CHEMICAL ENGINEERING TRANSACTIONS

VOL. 97, 2022

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Jeng Shiun Lim, Nor Alafiza Yunus, Jiří Jaromír Klemeš
Copyright © 2022, AIDIC Servizi S.r.l.
ISBN 978-88-95608-96-9; **ISSN** 2283-9216

DOI: 10.3303/CET2297020

The Relationship between Green Transportation and Leisure Travel Based on Social Media Data

Juhyeon Kwak[a], Joonsik Jo[a], Donggyun Ku[b], Seungjae Lee[b,*]

[a]Department of Transportation Engineering / Department of Smart Cities, University of Seoul, Korea
[b]Department of Transportation Engineering, University of Seoul, Korea
sjlee@uos.ac.kr

In recent years, the number of people embarking on leisure-travel trips has increased rapidly. Consequently, travel patterns have changed depending on the type of leisure trip. However, accurately establishing non-daily travel data using conventional travel diary surveys with the existing self-survey method is impossible. A possible solution is the use of social media, which is widely used, as a data source. Users upload various posts describing their leisure travels on social media, which include both structured data, such as posting date and location information, and unstructured data, such as content and images. The purpose of this study was to analyze the relationship between individual behavior and green transportation during leisure travel using location-based social media data. Social media data were collected using a web crawler and annual transportation data were collected from the green transportation promotion areas in Seoul, Korea. A text mining technique was applied to the content of the posts to analyze the types of leisure for each individual. The text data in the contents were classified in detail by applying latent Dirichlet allocation (LDA), a topic modeling technique. Subsequently, to identify individual travel behavior, spatial analysis was performed by comparing the location information of the post with the usage of various transportation by leisure type. Consequently, four leisure types, i.e., exercise, tourism, rest, and social, were identified to be prevalent in the northern, central, southern, and eastern regions. Social and tourism-type leisure activities involved the use of shared bicycles and public transport. In contrast, people pursuing exercise-type leisure activities tended not to use shared bicycles and public transportation. The results of this study indicate that there is a correlation between activity type and choice of transportation. Identifying the purpose of travel can help strengthen the connection with green transportation.

1. Introduction

Automobiles emit large amounts of air pollutants, and the global warming caused by greenhouse gas emissions from vehicles has become a serious social problem. Integrating sustainability into transportation modes has become a worldwide target, especially with the sharp increase in the urban population and trips (Bencekri et al., 2021). Accordingly, various studies have been conducted in the field of green transportation. Ku et al. (2021) analyzed the benefits of green transportation reflecting the wider impact based on the origin/destination data of shared bicycle users. Choi et al. (2021) demonstrated that personal mobility (PM) enables faster and greener movements in complex downtown Seoul. In several cities, eco-friendly policies have been introduced, and the number of green transportation facilities is increasing. The characteristics of urban space and travel behavior are closely related and influence each other. Spatial characteristics result in certain modes of transportation dominating. Moreover, the preference for a certain mode of transportation in certain areas affects the area itself. For instance, in India, the poor depend heavily on non-motorized transportation, such as walking and cycling, as their primary mode of travel. This could reduce their employment potential in cities concentrated in the Central Business District (Srinivasan and Rogers, 2005). Similarly, Lee et al. (2014) found that urban baby boomers make more recreational non-motorized transport and social, utilitarian, and transit commuting trips. Most of these differences seem to be primarily a result of the urban setting and not the particular preferences of boomers living in urban settings. Kwak et al. (2021) proposed that the government could produce more effective results by considering local characteristics to revitalize eco-friendly transportation. In other words, if the regional

Paper Received: 31 May 2022; Revised: 10 August 2022; Accepted: 17 September 2022
Please cite this article as: Kwak J., Jo J., Ku D., Lee S., 2022, The Relationship between Green Transportation and Leisure Travel Based on Social Media Data, Chemical Engineering Transactions, 97, 115-120 DOI:10.3303/CET2297020

characteristics and travel behavior of the region are identified, the connection between eco-friendly means of transportation can be strengthened and traffic volume using eco-friendly means can be increased.

Precise tracking of traffic is not possible with conventional travel diary surveys using the existing self-survey method. To understand specific travel patterns, unstructured data, such as those found on social media, must be analyzed instead of structured data. Recently, many people worldwide have been using social media. Users upload various posts on their leisure travel history to social media. Social media posts include both structured data, such as posting date and location information, and unstructured data, such as content and images. Using social media data, travel patterns can be analyzed. For instance, Georgiou et al. (2015) showed that drivers' emotions may have an impact on the observed social discussion volume using signals such as humorous remarks, swearing, frustration, and occasional warnings in their social media content. Therefore, this study aims to track leisure travel using social media data and identify characteristics related to the use of eco-friendly transportation by analyzing the relationship between type of leisure (purpose of travel) and traffic volume change (means of travel).

2. Methodology

First, data from Instagram posts is collected using web crawling techniques. Text data of posts is applied with text mining, and the location information of the post is mapped to coordinates. Pre-processed text data is assigned topics for each post by applying LDA. After that, the space distribution for the topics of the posts and the usage of green transportation by district are analyzed. Figure 1 illustrates the flow of this study.

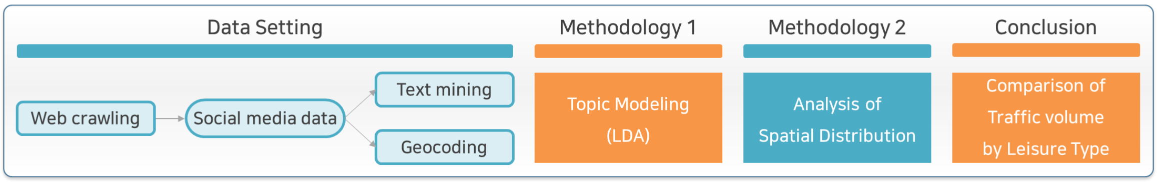

Figure 1: Framework of study

2.1 Social Media data

Instagram, a mobile photo capturing and sharing service, has emerged as a popular new data-sharing medium in recent years (Hu et al. 2014). Instagram offers users a post including up to 2,200 text characters, the ability to include multiple images or videos in a single post and the option to tag content with searchable hashtags (Carpenter et al., 2020). Individual posts typically pertain to location, date, time, content, comments, and likes. To efficiently collect social media data, information on posts was collected using a web crawling framework, Selenium. Using Instagram's hashtag function, 5,000 posts were crawled with the keyword 'Fortress Wall of Seoul', which is green transportation promotion area in Seoul, South Korea. The location name variable was converted into coordinate form by calling Kakao API key to extract latitude and longitude. Only posts uploaded in Fortress Wall of Seoul were extracted using the clip function, one of the GIS tools, and as a result, 1,973 data points were collected from 2020 to 2021. Considering the importance of data preprocessing for effective text data analysis, the text data were tokenized on a word basis, and the document was divided into word units. The words extracted in this manner were normalized by rule-based integration, case-insensitive integration, elimination of stopwords, and stemming. Subsequently, the Korean preprocessing package was used to ensure correct spelling. The structure of the social media data is shown in Table 1.

Table 1: Components of social media data

Components	Description	Property
Upload id	ID of user who uploaded the post	TEXT
Location name	The location where the post was created or artificially specified by the user	TEXT
Date	The date when user uploaded the post (year-month-day)	DATE
Time	The time when user uploaded the post (hour-minute-seconds)	TIME
Content	The content the user has written in the body	TEXT
Comments	The content of the comments in the post	TEXT
Likes	The number of people who clicked "Like" on the post	NUMBER

Kusmawan et al. (2014) used SNS messages to monitor the traffic conditions on a road by computing the degree of traffic congestion. Alkouz and al Aghbari (2020) proposed SNSJam, a system for detecting and predicting road traffic jams using Instagram data sources. In other words, it resulted in an effective method of collecting traffic information based on unstructured data and analyzing travel behavior through SNS.

2.2 Latent Dirichlet Allocation (LDA)

LDA is a representative algorithm for topic modeling, which is the process of identifying topics in a set of documents. In this study, topic modeling was performed under the assumption that the topic distribution of LDA is not fixed, and is randomly generated from the Dirichlet distribution. After extracting the topic distribution for a specific document, the topic value was randomly extracted based on the length of the document, and words were randomly generated from the word distribution corresponding to the topic. LDA performs topic classification of documents in the following manner: LDA receives the number of topics K under the assumption that they are distributed over all documents M of topics K. Therefore, the user must assign the number of topics K, which is a hyperparameter. LDA randomly allocates one of the topics K for every word in every document. Consequently, each document has a topic, which, in turn, has a word distribution; this result will be inaccurate because it is randomly assigned. When LDA allocates topics for each word, it assumes that each word w in a document is assigned to the wrong topic, but all other words are assigned to the correct topic. Under this assumption, LDA reassigns the topic of the word on the basis of two criteria: the ratio of the words corresponding to topic T among the words in document D and the ratio of all documents with word w. By repeating this process, the assignment of topics to all words is gradually unified into a convergent state. The graphical model representation of LDA is shown in Figure 2. The probability that topic $z_{d,i}$ of i-th word of the d-th document is assigned to the j-th is given by Eq(1):

$$P\left(z_{d,i} = j \middle| z_{-i}, w\right) = \frac{n_{d,k}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)} \times \frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{i=1}^{V}(v_{k,j}+\beta_j)} = AB \tag{1}$$

where, $n_{d,k}$ is frequency of words in the d-th document assigned to the k-th topic, $v_{k,w_{d,n}}$ is frequency of $w_{d,n}$ assigned to the k-th topic in the entire corpus, $w_{d,n}$ is the n-th word in the d-th document, V is total number of words in the corpus, α is frequency of words in the d-th document assigned to the k-th topic, β is dirichlet distribution parameter for creating word distributions in topics, A is the degree of association that the d-th document has with the k-th topic, B is the degree of association that $w_{d,n}$ has with the k-th topic.

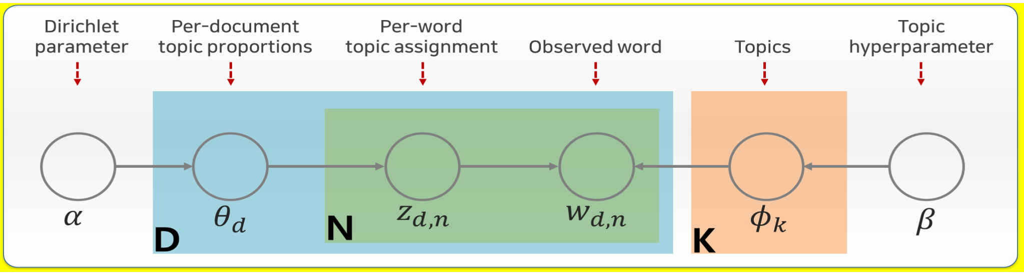

Figure 2: Graphical model representation of LDA

Zhao et al. (2020) adapted LDA to discover representative and interpretable activity categorization and latent activity patterns from transit smart card data in an unsupervised manner. In this study, individual posts on Instagram were considered as a single document and topics related to leisure activities were extracted. Subsequently, each document was allocated a distribution and weight of topics, which were entered as features of the document.

2.3 Influencing Factor in Travel Behavior

Various studies have been conducted to determine the factors that influence travel behavior. Kamargianni et al. (2015) studied transport network characteristics, such as the availability of a separate bicycle path, bicycle parking spaces, and sidewalk width, which significantly affected the choice of active transport in teenagers. The results indicated that a green lifestyle increases the probability that the individual chooses to travel by bus, while an active lifestyle increases the probability that the individual chooses active transport (walking and cycling). Saelens et al. (2003) suggested that residents from communities with higher density, greater connectivity, and more land use reported higher rates of walking/cycling for utilitarian purposes than low-density, poorly connected, and single land-use neighborhoods. In other words, the choice of transportation modes depends on the region and affects each person in different ways, depending on individual characteristics. However, the analysis of travel patterns based on existing structured data has limitations in estimating the purpose of travel and specific travel behavior. Nondaily travel data can be obtained from social media data, and activity types can be identified using unstructured data analysis. Therefore, this study observed its characteristics through geographical visualization and unstructured data analysis by matching the change in traffic volume by transportation modes and the number of leisure activities that appear locally.

3. Results

This section presents the results of the topic modeling and spatial analysis performed to establish a region-wise correlation between the type of leisure activity and the use of green transportation modes.

3.1 Latent Dirichlet Allocation (LDA)

In this study, LDA was used in the GenSim library developed in Python. A total of 1,973 pre-processed Instagram posts were used to extract topics using Gibbs sampling-based LDA algorithms, and the posts were reclassified according to the extracted topics. Four topics were set and the top 30 keywords were derived for each topic. The primary keywords and number of documents are listed in Table 2. For Topic 1, the main keywords were people, friends, love, cafés, and restaurants. For Topic 2, the main keywords were running, hiking, mountain, trail, and walking. For Topic 3, the main keywords were tour, travel, culture, guide, and museum. Finally, for Topic 4, the main keywords were flower, walk, outings, relaxation, and art. Therefore, topics 1 to 4 were labelled social, exercise, tourism, and rest.

Table 2: Result of LDA

Topic	Subject	Primary keywords	Num of documents	Perc of documents
1	Social	people, friends, love, cafe, restaurant	517	0.26
2	Exercise	running, hiking, mountain, trail, walking	444	0.23
3	Tourism	tour, travel, culture, guide, museum	433	0.22
4	Rest	flower, walks, outings, relaxation, art	579	0.29

The topic modeling results were visualized using the LDAvis library developed in Python. LDAvis shows the topic similarity and word distribution by topic by reducing the topic with a number of words into two dimensions using principal component analysis (PCA). The size of the circle indicates the number of words relevant to the topic and their distribution. The distance between the circles indicates the similarity between topics; that is, the closer it is, the more similar the topic. The bar shows the leading keywords that form the topic and the topics related to that keyword. Keyword extraction was based on two criteria, salience and discriminative power. Salience indicates that a word is important because it frequently appears in each document. In contrast, discriminative power indicates that a word appears repeatedly throughout a document and is consequently less important. In other words, a negative correlation exists between salience and discriminative power. Figure 3 shows the results of LDAvis.

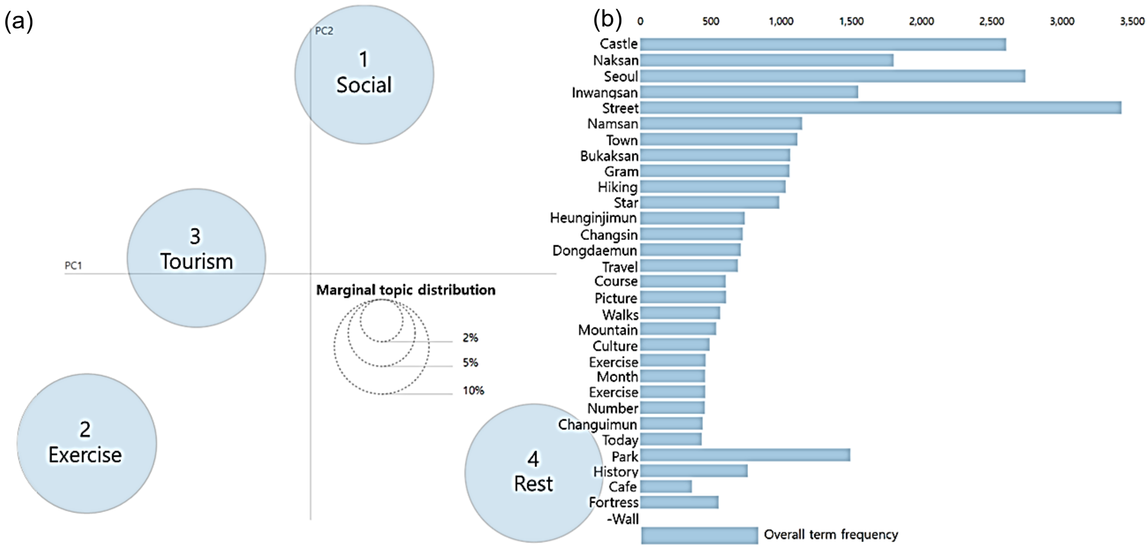

Figure 3: (a) Intertopic distance map (via multidimensional scaling) (b) Top-30 most salient terms

3.2 Analysis of Spatial Distribution

In this study, the 'ArcGIS' software was used for spatial analysis. Each post was mapped to a low-emission zone using longitude and latitude coordinates. The distribution of representative topics by region in the low-emission zone is shown in Figure 4. For instance, exercise-type leisure activities were found to predominate in the northern region. This observation is valid because this region comprises popular mountainous areas such as the Inwangsan and Bukaksan Mountains, where several people come to hike. In the central region, tourism was identified as the main leisure activity. This is confirmed by the fact that this region includes several cultural monuments and tourist attractions, such as the Changdeokgung Palace and Jongmyo Shrine. In the southern region, rest-type leisure activities were found to predominate. This is confirmed by the presence of several resting places, such as Namsan Cherry Blossom Road and Namsan Park, in this region. Finally, social-type activities were identified as the primary leisure activities in the eastern region. This is confirmed by the presence of trendy places such as Eulji-ro and Daehak-ro, which are frequented by youngsters.

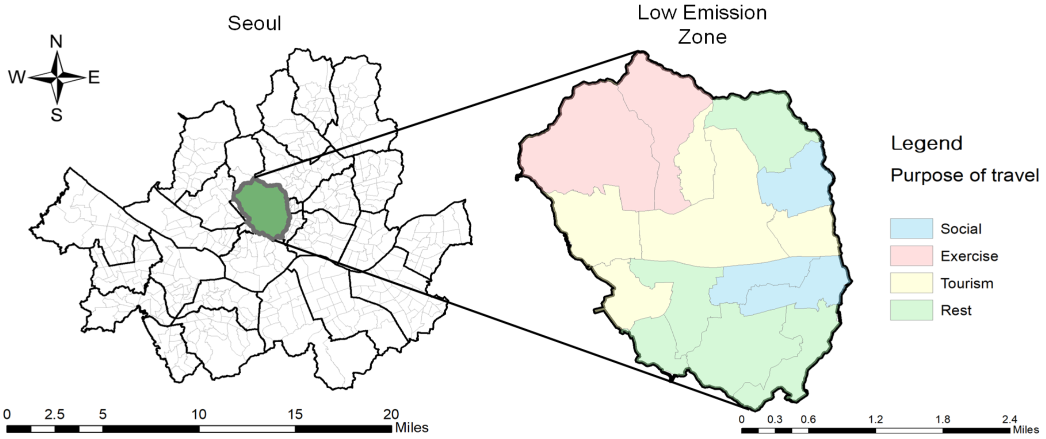

Figure 4: Distribution of representative topics by region

By dividing regions according to representative leisure types, the regional changes in traffic volume were confirmed. The data represents the period from 2020 to 2021. The traffic volume includes the number of shared bicycle rentals and public transportation rides, which is the sum of the number of bus and subway rides. The analysis showed that areas where social-type and tourism-type activities predominate, exhibit a higher number of bicycle rentals and public transportation rides than the average of all leisure activities. In particular, the number of shared bicycle rentals was the highest at 23,785 rentals/station in areas where social-type was the main leisure activity, and the number of public transportation riders was the highest at 2,345,624 people/station in areas where tourism-type was the main leisure activity. In other words, people travelling to the low emission zone for social and tourism purposes availed green transportation modes more frequently. Conversely, in areas where exercise-type was the main leisure activity, the number of bicycle rental and public transportation riders was 12,446 rental/station and 262,330 people/station, which were significantly lower than the average. In other words, most people who engage in intense exercise activities avail the services of a passenger car, which is a door-to-door transportation mode that requires less effort.

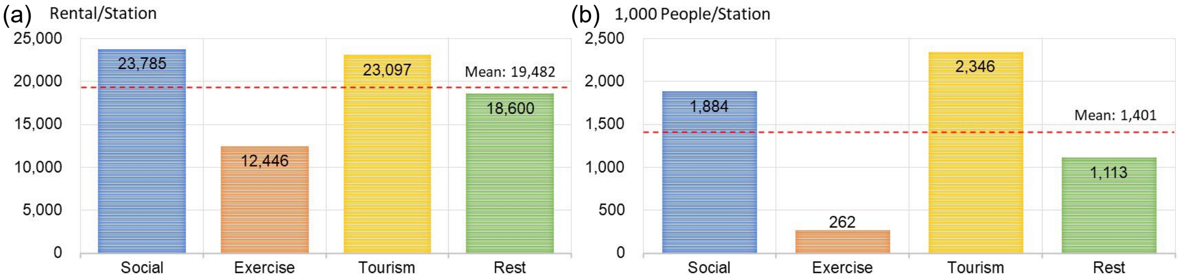

Figure 5: (a) Number of shared bicycle rental by region (b) Number of public transportation rides by region

4. Conclusions

To revitalize the government's green transportation policy, the individual behavior of travellers according to the purpose of travel must be analyzed. In this study, individual non-daily data was collected using Instagram. LDA was used to extract topics classified into four leisure types: social, exercise, tourism, and rest. The spatial distribution of these topics within the low-emission zone was determined. The representative topics were different for each region: northern, central, southern, and eastern regions. In addition, by analyzing the amount of use of transportation according to leisure activities by region, the correlation between the purpose of travel and the mode of transportation was identified. People who travelled for tourism-type leisure activities used public transportation the most. This seems to be a typical use of public transportation facilities for travelers without a car to visit tourist attractions. In contrast, people involved in exercise-type leisure tend not to use these modes of transportation. People who put a lot of energy into intense exercise activities mainly used a passenger car, because they did not want to use a lot of energy during travel. From this perspective, establishing a regional customized linkage plan for green transportation according to the major types of activities can maximize the effectiveness of green transportation activation strategies. For instance, policies that improve the infrastructure of social and tourism-type leisure activities can effectively revitalize green transportation in low-emission zones. In addition, if a differentiated activation strategy is prepared that reflects the main leisure types and characteristics of each region, the use of eco-friendly transportation can be promoted. Activating customized transportation according to the purpose of travel will promote eco-friendly cities.

Main contribution of this study is to identify individual traffic purposes by collecting and analyzing unstructured data from social media rather than the conventional self-survey method, and find a relationship with eco-friendly transportation. People recorded their leisure information about low-emission zone on social media. Using unstructured data stored on social media, the main purpose of travel different for each region was identified. The purpose of travel affected visitors' choice of means of travel, and significant differences in the usage of green transportation were confirmed for each purpose of travel. The target area of this study was low-emission zone, and if keywords are used differently when collecting posts, this framework can be applied to various areas. There are two limitations that the keyword must be set by users themselves, and it is difficult to collect various posts without keywords uploaded in the region. In the future, based on the results of this study, it can be a great help in predicting traffic demand by precisely tracking individual traffic behaviors.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019K1A4A7A03112460). This work was also supported by the Korea Ministry of Land, Infrastructure, and Transport (MOLIT) as an Innovative Talent Education Program for Smart City.

References

Alkouz B., Al Aghbari Z., 2020, SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks, Information Processing & Management, 57(1), 102139.

Bencekri M., Ku D., Kwak J., Kim J., Lee S., 2021, Review of eco-friendly guidance of transport infrastructure: Korea and the World, Chemical Engineering Transactions, 89, 235–240.

Carpenter J.P., Morrison S.A., Craft M., Lee M., 2020, How and why are educators using Instagram?, Teaching and Teacher Education, 96, 103149.

Cho M., Ku D., Lee S., Lee S., 2021, Environmental Impact of Personal Mobility in Road Managements, Chemical Engineering Transactions, 89, 331–336.

Georgiou T., Abbadi A. el, Yan X., George J., 2015, Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application, Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 25th-28th August, Paris, France, 330–335.

Hu Y., Manikonda L., Kambhampati S., 2014, What We Instagram: A First Analysis of Instagram Photo Content and User Types, Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 1st-4th June, Ann Arbor, USA, 595-598.

Kamargianni M., Dubey S., Polydoropoulou A., Bhat, C., 2015, Investigating the subjective and objective factors influencing teenagers' school travel mode choice – An integrated choice and latent variable model, Transportation Research Part A: Policy and Practice, 78, 473–488.

Ku D., Kwak J., Na S., Lee S., Lee S., 2021, Impact assessment on cycle super highway schemes, Chemical Engineering Transactions, 83, 181-186.

Kusmawan P.Y., Hong B., Jeon S., Lee J., Kwon J., 2014, Computing Traffic Congestion Degree Using SNS-based Graph Structure, 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), 10th-13th November, Doha, Qatar, 397–404.

Kwak J., Oh H., Jeong I., Shin S., Ku D., Lee S., 2021, Changes in shared bicycle usage by COVID-19, Chemical Engineering Transactions, 89, 169–174.

Lee J.S., Christopher Zegras P., Ben-Joseph E., Park S., 2014, Does urban living influence baby boomers' travel behavior?, Journal of Transport Geography, 35, 21–29.

Saelens B.E., Sallis J.F., Frank L.D., 2003, Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures, Annals of Behavioral Medicine, 25(2), 80–91.

Srinivasan S., Rogers P., 2005, Travel behavior of low-income residents: studying two contrasting locations in the city of Chennai, India, Journal of Transport Geography, 13(3), 265–274.

Zhao Z., Koutsopoulos H.N., Zhao J., 2020, Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model, Transportation Research Part C: Emerging Technologies, 116, 102627.