org/10.1109/MPRV.2008.71.

Guzman, L.A., Arellana, J., Oviedo, D., Moncada Aristizábal, C.A., 2021. COVID-19, activity and mobility patterns in Bogotá. Are we ready for a '15-minute city'? Travel Behav. Soc. 24, 245–256. https://doi.org/10.1016/j.tbs.2021.04.008.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geolocation data. Transportation Research Part C: Emerging Technologies 44, 363–381. https://doi.org/10.1016/j.trc.2014.04.003.

Hasan, S., Zhan, X., Ukkusuri, S.V., 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2505821.2505823.

Hasnat, M., 2018. Analysing Destination Choices of Visitors and Residents from Location Based Social Media Data. Electronic Theses and Dissertations 2004–2019. https://stars.library.ucf.edu/etd/5774.

Hasnat, M.M., Hasan, S., 2018. Identifying visitors and analysing spatial patterns of their destinations from location-based social media data. Transportation Research Part C: Emerging Technologies 96 (January), 38–54. https://doi.org/10.1016/j.trc.2018.09.006.

Huang, X., Li, Z., Jiang, Y., Li, X., Porter, D., Gao, S., 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS One 15 (11). https://doi.org/10.1371/JOURNAL.PONE.0241957.

Ichimura, T., Kamada, S., 2012. A generation method of filtering rules of Twitter via smartphone based participatory sensing system for visitor by interactive GHSOM and C4.5. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, pp. 110–115. https://doi.org/10.1109/ICSMC.2012.6377685.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithms: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 881–892. https://doi.org/10.1109/TPAMI.2002.1017616.

Lee, J.H., Davis, A.W., Yoon, S.Y., Goulias, K.G., 2016. Activity space estimation with longitudinal observations of social media data. Transportation 43 (6), 955–977. https://doi.org/10.1007/s11116-016-9719-1.

Mai, E.C., Hranac, R.C., 2013. Twitter Interactions as a. Data Source for Transportation Incidents. Undefined.

Majid, A., Chen, L., Chen, G., Mirza, H.T., Hussain, I., Woodward, J., 2013. A context-aware personalised travel recommendation system based on geotagged social media data mining. Int. J. Geogr. Inf. Sci. 27 (4), 662–684. https://doi.org/10.1080/13658816.2012.696649.

Manca, M., Boratto, L., Morell, V., Roman, Martori, i, Gallissà,, O.,, &, Kaltenbrunner,, A., 2017. Using social media to characterise urban mobility patterns: State-of-the-art survey and case-study. Online Social Networks and Media 1, 56–69. https://doi.org/10.1016/j.osnem.2017.04.002.

McNeill, G., Bright, J., Hale, S.A., 2017. Estimating local commuting patterns from geolocated Twitter data. EPJ Data Sci. 6 (1), 24. https://doi.org/10.1140/epjds/s13688-017-0120-x.

Popescu, A., Grefenstette, G., 2011. Mining social media to create personalised recommendations for visitor visits. ACM International Conference Proceeding Series 1–6. https://doi.org/10.1145/1999320.1999357.

Pozdnoukhov, A., & Kaiser, C. (2011). Space-time dynamics of topics in streaming text. 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2011 - Held in Conjunction with the 19th ACM SIGSPATIAL GIS 2011, 1. https://doi.org/10.1145/2063212.2063223.

Safavian, S.R., Landgrebe, D., 1991. A Survey of Decision Tree Classifier Methodology. IEEE Trans. Syst. Man Cybern. 21 (3), 660–674. https://doi.org/10.1109/21.97458.

R.J. Steur Twitter as a spatio-temporal source for incident management 2015 http://dspace.library.uu.nl/handle/1874/303174.

Sun, Y., Fan, H., Helbich, M., Zipf, A., 2013. Analysing human activities through volunteered geographic information: Using flickr to analyze spatial and temporal pattern of visitor accommodation 9783642342028, 57–69. https://doi.org/10.1007/978-3-642-34203-5_4.

Tirachini, A., Cats, O., 2020. COVID-19 and Public Transportation: Current Assessment, Prospects, and Research Needs. J. Public Transp. 22, 1–34. https://doi.org/10.5038/2375-0901.22.1.1.

UNDP. (n.d.). Social media For Youth & Civic Engagement in India.

Vechtomova, O., 2009. Introduction to Information Retrieval Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (Stanford University, Yahoo! Research, and University of Stuttgart) Cambridge: Cambridge University Press, 2008, xxi+482 pp; hardbound, ISBN 978-0-521-86571-5, $60.00. Comput. Linguist. 35 (2), 307–309.

Ward, J.H., 1963. Hierarchical Grouping to Optimise an Objective Function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Xiang, Z., Gretzel, U., 2010. Role of social media in online travel information search. Tour. Manag. 31 (2), 179–188. https://doi.org/10.1016/j.tourman.2009.02.016.

Zannat, K.E., Bhaduri, E., Goswami, A.K., Choudhury, C.F., 2021. The tale of two countries: modeling the effects of COVID-19 on shopping behaviour in Bangladesh and India. Transportation Letters 1–13. https://doi.org/10.1080/19427867.2021.1892939.