IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 July 2022, accepted 12 September 2022, date of publication 15 September 2022,
date of current version 22 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206859

RESEARCH ARTICLE

Big Data-Driven Approach to Analyzing Spatio-Temporal Mobility Pattern

MUNAIRAH ALJERI, (Member, IEEE)
Kuwait Institute for Scientific Research, Safat 13109, Kuwait
e-mail: mujeri@kisr.edu.kw

ABSTRACT It is imperative to understand human movement and behavior, from epidemic monitoring to complex communications. So far, most research and studies on investigating and interpreting human movements have traditionally depended on private and accumulated data such as mobile records. In this work, social network data is suggested as a proxy for human mobility, as it relies on a large amount of publicly accessible data. A mechanism for urban mobility mining and extraction scheme is proposed in this research to shed light on the importance and benefits of the publicly available social network data. Given the potential value of the Big Data obtained from social network platforms, we sought to demonstrate the process of analyzing and understanding human mobility patterns and activity behavior in urban areas through the social network data. Human mobility is far from spontaneous, follows well-defined statistical patterns. This research provides evidence of spatial and temporal regularity in human mobility patterns by examining daily individual trajectories of users covering an average time span of three years (2018 to 2020). Despite the diversity of individual movements history, we concluded that humans follow simple, reproducible patterns. Additionally, we studied and evaluated the effect of COVID-19 on human mobility and activity behavior in urban areas and established a strong association between human mobility and COVID-19 spread. Numerous years of mobility data analysis can reveal well-established trends, such as social or cultural activities, which serve as a baseline for detecting anomalies and changes in human mobility and activity behavior.

INDEX TERMS Big data, COVID-19, data mining, mobility pattern, social network.

I. INTRODUCTION

The movement of people is intimately related to distribution patterns of land and space. The flowing movement of the population over spatial and temporal boundaries significantly influences communities and cities' economy, society, and environment. Human mobility occurs on a daily basis in metropolitan regions, where the majority of the population interacts with many urban components and activities, including stores, malls, offices, and public facilities. Understanding how people move and behave in urban areas is imperative of addressing various challenges such as: discovering the physical and social character of cities [20], predicting and understanding road traffic and congestion by estimating population flow [21], improving the control of disease outbreaks [18], and advances transportation planning by investigating resident's daily routes [12].

Early studies of human mobility mostly depend on census data, which was hard to gain inclusive results due to data quantity and geographic coverage limitations. Today, with the well-developed mobile communication and tracking technologies such as mobile devices and GPS, it is possible to gather and acquire large-volume and diverse data related to human mobility. The investigation of human mobility thus becomes heavily data-driven and spreads into countless directions of researches and studies. Data from network carriers such as cell services companies, on the other hand are aggregated and generalized, and impossible for the general public to obtain.

Recent research have proposed that social network data can be used instead of mobile phone logs [14], [15], [25], [29], thanks to its widespread availability, location accuracy, and massive size. Social networking has become a popular

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

activity in the developed society, as it allows people to remain socially connected with friends, family, and colleagues. Social networks can provide multiple purposes to their users, including business and social purposes, through websites and platforms like Twitter and Instagram (social media applications). Active users on social media reached 3.69 billion worldwide as of 2021, with a total of 58% of the world's population [28]. Social media data has become one of the most representative and relevant data sources for *Big Data*, thanks to its massive data size and widespread availability. Big Data has become an essential issue for a large number of research areas, including pattern mining. The vast amount of heterogeneous data generated every day from social media applications can represent a footprint of daily users' mobility, which represents their travel patterns and activity behaviors.

In recent years, social network data has been extensively utilized in a variety of areas, including public health. During the outbreak of COVID-19, social networking data like Twitter and Weibo were used to assess public interest [27], anticipate spread and second waves [24], and conduct analysis on human emotions [23].Such evidence in disaster studies will help researchers better analyze the epidemic in terms of time and space, as well as play a key role in developing control strategies.

This article proposes a mechanism for urban mobility mining and extraction that sheds light on the importance and benefits of the publicly available social network data. Given the potential value of the Big Data obtained from social media platforms, we sought to demonstrate the process of analyzing and understanding human mobility patterns and activity behavior in urban areas through social media data. The primary objective of this work is to give an in-depth study and analysis of human mobility patterns in the State of Kuwait using big social network data. This paper's major contributions are outlined as follows:

- Design a novel mobility detection and pattern recognition scheme for Spatio-temporal social networks.
- A thorough evaluation of different mobility characteristics and pattern behavior analysis during a time-span of three years.
- A comprehensive mobility analysis of human activity patterns during the current global pandemic and its impact on individual's mobility and regional trajectories.
- Relationship identification between human mobility and the spread of COVID-19 in urban areas.

A brief illustration of the suggested workflow is displayed in Fig. 1.

The remainder of this article is organized as follows. First, an overview of the background is described in section II. Followed by the structure of data acquisition and pre-processing technique in section III. Then the proposed mobility pattern identification scheme is described in section IV. In section V, we discuss the impact of the global pandemic on human mobility. Section VI describes the article's shortcomings and limitations. Finally, section VII reports the conclusion and future direction.

II. BACKGROUND

This section discusses relevant research conducted to analyze human mobility patterns through the use of social media data. The section consists of two main parts—first, an overview of the general studies and research analyzing and characterizing mobility patterns and user behavior. Second, research conducted during the current COVID-19 pandemic investigates human mobility patterns in urban areas from derived social media data.

Human mobility describes the movement of individuals from original location to multiple destinations during a specific period. Human mobility has recently emerged as a significant and intriguing study subject across a wide range of disciplines, ranging from computer technology to social science and geography. The concept can be applied to various studies, including smart cities, crisis decision-making, migration, and transportation planning. Understanding and analyzing human mobility becomes a critical research question that requires integrating different data sources that are publicly available and widely used worldwide, such as social network data. The following are some different research studies that considered social network data in analyzing the spatial-temporal mobility patterns.

Ullah *et al.* [31] analyzed spatial and temporal data from location-based social networks (Weibo) to show the impact of people in green spaces. Using the kernel density estimation approach, they examined the environmental characteristics and external factors linked to the spatiotemporal distribution of user's movements in urban green areas. They concluded that spatial agglomeration in urban parks was more significant in the center due to high human activity levels.

Ebrahimpour *et al.* [7] investigated social-geographic human mobility using social media applications and tracked user's behavior in Shanghai. They validated the use of geo-location data in detecting human mobility patterns and discovered Spatio-temporal relations between user movement and the hour of the day. They proved a linked connection between the geographic distribution of the city and the mobility of social media users. They analyzed the spatial pattern of user's movement towards points of interest in the city, concluding the result of their research is helpful in urban planning and the design of the city's infrastructure.

Xuan *et al.* [33] mined the movement patterns of various groups using smart card data along with social network data. Through quantifying spatial and temporal features of students and visitor passengers, they shed light on useful insights into mobility movement. The K-means algorithm was used to categorize consumers and examine how their destinations and journeys changed over the course of the week. According to the amount of tap-ins and tap-outs, the findings indicated a major hub in the travel metro system.

As multiple recent studies have shown [8], [10], [34], human mobility data is critical in the battle against the spread

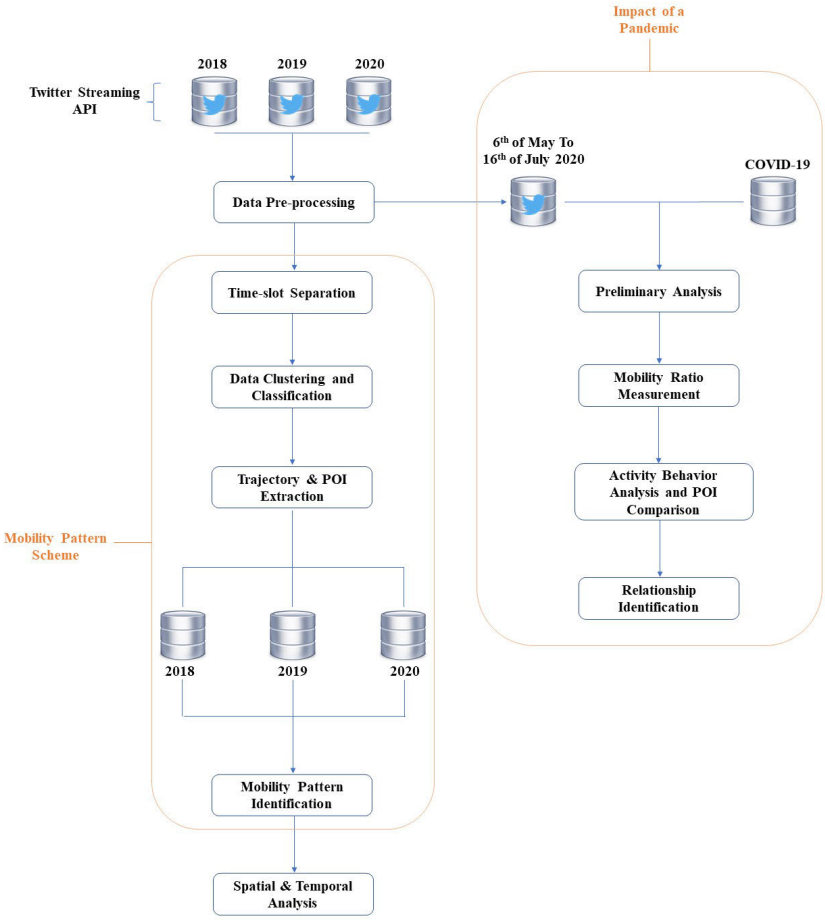


FIGURE 1. Overview of the proposed workflow.

of the COVID-19 pandemic. Many experts and policymakers have begun to target and analyze human movement as one of the primary instruments in monitoring the COVID-19 epidemic, as well as a central element influencing the virus's dissemination on a national and global scale. The following research examined mobility from social network applications in relation to the latest corona-virus outbreak:

Huang *et al.* [11] investigated the effect of multinational collaborative activities on human movements and travel by reviewing data from over 500 million tweets from around the globe. As a versatility measure, the writers suggested two forms of distances: day-to-day time and three-day prior time. The planned distances reveled a significant change in mobility journeys and travel, indicating that government-imposed policies have had a direct effect on users' everyday habits and that mobility improvements is well correlated with the restricted measures implemented in the United States.

Xu *et al.* [32] introduced the Twitter mobility social index as a metric of social distancing focused on Twitter collected information. They calculated the social index for several days from January 2019, to April 2020, and estimated travel distance of a person during the week. They discovered a significant decrease in movement in the US following the introduction of social distancing interventions. The improvement and changes in travel, on the other hand, differed greatly by state.

According to Bisanzio *et al.* [3], the dissemination of Spatial temporal corona-virus worldwide occurred within the first few weeks of the pandemic, when over 60% of the population in Wuhan, China migrated to other countries. Analysis of Twitter data from 2013 to 2014 was used to predict mobility trends. Bisanzo *et al.* was able to categorize countries with low and high IDVI frequencies and forecast positive cases among places travelled by a cohort's users by measuring the infectious disease risk index (IDVI).

Even though there may be some parallels between the methodologies indicated above and the work presented. Some significant discrepancies can be highlighted. To begin, we perform a detailed examination of various mobility features and pattern behavior analysis during three years. Second, our proposed research compares spatial and temporal mobility patterns as well as activity behavior using recent

Twitter data as a baseline (2018-2019). Furthermore, we evaluate the spatiotemporal variation in hot-spots and points of interest during the course of the study. Finally, we conduct an in-depth analysis of the link between verified positive coronavirus cases and social media users' daily mobility percentages.

To fully comprehend the study's findings and interpretation, it's essential to first consider some of the main steps and prohibitions implemented by the MOH in combating the spread of COVID-19 cases, as well as the spatial distribution of Kuwait's urban areas. Like much of the Arab countries, with primarily dry landscapes, Kuwait has the majority of its land filled by desert and its cities are found along the coastline. Kuwait's financial and commercial activities take place in the heart of the region, which is found in the administrative district in the governorate's capital.

According to a recent Twitter data in Kuwait, the peak time of activity occur at 6 p.m. and 7 p.m. on weekdays, and the most engaged area is Sharq city, which is situated in the capital region (Al-Asimah). In 2019, more than 1,400,000 Twitter data were generated, with January and November being the busiest months. In a previous analysis [1], we studied spatiotemporal data derived from Twitter and identified user movement and activity patterns. We monitored and discovered the general movement trend and activity of users during the day by identifying major hotspots. Noting that during 3 p.m. and 9 p.m., shopping centers, restaurants, and cafés are the primary destinations for user mobility, while user mobility peaks in suburban areas during the early morning and night hours.

III. DATA ACQUISITION AND PRE-PROCESSING

This section discusses the extraction technique used to obtain the required information and data, and the method of processing the data for examination. In addition to defining and characterizing the general features and information offered by the data source.

A. BIG DATA OF SOCIAL MEDIA

The massive growth of continuously generated data from social media over the past few years has resulted in a growing interest in the efficient means of collecting, analyzing, and querying the large volume of data generated every day. Social media data is considered as one of the primary sources of Big Data thanks to its widespread availability and massive size. Studies and researches have proved that social media data can be of great value in many fields, including business [13], public health [5], crisis management [9], and urban planning [26].

In this work, social media data, Twitter, is used to analyze social-geographic human mobility in Kuwait to track citizen's behavior and activity patterns in urban areas. This study aims to identify and investigate spatial-temporal human mobility patterns and behavior activity with the support of social media networks.

1) DATA ACQUISITION AND STUDY PERIOD

Geo-tagged tweets were extracted using Twitter Streaming API [30] with the use of python script. The captured data represented real-time location of Twitter users in Kuwait. The gathered data were saved as a JSON format, each pulled data record (post) included the following information: *'Tweet Text Language, Tweet ID, Date and Time of the Tweet, Tweet Text, Application URL, Hashtag, Coordinates, Place Name (geo-tagged place), Screen Name, Country, User ID, Place Type, and Photo URL'*.

The data retrieved were constrained by Kuwait's latitude and longitude coordinates. The extraction approach of social media data started from our previous work in late 2017 [1] till present. Currently, the database holds more than three years of fetched Twitter data (2018 – till present) with more than 3 million records. Each year of Twitter data is saved as a separate database. Since one of the main objectives of this study is to analyze the pandemic's impact on human behavior and understand its effects on social activity, data extracted from March till November was selected as the primary focus period of this study (as the virus started escalating in late February).

The data collected has been cleaned in order to find and delete missing, damaged and noisy data, to eliminate and avoid any mistakes in the study. The cleaning process began with the elimination of tweets from beyond Kuwait's geographical boundaries, tweets with missing latitude and longitude records (due to lost connection), and spam users. Spammers were discovered and located based on their consecutive geotagged posts. The pace and distance of two successive posts sent by the same person on the same day were the primary identifiers of spam users; users who seemed to be moving faster than 3.6km/h were marked as spammers.

2) DATA CHARACTERISTICS AND PRELIMINARIES

Data gathered for this study represent social media users' daily activity points during a nine-month duration (March – November) for the years 2018,2019, and 2020. Each data point was assigned a weekend or weekday classification depending on the date of its post. Moreover, each data point was assigned to its district and governorate area allocated by the point's location. Table 1 shows some main statistics and activity trends differences between the three databases (2018,2019,2020). As shown from the table, 2018 and 2019 data almost share similar activity trends (most tweeted hour, most tweeted day, and most tweeted area (city center)). Noticing that in 2018 the most engaged month is June, followed by April, and in 2019 data records, April was the most tweeted month, followed by October. Interestingly, October is the most tweeted month in 2020 data, even though the virus started spreading in March and escalating in June.

The activity trends of 2020 data are different from the previous years, starting with lower data points, fewer users, and different activity areas. Moreover, the most tweeted day is Monday. The most-tweeted hour is 8 p.m.; this is when

TABLE 1. Preliminary statistics and general activity trends of the gathered data.

	2018	2019	2020
Cleaned data	758,740	696,272	202,469
Total users	17,670	17,043	13,984
Most tweeted hour	6:00 p.m.	5:00 p.m.	8:00 p.m.
Most tweeted day	Thursday	Thursday	Monday
Most tweeted month	June	April	October
Weekday%	52.84%	51.45%	57.04%
Weekend%	47.16%	48.55%	42.91%
Most tweeted area	Sharq (city-center)	Sharq (city-center)	Qibla (city-center)

the weekly corona virus updates from MOH are usually announced at this time and day via TV and social media applications.

One of the key purposes of this research, as previously stated, is to establish the relationship between human mobility and COVID-19 spread in Kuwait. As illustrated in Fig. 1, a separate database was generated by extracting the required study time from the accessible COVID-19 cases. The database contains Twitter data collected between 06^{th} May 2020 to 15^{th} July 2020. About 6,700 individual users were detected over the course of the 70-day sampling period, and a total of 46,231 tweets were pulled, representing a median of 583 generated posts on a daily basis. extracted data were spread through the study area, with the capital governorate reporting the most, following that Hawalli and Al-Jahra governorates.

B. DATA PREPARATION AND PRE-PROCESSING

The primary component of the proposed human mobility detection scheme is mining individual user movement tracks from Twitter data for analysis and correlation with the spread of COVID-19 using only the geospatial data described in the coming sub-sections:

1) GEO-TAGGED TWEETS

A geo-tagged tweet is identified as the precise location of the mobile user at the moment the message was sent, as determined by the geographic location (x, y).

Each data record is characterized by a set of parameters describing the tweet. $Rn = (u, p, t)$, where u denotes the person who tweeted, p denotes the place from which u posted, and t denotes the time of the posted message. We identified a collection of point-locations from which messages were sent, reflecting a list of places visited by mobile users, from the geo-tagged posts collected for the analysis.

Each user can frequently visit multiple locations within a given area; for each user, we join their individual activity tracks, referred to as trajectories:

2) TRAJECTORIES

A trajectory is described as the spatial-temporal order of a user's visited places over the course of a single day. Three data sets based on user movements were developed. The first dataset is a row matrix containing information about each

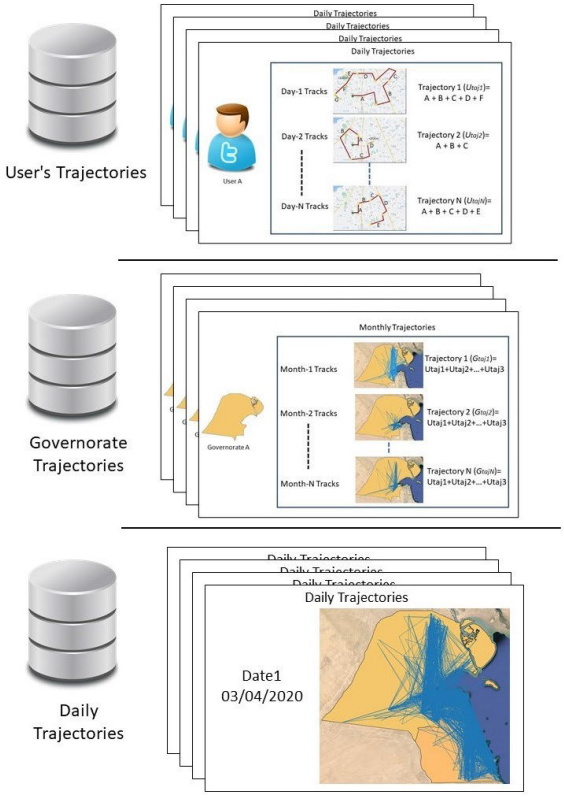


FIGURE 2. Data sets used to derive spatio-temporal movements of Twitter users.

user's movement over the given time span. The second dataset collection contains overall users visited places inside a governorate. Finally, as shown in Fig. 2, daily trajectories are stored in the last-row of the matrix.

To prevent any disruptions when analyzing users' movements, which does not correspond to human activity in specific aspects, gathered data must be aggregated to reflect accurate user mobility in real world scenarios.

3) AGGREGATION

Data aggregation relates to the merger and combination phase of sets of trajectories record that sharers some conditions and displaying them as a single data record.

The term ''aggregated data'' refers to data that is shared on the exact same day by the same Twitter user in terms of time

TABLE 2. Time slot separation details.

Time Slot	Time Range
Early-morning	00:00 – 08:00
Morning	08:00 – 12:00
Evening	12:00 – 16:00
Late-evening	16:00 – 21:00
Night	21:00 – 00:00

and space. When two or more data records (Ri, $R\,j$) overlap, the following requirements are met:

- $u_i \neq u_j$
- Distance between p_i and $p_j \leq 300m$
- Time difference between t_i and $t_j \leq 1$ hour
- Both R_i and R_j are tweeted the same day

following that, mobility was created by connecting individual user trajectories. Then, a sequence of time-series mobility was constructed by linking each data record in each dataset, covering every day movement of individual users, monthly movement of each region, as well as overall mobility of each user.

IV. MOBILITY PATTERN IDENTIFICATION SCHEME

The proposed mobility pattern detection scheme consists of three main phases. First, data will be divided into five time slots representing different times thought-out the full day. Second, a clustering algorithm will be applied to each time slot for each data set to group homogeneous data points. And finally, mining and extracting the general mobility pattern and the identification of social activities.

A. TIME SLOT SEPARATION

Mobility pattern recognition is based on the premise that user activities and daily living are not the same, but it changes over time. The first step in identifying space-time activity patterns is to split the 24 hours of the day into different time slots. After the data cleaning process, each database is divided into five data-sets representing each time slot. Table 2 summarizes the specifics of each time slot.

B. CLUSTERING ALGORITHM

The following step in identifying mobility patterns is to apply a clustering algorithm to the five data-sets in each database. The clustering algorithm technique is the most popular technique used in data science, specifically in the data mining field. It aims to divide similar data points into groups or clusters in large data-sets. Data points that are closed to each other are positioned in the same group, where each group contains more similar data points to each other (greatest similarities) than to those in the other groups (greatest dissimilarities). The clustering algorithm aims to cluster homogeneous data points to identify and extract social activities and points of interest in each database.

In this work, a comparison between two of the most common and most popular partitional clustering techniques in the

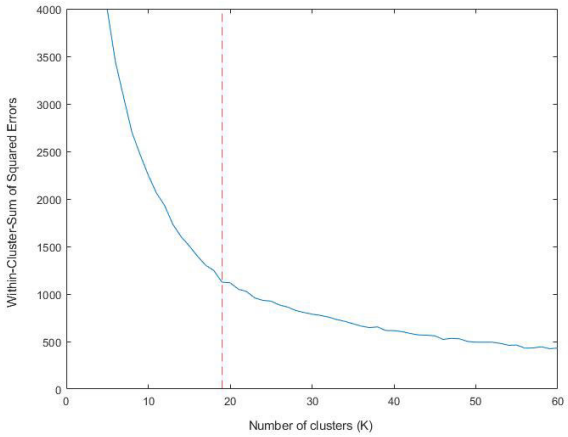


FIGURE 3. Elbow method showing the optimal number of clusters for the morning slot (2018 database).

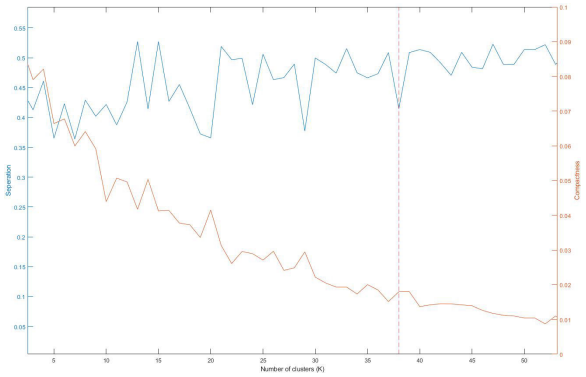


FIGURE 4. Optimal number of clusters for the morning slot (2018 database).

data mining field are performed on each data-set, K-means and Fuzzy clustering algorithms. While, Silhouette index (SI), Davies–Bouldin index (DB), and Calinski–Harabasz index (CH) are used as validity indices assessment in comparing between the two clustering algorithms.

1) K-MEANS CLUSTERING ALGORITHM

The global k-means clustering algorithm is the most straightforward unsupervised learning algorithm [16]. It is an iterative machine learning technique that partitions the unlabeled data-set into non-overlapping clusters. It starts by placing centroids (K-pre-defined) in random locations in space. Then, iteratively finds and assigns the nearest centroid for each point by calculating the distance (e.g., Euclidean distance) between the point and the centroids. Finally, it moves each centroid to the mean of points assigned to it. The algorithm continues iterating until no points are changing its assigned centroid cluster.

The ''choosing K'' step is the crucial element in any unsupervised clustering algorithm; it improves the efficiency and effectiveness of the clustering performance in processing a large amount of data. The Elbow method is the most well-known method for determining the optimal number of clusters (K) for the k-means. It calculates the

within-cluster-sum-of-errors for different values of K and chooses the K for which the sum of squared errors starts to minimize (visible as an elbow).

The elbow method is applied for each database by running the k-means clustering on each data-set for a range of values from 5 to 60; the Euclidean method was used as the distance measure between data points. The plot in Fig. 3 illustrates the method's result for the 2018 database (morning data-set). The plot looks like an arm with an apparent elbow at K = 19. Some of the data-sets were ambiguous, not showing a clear elbow curve. Therefore, the Intra-cluster cohesion and Inter-cluster separation were used instead to determine the optimal value of clusters (K).

2) FUZZY CLUSTERING ALGORITHM (FCM)

The idea behind any clustering approach is to uncover the structure or specific groups within the data set. A fuzzy clustering algorithm [2] recognizes that clusters of data points are not always perfectly separated, and data points may belong to one or more groups. Thus, the algorithm works by assigning membership degrees to each data point between 0 and 1. Membership indicates the degree to which a data-point record is associated with a particular cluster. Points that are further away from the middle of the cluster are affiliated with a low degree membership. Cluster centers iteratively change position to the correct location within the data points by minimizing the objective function, represented by the Euclidean distance of any data point to the center of the cluster weighted by the point's membership degree to the center.

The optimal number of clusters (K) for each database was determined based on two parameters: The Intra-cluster cohesion (compactness) parameter and the Inter-cluster separation parameter. Compactness or cluster cohesion measures how close the points within the same cluster (average distance between within data points). Cluster separation measures how well-separated cluster centers are from each other. Clustering is efficient if the compactness between within data points is minimized and separation between clusters centroid is maximized. Fig. 4 illustrates the cluster validation result for the morning slot (2018 database), where the optimal number of clusters (K) is set to 38.

3) VALIDITY INDICES COMPARISON

Determining which clustering algorithm is best suited for this research, three validity indices were used to compare between the two algorithms.

- Silhouette index (*SI*)
  The silhouette index [17] computes each data point's width value in any cluster based on its membership degree. The width is a value measure of how similar data point is to its own cluster compared to other clusters. It ranges between [-1,1], where values close to 1 indicate a well-matched point to its cluster—a higher value of the index results in a better clustering of the data points.
- Davies-Bouldin index (*DB*)
  The DB index [6] is defined as the ratio between the within-cluster compactness and between-cluster separation. DB's objective is to obtain clusters with the lowest average distance between within-points and the maximum average distance between centroids of clusters. Lower values of DB indicate better clustering results.
- Calinski-Harabasz index (*CH*)
  The index [4] evaluates the clustering result based on two matrices, the cluster scatters matrix (BSCM) and the within-cluster scatters matrix (WCSM). The object of the index is to obtain a maximum BSCM and a minimum WCSM. The CH value indicates the best suitable clustering result when the ratio between the two matrices is maximized.

Fig. 5 is the result of the three validity indices evaluation on each data-set for each database after applying the FCM and k-means clustering algorithm. The clustering evaluation result concluded that the FCM algorithm is best suitable as a clustering method to be used in the three databases (2018,2019,2020).

TABLE 3. FCM details in each database.

Year	# of Slots	# of Clusters	# of Data	% of Highly active users
2018	slot 1	33	89,181	10%
	slot 2	38	94,809	8%
	slot 3	38	147,332	13%
	slot 4	40	365,227	20%
	slot 5	34	62,191	14%
2019	slot 1	36	78,777	7%
	slot 2	38	78,463	9%
	slot 3	40	125,364	10%
	slot 4	44	279,852	13%
	slot 5	37	33,816	11%
2020	slot 1	39	19,569	16%
	slot 2	41	20,705	19%
	slot 3	45	28,317	14%
	slot 4	48	60,422	8%
	slot 5	39	29,340	9%

The primary step in extracting the mobility pattern was to cluster similar data points in each timeslot by applying the fuzzy clustering algorithm. Table 3 shows each time slot cluster's details in each database after applying the FCM algorithm, including percentages of highly active users. Observing from the table that the most number of clusters are always produced in the fourth time slot in each database. Moreover, the percentage of highly active users is usually within the mid-day to end-day hours in ordinary circumstances, but in 2020 we notice that highly active users are more involved in the early morning to mid-day hours.

C. MOBILITY IDENTIFICATION AND RECOGNITION

Extracting and identifying the mobility pattern of social media users is achieved through three main tasks. First, classifying each slot's activity and determining for each user the most suitable cluster in each time slot. Second, extraction of individual flow movement and overall mobility patterns of

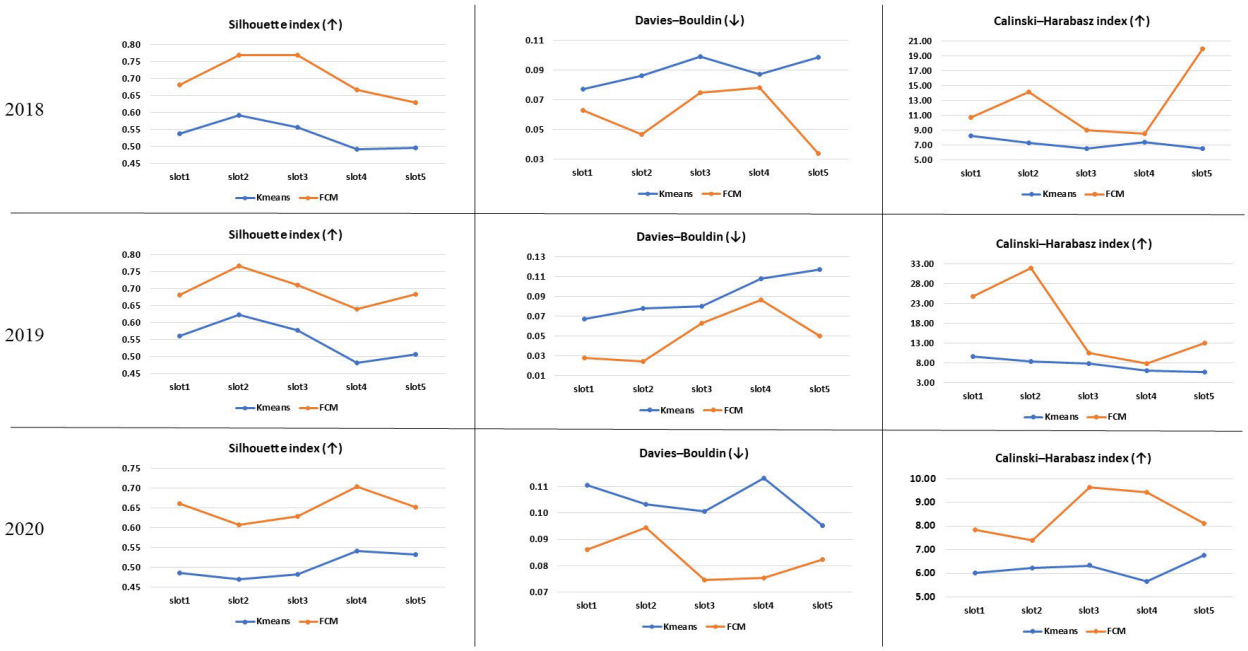


FIGURE 5. Validity indices evaluation on each database.

active social media users in each database. And lastly, the identification of hotspots and social activity of social media users throughout the full day.

1) CLASSIFICATION OF USER's ACTIVITY

The main point of classifying users based on their activity level is to find each user's representative cluster in each timeslot. Each user is classified based on their activity level into three classes: High active, medium active, and low active. For the purpose of extracting the general mobility pattern of social media users, high active users were only considered in this task. For each user in each time slot, each cluster's activity percentage is calculated in the timeslot. The cluster with the highest activity level rate is selected as the user's most representative cluster in that timeslot, meaning that the user tends to tweet more in that specific cluster area.

2) MOBILITY EXTRACTION

After assigning each user the most representative cluster in each time slot, each user ends up with five variables representing individual flow movement through the day (five timeslots). Table 4 demonstrates a sample of the individual mobility flow of users. Illustrating more on individual mobility flow, user3 in Table 4 tends to tweet more in cluster A1 in the early morning slot than the rest of the other clusters in that timeslot. If there are no data for a user in one of the timeslots, a null value will represent the variable in that timeslot, similarly with user4 in the night slot.

The general movement pattern is achieved by an examination of the flow movement from single slot to the next in the resulting human mobility flow table (e.g., third to forth slot).First, total number of occurrences for each cluster is determined in each timeslot. Then, the cluster with the highest count is selected to represent the timeslot. And finally, the percentage of users (in each represented cluster) flowing from one timeslot to another is calculated.

TABLE 4. Sample of the individual mobility flow of users between time-slots.

Users	Early-morning	Morning	Evening	Late-evening	Night
User-1	A1	A2	C3	D4	C5
User-2	D1	B2	B3	A4	L5
User-3	A1	E2	L3	A4	C5
User-4	B1	A2	L3	A4	—

3) HOTSPOTS IDENTIFICATION

A density heat-map is created for each timeslot (representative cluster) to identify hotspots and points of interest. The main idea of the heat-map analysis is to understand and visualize users' behavior and social activity in each timeslot. Each identified POI falls into the five different categories explained earlier. The identified POI gives a precise visualization of social activity and behavior throughout the full day.

D. SPATIO-TEMPORAL ANALYSIS OF HUMAN MOBILITY

The extraction and identification of mobility patterns are essential in understanding how people behave in an urban environment, especially for detecting anomalies. The data gathered for this study went through different processes and phases to detect individual and general mobility patterns of social media users in Kuwait. One of this study's core

objectives was to identify and analyze users' behavior during the current pandemic by comparing the recent extracted mobility pattern against mobility patterns mined from a non-crisis situation (pre-pandemic). Fig. 6 illustrates the extracted mobility pattern from the years 2018, 2019, and 2020. The figure shows that 2018 and 2019 have almost similar mobility patterns, indicating that humans follow simple reproducible patterns. While the 2020 pattern shows some different activity behavior due to the current pandemic, demonstrating the value of social network data in detecting anomalies in human mobility and activity behavior. The main comparison points between the three mobility patterns are as follows:

- Residential areas in the early morning slots (2020) have increased in a noticeable percentage than in previous years. The main reason for the increased percentage goes to the government reducing the number of employees to 30% in workplaces and giving off-work to workers with reduced immunity. Moreover, residential areas are mainly in the top 3 of user's activity in the rest of the slots, indicating that most of the time, users are staying home and committed to going out only when necessary.
- Hospitals and clinics have increased as hotspots in the year 2020, where more than 43% of data points are categorized as others. Noticing in the morning slot, hospitals and clinics represent 40% of the morning hours' activities.
- The interest in restaurants and cafes remains the same as a daily activity in all three years.

V. IMPACT OF A GLOBAL PANDEMIC ON HUMAN MOBILITY

Following the initial outbreak of COVID-19, researchers and scientists have made several attempts to study various causes and variables in order to gain timely knowledge on the other aspects of the corona-dissemination. Human movement is a key element in the geographical dissemination and outbreak of infectious epidemic such as the corona-virus. Understanding and studying human movement during a pandemic will help researchers and decision makers to further understand the implications and consequences of mobility constraints on COVID-19 distributed in communities and nations.

Analyzing how individuals travel and respond during an outbreak is critical, not just for tracking the efficacy of lock-down and restraint measures, but also for studying and comprehending the connection between human mobility and the transmission of COVID-19. This segment examines the effect of corona-virus containment on users' social mobility in Kuwait's regional areas during complete lockdown and curfew periods. Beginning with an examination of the detailed mobility data derived from social network apps, we would attempt to examine and deduce the relationship between COVID-19 spreads and everyday human mobility. Additionally, we will analyze and evaluate the effect on human activity and movement patterns of the restrictions and warning actions implemented by the MOH to avert the corona-virus epidemic. This is accomplished by analyzing several factors, including variations in regular mobility ratios, and studying changes in Points of Interest (POI) by comparing the findings to related details in standard circumstances (i.e., pre-pandemic data). Table 5 outlines the three stages that comprise the study's primary focus time.

The remainder of this section is organized as follows. First, a description of COVID-19 data characteristics and the obtained social media data. Second, Mobility habits and behaviour during the full lock-down and partial curfew. Third, the impact of COVID-19 on communities and human behavior. And finally, examining the relationship between the corona-virus spread and human mobility.

A. DATA CHARACTERISTICS

The publicly accessible dataset on corona-virus for the state of Kuwait was acquired from the official ministry of health (MOH) corona website [22]. The official website offers daily reports on corona-virus updates. The obtained COVID-19 data information included: date, daily positive cases, and cumulative reported cases. For a short period of time, MOH offered detailed information regarding the breakdown of positive cases in each of the five health regions in Kuwait, from 06^{th} May 2020 to 15^{th} July 2020. Generally speaking, well functioning health sectors(areas) are characterized as being concerned with routine medical treatment in only one governorate region. With the exception of the Al-Ahmadi health sector, linking two governorates regions (Al-Ahmadi and Mubarak Al-Kabeer governorates), each of Kuwait's four governorates region has its own health area. Through the rest of this paper, COVID-19 data positive cases retrieved from 06^{th} May 2020 to 15^{th} July 2020 period will only be considered for the analysis, as it's the only period with a detailed breakdown of corona-virus cases that were publicly available for researchers.

A total of 50,311 positive cases of the corona-virus was reported from 06^{th} May 2020 to 15^{th} July 2020. Details of reported cases in each health area are shown in Table 6. Daily numbers of positive cases in each health area are illustrated in Fig. 7. As the figure indicates, the virus proceeded to propagate daily, reaching a height on $19th$ of May 2020 in Al-Farwaniya with 397 new cases. Then gradually reduce to a level that permitted the start of the second partial curfew, and gradually increase again. More details on this drop are discussed in the upcoming sections.

B. MOBILITY RATIO MEASUREMENTS

The data collection used in this analysis is a four-month series of regular movement travels of Twitter users across Kuwait's borders, beginning on the $13th$ of April 2020 and finishing on the $10th$ of July 2020. Over 40,231 tweets were generated in over 7,198 separate locations with 6,734 apps users and a total of 14,284 track movements.

By examining the mobility ratio in greater details for each health sector, Fig. 8 shows the changes in regular rate

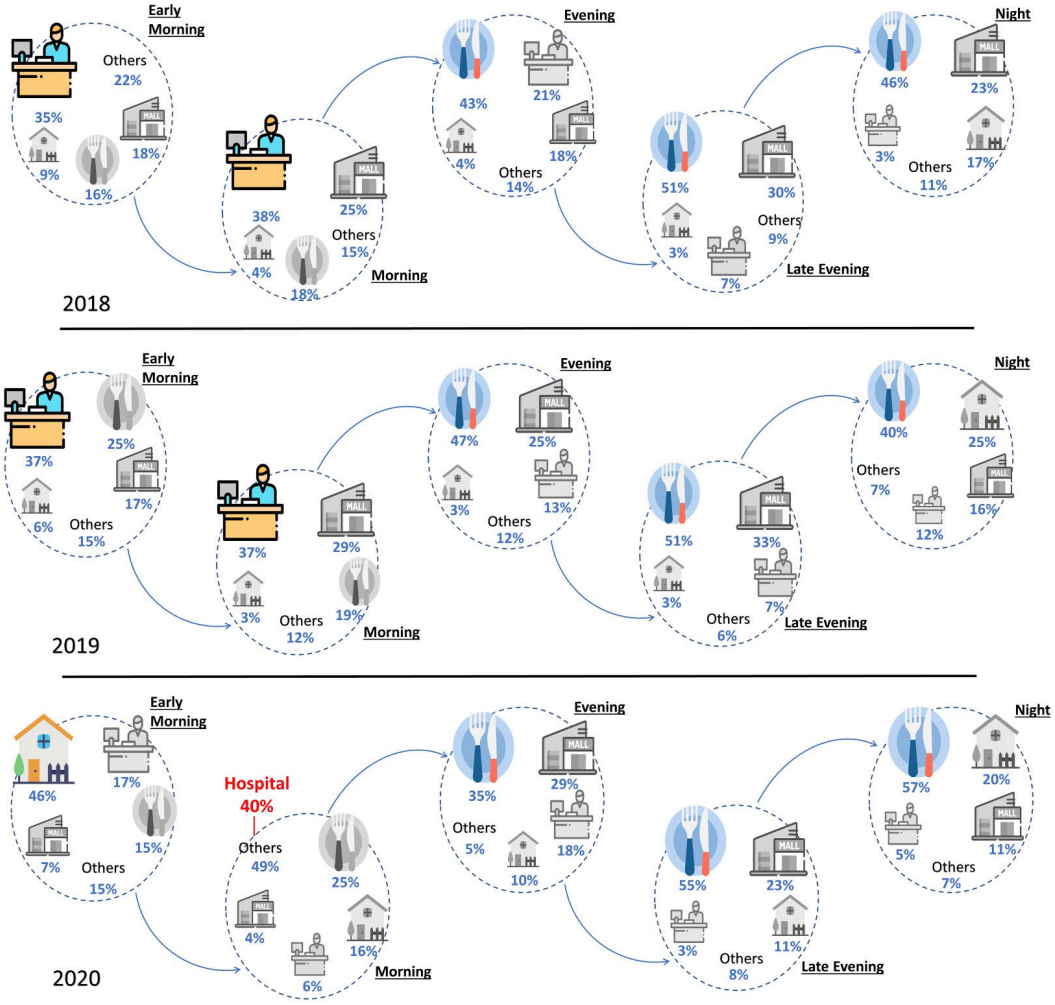


FIGURE 6. Mobility pattern and activity of social media users through-out the full day.

TABLE 5. Details on partial curfews and complete lock-downs.

	1^{st}Partial curfew	Full lock-down	2^{st}Partial curfew
Period	18^{th} April to 9^{th} May	10^{th} May to 30^{th} May	31^{st} May to 20^{th} June
Duration	22 days	21 days	22 days
Open Hours	5 p.m. to 4 a.m.	-	6 p.m. to 6 a.m.

TABLE 6. Reported positive cases of corona-virus data in each health sector.

Health areas	Total positive cases
Al-Asimah	5,332
Hawalli	8,117
Al-Farwaniya	14,717
Al-Ahmadi	12,905
Al-Jahra	9,240

of mobility in each health sector. As seen in the graph, COVID-19 had a significant impact on users' movement tracks, resulting in a 30-90% decrease in mobility ratio compared to regular days. The health sectors of Al-Farwaniya, Al-Asmiah, Hawalli, and Al-Ahmadi all display a nearly identical mobility ratio shift. Al-Jahra, on the other hand, has a peculiar mobility habit than the rest of the governorates owing to different user activity, distinct behaviors, and unusual mobility habits. This, though, is outside the reach of this article.

C. MOBILITY DURING FULL LOCK-DOWN AND PARTIAL CURFEW

One of the primary goals of this research is to investigate human activity during the current pandemic, especially during lockdowns and partial curfews. The data was split into

TABLE 7. Data-sets details.

	Duration	Total data	Total users
1^{st} partial curfew	22 days	11,299	3,145
Full lock-down	21 days	7,646	2,061
2^{nd} partial curfew	22 days	21,008	3,240

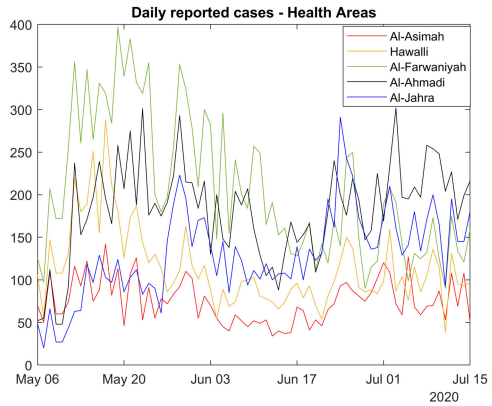


FIGURE 7. Day-to-day cases of corona-virus in each health sector.

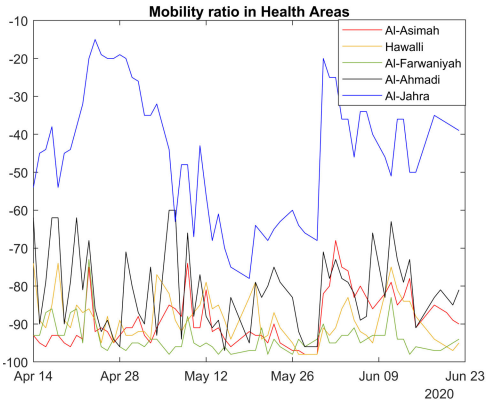


FIGURE 8. Day-to-day mobility ratio in each health sector.

three data sets to reflect the three stages of MOH's restriction steps. Table 7 contains detailed information about each data collection.

To fully understand users' behavior and change of mobility during the lock-down and partial curfew, users were divided into active (mobile) and inactive (non-mobile) users depending on their degree of activity (less-mobile). The following criteria are used to classify social media users: 1) The individual tweets every two days or so, and 2) the interval between successive tweets is greater than 300m. Only data from mobile users is used to calculate the mobility ratio between the three stages, since this reflects real mobility on everyday basis. Fig. 9 shows how the user's movements changes over the course of the three stages. According to the results, mobility began to decline in Full lock-down, showing the dedication of social media users to the complete ban and only going-out when absolutely required. Furthermore,

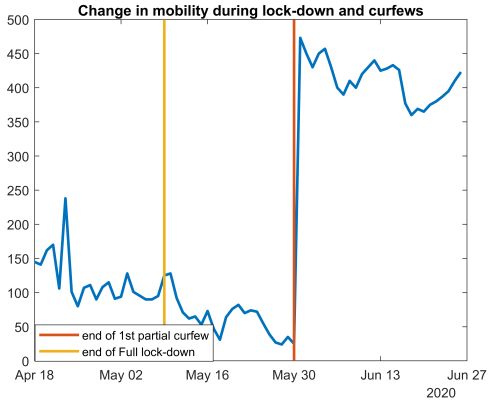


FIGURE 9. Changes in mobility patterns as a result of complete lockdown and partial curfews.

as predicted, mobility improves dramatically after removing the complete lock-down, with a 52% improvement in mobility. The improved versatility rate corresponds to the opening of shopping centers and restaurants.

D. IMPACT ON COMMUNITY AND MOBILITY HABITS

We studied the activity of users during the pandemic by finding points of concern in data sets using heat maps, in order to see the effect of COVID-19 on society and human behaviour. The points of interest listed are classified into five categories: job and research zones, shopping centers, cafes and restaurants, and suburban districts. The study showed that during the complete lock-down, point of interest in suburban districts increased to more than 3%, implying that users adhered to the lock-down and remained at home, while after entering the 2^{nd} conditional curfew, POI in cafes and restaurants increased by 18%. To get a clearer understanding of how human behaviour and mobility patterns have changed over time (current pandemic, 2019 Twitter data), we compared two density heat-maps that covered the very same geographic region during the study timeframe (April to July) using a change detection technique.

Fig. 10 illustrates the outcomes of contrasting heat maps of the collected data for the years 2020 and 2019. It depicts COVID-19's impact on mobility trends and activity behaviors. Red zones have more than normal usage activities and interactions, while blue zones have fewer than usual behavior. As its obvious, we see more users in suburban areas and hospitals than we do in usual situations, and less users in shopping centers, which used to be the hubs of user interaction in regular days.

E. RELATIONSHIP BETWEEN HUMAN MOBILITY AND THE SPREAD OF COVID-19

To explain the relationship between activity and positive COVID cases, we must consider the incubation period as well as the interval between taking the test. Given that studies measure the average reporting delay of the incubation period as 12 days with a median of 5 days [19], the amount of new confirmed corona-virus cases in a single day would be

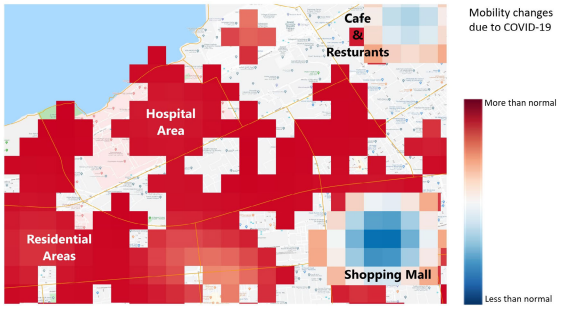


FIGURE 10. Changes in user activity in visited POI (heatmap difference between 2019 and 2020).

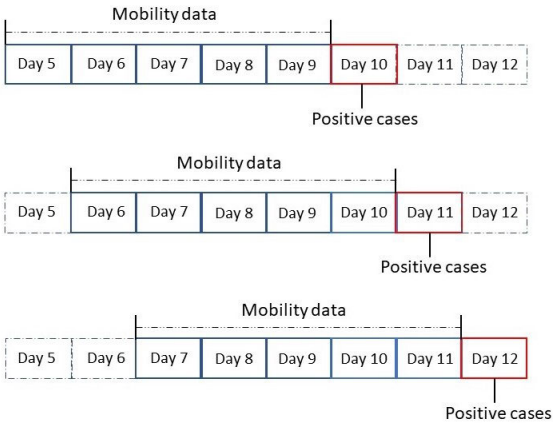


FIGURE 11. Sliding time window.

proportional to the cumulative mobility of users who made more than a single movement track "5" days prior. As shown in Fig. 11, the sliding time window method was used to transfer through mobility data according to a fixed interval (5 days back) in order to compare mobility with confirmed cases within the given time span. Our analyzes showed a strong association between mobile data derived from Twitter and the positive corona-virus cases in the study area, and a coefficient of correlation greater than 0,6.

VI. CHALLENGES AND LIMITATION

Using social media gives you a massive data haul on human traits,behaviour and trends. Many researchers have used the data in various fields such as crowd control, smart city design, anomaly detection, and more. In this article, we have examined the latest global pandemic epidemic through the lens of social network data in an attempt to define and compare the association between human mobility and the transmission of COVID-19. Several problems and concerns remain, however, in the investigation of human mobility and its effects on COVID-19 outbreaks.

To begin, the corona-virus data used in this study were the publicly available information provided by the Ministry of Health; comprehensive details on corona-virus in Kuwait are not yet available to the public. Additional information of the ethnicity, sex, home and near contacts of the infected person could help to predict the spread of corona-virus in particular areas. Furthermore, it is important to continue investigating the virus's dissemination in the event of a second wave, as the world has seen a second sharp increase in the amount of affected cases in recent weeks.

Second, variables correlated with corona-virus should not be restricted to mobility data alone; additional parameters, such as population density, air temperature, and literacy rates, should be studied and assumed to be involved in causing the virus's dissemination. At the moment, those details are open to the media, but not at in depth level.

Last but not least, the connectivity data obtained from 2018 and 2019 social network application could not represent mobility trends and activity in 2020. In other words, evaluating past historic mobility in a crisis or non-crisis situations may prove fruitful, especially for identifying and evaluating anomalies in human mobility. As a result, the need to combine various data sources enables the provision of reliable and detailed knowledge regarding human trajectories.

As additional data sets become available and our understanding of the geographic distribution of COVID-19 improves, social network data can play a crucial role in measuring the behavior and movement habits of individuals and communities. Researchers and public health officials can expand their field in human mobility by observing mobility across spatio-temporal scales of social networks data, providing an additional lens to identifying outbreaks and understanding how people react during a pandemic.

VII. CONCLUSION

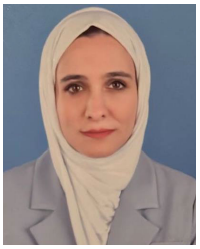
A broad deployment of human mobility is imperative for the advancement of disease detection and forecasting as well as for communication networks. Until now, the majority of analysis on human movements and travel has relied on private and aggregated records, such as mobile phone data. In this study, social network data were suggested as a proxy for human mobility since it is based on large amount of open and accessible data. Using the most recent mobility data derived from social media users, we monitored and evaluated users' regional movement and its ratio during full lockdowns and partial curfews in Kuwait, demonstrating how user mobility is associated with positive COVID-19 records. Furthermore, we analyzed and comprehended the impact of COVID-19 on human behavior and activity patterns in societies and metropolitan areas. Additionally, we investigated and examined social network users' spatiotemporal mobility trends and behavior activity development from 2018 to 2020. Numerous years of mobility data analysis may reveal well-established trends, such as social or cultural events, that serve as a measure for detecting anomalies and changes in human mobility.

One potential future study area is fusing and combining data from various sources, including different social networks applications, cell trackers, and other sensory gadgets, to gain a high-level understanding of human mobility. In addition to incorporating contextual features to the proposed mobility

detection scheme to be used for sentiment analysis. Additionally, numerous directions for potential research area are feasible, such as the creation of a user-based mobility risk framework, to validate the flow movements and to deepen the spatio-temporal distribution of the corona-virus. Finally, developing a forecasting model for a multi-dimensional mobility pattern of user movements and travel routes.

REFERENCES

[1] M. Al-Jeri, "Towards human mobility detection scheme for location-based social network," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1–7.
[2] J. C. Bezdek, W. Full, and R. Ehrlich, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.
[3] D. Bisanzio, M. U. Kraemer, I. I. Bogoch, T. Brewer, J. S. Brownstein, and R. Reithinger, "Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale," *Geospatial Health*, vol. 15, no. 1, pp. 1–6, 2020.
[4] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[5] Q. Chen, C. Min, W. Zhang, G. Wang, X. Ma, and R. Evans, "Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis," *Comput. Hum. Behav.*, vol. 110, Sep. 2020, Art. no. 106380.
[6] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[7] Z. Ebrahimpour, W. Wan, O. Cervantes, T. Luo, and H. Ullah, "Comparison of main approaches for extracting behavior features from crowd flow analysis," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 10, p. 440, 2019.
[8] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse, "Mapping county-level mobility pattern changes in the United States in response to COVID-19," *SIGSpatial Special*, vol. 12, no. 1, pp. 16–26, 2020.
[9] A. G. Gulnerman, H. Karaman, and A. Basiri, "New age of crisis management with social media," in *Open Source Geospatial Science for Urban Studies*. Cham, Switzerland: Springer, 2021, pp. 131–160.
[10] G. M. Hadjidemetriou, M. Sasidharan, G. Kouyialis, and A. K. Parlikad, "The impact of government measures and human mobility trend on COVID-19 related deaths in the U.K.," *Transp. Res. Interdiscipl. Perspect.*, vol. 6, Jul. 2020, Art. no. 100167.
[11] X. Huang, Z. Li, Y. Jiang, X. Li, and D. Porter, "Twitter, human mobility, and COVID-19," 2020, *arXiv:2007.01100*.
[12] Z. Huang, X. Ling, P. Wang, F. Zhang, Y. Mao, T. Lin, and F. Y. G. Wang, "Modeling real-time human mobility based on mobile phone and transportation data fusion," *Transp. Res. C, Emerg. Technol.*, vol. 96, pp. 251–269, Nov. 2018.
[13] J. Jacobson, A. Gruzd, and Á. Hernández-García, "Social media marketing: Who is watching the watchers?" *J. Retailing Consum. Services*, vol. 53, Mar. 2020, Art. no. 101774.
[14] S. Jiang, J. Ferreira, and M. C. González, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017.
[15] C. Kadar and I. Pletikosa, "Mining large-scale human mobility data for long-term crime prediction," *EPJ Data Sci.*, vol. 7, pp. 1–27, Dec. 2018.
[16] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[17] L. Kaufman and P. J. Rousseeuw, "*Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344. Hoboken, NJ, USA: Wiley, 2009.
[18] A. Khatua, T. K. Kar, S. K. Nandi, S. Jana, and Y. Kang, "Impact of human mobility on the transmission dynamics of infectious diseases," *Energy, Ecology Environ.*, vol. 5, pp. 389–406, 2020.
[19] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application," *Ann. Internal Med.*, vol. 172, no. 9, pp. 577–582, 2020.
[20] W. Li, S. Wang, X. Zhang, Q. Jia, and Y. Tian, "Understanding intra-urban human mobility through an exploratory spatiotemporal analysis of bike-sharing trajectories," *Int. J. Geograph. Inf. Sci.*, vol. 34, no. 12, pp. 2451–2474, 2020.
[21] Z. Liu, Z. Li, K. Wu, and M. Li, "Urban traffic prediction from mobility data using deep learning," *IEEE Netw.*, vol. 32, no. 4, pp. 40–46, Aug. 2018.
[22] Kuwait Ministry of Health. (2020). *Kuwait Corona-Virus Updates*. [Online]. Available: https://corona.e.gov.kw/
[23] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, 2021.
[24] L. Qin, Q. Sun, Y. Wang, K.-F. Wu, M. Chen, B.-C. Shia, and S.-Y. Wu, "Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index," *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, p. 2365, 2020.
[25] N. W. Ruktanonchai, C. W. Ruktanonchai, J. R. Floyd, and A. J. Tatem, "Using Google location history data to quantify fine-scale human mobility," *Int. J. health Geograph.*, vol. 17, no. 1, pp. 1–13, 2018.
[26] Z. Shao, N. S. Sumari, A. Portnov, F. Ujoh, W. Musakwa, and P. J. Mandela, "Urban sprawl and its impact on sustainable urban development: A combination of remote sensing and social media data," *Geo-Spatial Inf. Sci.*, vol. 24, no. 2, pp. 1–15, 2020.
[27] A. Silver and J. Andrey, "Public attention to extreme weather as reflected by social media activity," *J. Contingencies Crisis Manage.*, vol. 27, no. 4, pp. 346–358, 2019.
[28] Statista. (2021). *Social Media Usage Statistics*. [Online]. Available: https://backlinko.com/social-mediausers
[29] M. W. Traunmueller, N. Johnson, A. Malik, and C. E. Kontokosta, "Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities," *Comput., Environ. Urban Syst.*, vol. 72, pp. 4–12, Nov. 2018.
[30] Twitter. (2020). *Streaming Tweets in Real-Time*. [Online]. Available: https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time
[31] H. Ullah, W. Wan, S. A. Haidery, N. U. Khan, Z. Ebrahimpour, and T. Luo, "Analyzing the spatiotemporal patterns in green spaces for urban studies using location-based social media data," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 11, p. 506, 2019.
[32] P. Xu, M. Dredze, and D. A. Broniatowski, "The Twitter social mobility index: Measuring social distancing practices from geolocated tweets," 2020, *arXiv:2004.02397*.
[33] Y. Yang, A. Heppenstall, A. Turner, and A. Comber, "Who, where, why and when? Using smart card and social media data to understand urban mobility," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 6, p. 271, 2019.
[34] Y. Zhou, R. Xu, D. Hu, Y. Yue, Q. Li, and J. Xia, "Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: A modelling study using mobile phone data," *Lancet Digit. Health*, vol. 2, no. 8, pp. e417–e424, 2020.



MUNAIRAH ALJERI (Member, IEEE) received the B.S. degree in computer science from Gulf University for Science and Technology (GUST), Kuwait, in 2011.

Since 2014, she has been working as a Senior Research Associate with the Kuwait Institute for Scientific Research (KISR), Kuwait. Her research interests include big data analysis, prediction models, data mining techniques, mobility pattern analysis, and GIS application development.

Mrs. Aljeri awards include two Best Paper Awards in Proceedings of the 18th ACM Symposium on Mobility Management and Wireless Access, and IEEE Symposium on Computers and Communications.