Knowledge and Information Systems
https://doi.org/10.1007/s10115-024-02112-8

BRIEF REPORT


# Big data in transportation: a systematic literature analysis and topic classification

**Danai Tzika-Kostopoulou[1] · Eftihia Nathanail[1] · Konstantinos Kokkinos[2]**

Received: 6 August 2023 / Revised: 26 January 2024 / Accepted: 21 March 2024


**Abstract**
This paper identifies trends in the application of big data in the transport sector and categorizes research work across scientific subfields. The systematic analysis considered literature published between 2012 and 2022. A total of 2671 studies were evaluated from a dataset of 3532 collected papers, and bibliometric techniques were applied to capture the evolution of research interest over the years and identify the most influential studies. The proposed unsupervised classification model defined categories and classified the relevant articles based on their particular scientific interest using representative keywords from the title, abstract, and keywords (referred to as top words). The model's performance was verified with an accuracy of 91% using Naïve Bayesian and Convolutional Neural Networks approach. The analysis identified eight research topics, with urban transport planning and smart city applications being the dominant categories. This paper contributes to the literature by proposing a methodology for literature analysis, identifying emerging scientific areas, and highlighting potential directions for future research.

**Keywords** Big data · Transportation · Topic model · Classification · Term frequency–inverse document frequency method

## 1 Introduction

Urbanization, ongoing changes in mobility patterns, and rapid growth in freight transportation pose significant challenges to stakeholders and researchers within the transportation sector. New policies have focused on promoting sustainability and reducing emissions through smart

✉ Danai Tzika-Kostopoulou
danaitzika@uth.gr

Eftihia Nathanail
enath@uth.gr

Konstantinos Kokkinos
kokkinos@uth.gr

[1] Department of Civil Engineering, University of Thessaly, Volos, Greece

[2] Department of Digital Systems, University of Thessaly, Larissa, Greece

Published online: 08 May 2024

and multimodal mobility [1]. To address these challenges, authorities are developing strategies that employ technological advances to gain a deeper understanding of travel behavior, produce more accurate travel demand estimates, and enhance transport system performance.

Undoubtedly, the development of Intelligent Transport Systems (ITS) and recent advances in Information and Communication Technology (ICT) have enabled the continuous generation, collection, and processing of data and the observation of mobility behavior with unprecedented precision [2]. Such data can be obtained from various sources, including ITS, cell phone call records, smart cards, geocoded social media, GPS, sensors, and video detectors.

Over the past decade, there has been increasing research interest in the application of big data in various transportation sectors, such as supply chain and logistics [3], traffic management [4], travel demand estimation [5], travel behavior [6], and real-time traffic operations [7]. Additionally, numerous studies in the field of transport planning and modeling have applied big data to extract vital attributes including trip identification [8] and activity inference [9]. Despite research efforts and existing applications of big data in transportation, many aspects remain unknown, and the prospects of big data to gain better insights into transport infrastructure and travel patterns have not yet been explored.

To maximize the benefits of big data in traffic operations, infrastructure maintenance, and predictive modeling, several challenges remain to be addressed. These include handling the high velocity and volume of streaming data for real-time applications, integrating multiple data sets with traditional information sources, ensuring that the data is representative of the entire population, and accessing big data without compromising privacy and ethical concerns. In terms of transport modeling and management, the research focuses on achieving short-term stability by incorporating more comprehensive data sets that cover hourly, daily, seasonal, or event-based variations, and on enhancing mobility on demand through real-time data processing and analysis. Additionally, it has become necessary to further investigate the methodology for processing large amounts of spatial and temporal information that has primarily been intended for non-transport purposes and to reconsider the existing analytical approaches to adapt to the changing data landscape.

There is an intention among policymakers, transport stakeholders, and researchers to better understand the relationship between big data and transport. The first step in this direction is to identify the key areas of big data utilization in the transport sector. Therefore, this study attempts to map big data applications within the transport domain. It provides a broad overview of big data applications in transport and contributes to the literature by introducing a methodology for identifying emerging areas where big data can be successfully applied and subfields that can be further developed.

The scope of the current study is twofold. First, a holistic literature analysis based on bibliometric techniques complemented by a topic classification model covering the complete domain of big data applications in the transportation sector was implemented. Despite numerous studies attempting to review the relevant literature, such as big data in public transport planning or transport management [10, 11] to the best of our knowledge, no such investigation has produced a comprehensive and systematic cluster of multiple big data applications across the entire transportation domain based on a significant number of literature records. Therefore, the primary objective of this study is to classify the literature according to its particular interest and to pinpoint evolving scientific subfields and current research trends.

Second, as multiple studies have been conducted in this domain, the need to identify and assess them prior to running one's own research through a thorough literature review is always necessary. However, the analysis and selection of appropriate studies can be challenging, particularly when the database is large. Therefore, this study aims to provide a comprehensive

approach for evaluating and selecting appropriate literature that could be a methodological tool in any research field. Bibliometric methods have been widely applied in several scientific fields. Most of these studies use simple statistical methods to determine the evolution of the number of publications over the years, authors' influence, or geographical distribution. There are also research works that attempt to categorize the literature mainly by manual content analysis [12, 13] or by co-citation analysis, applying software for network and graphical analyses [14, 15]. In this study, the review process also included an unsupervised topic model to classify literature into categories.

This paper presents a comprehensive evaluation of up-to-date published studies on big data and their applications in the transportation domain. A total of 2671 articles from Elsevier's Scopus database, published between 2012 and 2022 were analyzed. Bibliometric techniques were applied to capture the evolution of research over time and uncover emerging areas of interest. In addition, the focus of this study is to define categories and classify relevant papers based on their scientific interests. To achieve this, unsupervised classification was applied using the topic model proposed by Okafor [16] to identify clusters, extract the most representative topics, and group the documents accordingly.

The current study attempts to answer the following questions:

(1) Which studies contribute the most to the field of big data in transportation?
(2) What is the evolution of research over time in this field of interest?
(3) What are the main research areas that have potential for further exploration?
(4) What are the directions of future research?

This paper consists of six sections. Following the introduction, Sect. 2 provides a summary of previous research in this subject area. Section 3 outlines the methodology applied in this research. This includes the process of defining the eligible studies, bibliometric techniques utilized, and the topic model employed for paper classification. Section 4 presents the initial statistical results and the classification outcomes derived from the topic model. In Sect. 5, the findings are summarized, and the results associated with the research questions are discussed. The final Section presents the general conclusions and research perspectives of the study.

## 2 Literature review

Due to the significant benefits of big data, several studies have been conducted in recent years to review and examine the existing applications of different big data sources in transportation. Most of these focus on a specific transport domain, such as transport planning, transport management and operations, logistics and supply chain and Intelligent Transportation Systems.

In the context of transport planning and modeling, Anda et al. [2] reviewed the current application of historical big data sources, derived from call data records, smart card data, and geocoded social media records, to understand travel behavior and to examine the methodologies applied to travel demand models. Iliashenko et al. [17] explored the potential of big data and Internet of Things technologies for transport planning and modeling needs, pointing out possible applications. Wang et al. [18] analyzed existing studies on travel behavior utilizing mobile phone data. They also identified the main opportunities in terms of data collection, travel pattern identification, modeling and simulation. Huang et al. [19] conducted a more specialized literature review focusing on the existing mode detection methods based on mobile phone network data. In the public transportation sector, Pelletier et al. [20] focused on the application of smart card data, showing that in addition to fare collection, these data can

also be used for strategic purposes (long-term planning), tactical purposes (service adjustment and network development), and operational purposes (public transport performance indicators and payment management). Zannat et al. [10] provided an overview of big data applications focusing on public transport planning and categorized the reviewed literature into three categories: travel pattern analysis, public transport modeling, and public transport performance assessment.

In traffic forecasting, Lana et al. [21] conducted a survey to evaluate the challenges and technical advancements of traffic prediction models using big traffic data, whereas Miglani et al. [22] investigated different deep learning models for traffic flow prediction in autonomous vehicles. Regarding transport management, Pender et al. [23] examined social media use during transport network disruption events. Choi et al. [11] reviewed operational management studies associated with big data and identified key areas including forecasting, inventory management, revenue management, transportation management, supply chain management, and risk analysis.

There is also a range of surveys investigating the use of big data in other transport subfields. Ghofrani et al. [24] analyzed big data applications in railway engineering and transportation with a focus on three areas: operations, maintenance, and safety. In addition, Borgi et al. [25] reviewed big data in transport and logistics and highlighted the possibilities of enhancing operational efficiency, customer experience, and business models.

However, there is a lack of studies that have explored big data applications attempting to cover a wider range of transportation aspects. In this regard, Zhu et al. [26] examined the features of big data in intelligent transportation systems, the methods applied, and their applications in six subfields, namely road traffic accident analysis, road traffic flow prediction, public transportation service planning, personal travel route planning, rail transportation management and control, and asset management. Neilson et al. [27] conducted a review of big data usage obtained from traffic monitoring systems crowdsourcing, connected vehicles, and social media within the transportation domain and examined the storage, processing, and analytical techniques.

The study by Katrakazas et al. [28], conducted under the NOESIS project funded by the European Union's (EU) Horizon 2020 (H2020) program, is the only one we located that comprehensively covers the transportation field. Based on the reviewed literature, the study identified ten areas of focus that could further benefit from big data methods. The findings were validated by discussing with experts on big data in transportation. However, the disadvantage of this study lies in its dependence on a limited scope of the reviewed literature.

The majority of current review-based studies concentrate on one aspect of transportation, often analyzing a single big data source. Many of these studies rely on a limited literature dataset, and only a few have demonstrated a methodology for selecting the reviewed literature. Our review differs from existing surveys in the following ways: first, a methodology for defining the selected literature was developed, and the analysis was based on a large literature dataset. Second, this study is the only one to employ an unsupervised topic classification model to extract areas of interest and open challenges in the domain. Finally, it attempts to give an overview of the applications of big data across the entire field of transportation.

## 3 Research methodology

This study followed a three-stage literature analysis approach. The first stage includes defining the literature source and the papers' search procedures, as well as the "screening" to select the reviewed literature. The second stage involves statistics, which are widely employed in bibliometric analysis, to capture trends and primary insights. In the third stage, a topic classification model is applied to identify developing subfields and their applications. Eventually, the results are presented, and the findings are summarized.

### 3.1 Literature selection

The first step in this study was to define the reviewed literature. A bibliographic search was conducted using the Elsevier's Scopus database. Scopus and Web of Science (WOS) are the most extensive databases covering multiple scientific fields. However, Scopus offers wider overall coverage than WoS CC and provides a better representation of particular subject fields such as Computer Sciences [29] which is of interest in this study. Additionally, Scopus comprises 26591 peer-reviewed journals [30] including publications by Elsevier, Emerald, Informs, Taylor and Francis, Springer, and Interscience [15], covering the most representative journals in the transportation sector.

The relevant literature was identified and collected using the Scopus search API, which supports Boolean syntax. Four combinations of keywords were used in the "title, abstract, keywords" document search of the Scopus database including: "Big data" and "Transportation", "Big data" and "Travel", "Big data" and "Transport", "Big data" and "Traffic". The search was conducted in English as it offers a wider range of bibliographic sources. Only the last decade's peer-reviewed research papers published in scientific journals and conference proceedings have been collected, written exclusively in English. Review papers and document types such as books and book chapters were excluded. As big data in transport is an interdisciplinary field addressed by different research areas, in order to cover the whole field of interest, the following subject areas were predefined in the Scopus search: computer sciences; engineering; social sciences; environmental sciences; energy; business, management and accounting. Fields considered irrelevant to the current research interest were filtered out.

The initial search resulted in a total of 5234 articles published between the period 2012–2022. The data was collected in December 2021 and last updated on the 5th of September 2023. The results were stored in a csv format, including all essential paper information such as paper title, authors' names, source title, citations, abstracts, year of publication, and keywords. After removing duplications, a final dataset of 3532 papers remained.

The paper dataset went through a subject relevance review, at the first stage, by checking in the papers' title or keywords the presence of at least one combination of the search terms. If this condition was not met, a further review of the paper abstracts was conducted. From both stages, a filtered set of papers was selected, based on their relevance to the current study's areas of interest, associated with the search items. A total of 2671 selected papers formed the dataset which was further analyzed, evaluated, and categorized, based on clustering techniques.

### 3.2 Initial statistics

Once the dataset was defined, statistical analysis was performed to identify influential journals and articles within the study field. The first task was to understand the role of the different journals and conference proceedings. Those with the most publications were listed and further

analyzed according to their publication rate and research area of interest. Second, the number of citations generated by the articles was analyzed as a measure of the quality of the published studies, and the content of the most cited articles was further discussed. The above provided essential insights into research trends and emerging topics.

### 3.3 Topic classification

A crucial step in our analysis was to extract the most representative sub-topics and classify the articles into categories by applying an unsupervised topic model [16]. Initially, the Excel file with the selected papers' data (authors, year, title, abstract, and keywords) was imported into the model. Abstracts, titles, and keywords were analyzed and text-cleaning techniques were applied. This step includes normalizing text, removing punctuations, stopwords, and words of length less than three letters, as well as the lemmatization of the words. The most popular software tools/libraries used for text mining and cleaning, as well as natural language processing (NLP) in the topic model process, are implemented in Python programming and include NLTK (https://www.nltk.org/), spaCy (https://spacy.io/), Gensim (https://radimrehurek.com/gensim/), scikit-learn (https://scikit-learn.org/stable/), and Beautiful Soup (https://www.crummy.com/software/BeautifulSoup/). NLTK is a powerful NLP library with various text preprocessing functions, while spaCy handles tokenization, stop word removal, stemming, lemmatization, and part-of-speech tagging. Gensim is a popular library for topic labeling and document similarity analysis. To process textual data, scikit-learn is a machine learning library with text preprocessing functions. Finally, Beautiful Soup is a web-based library for parsing HTML and XML documents. For the approach explained in the following sections, NLTK and Beautiful Soup were used to parse web metadata for the research papers. Moreover, bigrams and trigrams of two or three words, frequently occurring together in a document, were created.

The basic aim was to generate clusters (topics) using a topic model. The proposed model extracts representative words from the title, abstract, and keyword section of each paper, aiming to cluster research articles into neighborhoods of topics of interest without requiring any prior annotations or labeling of the documents. The method initially constructs a word graph model by counting the Term Frequency – Inverse Document Frequency (TF-IDF) [31].

The resulting topics are visualized through a diagram that shows the topics as circular areas. For the aforementioned mechanism, the Jensen-Shannon Divergence & Principal Components dimension reduction methodology was used [32]. The model is implemented in the LDAvis (Latent Dirichlet Allocation) Python library, resulting in two principal components (PC1 and PC2) that visualize the distance between the topics on a two-dimensional plane. The topic circles are created using a computationally greedy approach (the first in-order topic gets the largest area, and the rest of the topics get a proportional area according to their calculated significance). The method picks the circle centroids randomly for the first one (around the intersection of the two axes), and then the distance of the circle centroids is relevant to the overlapping of the top words according to the Jensen-Shannon Divergence model. The former model was applied for numerous numbers of topics, while its performance was experimentally verified using a combination of Naïve Bayesian Networks and a Convolutional Neural Network approach.

In the sequence of the two machine learning methodologies, supervised (Bayesian Neural Networks) and unsupervised (Deep Learning) research article classification were considered via sentiment analysis. For the supervised case, the relations among words were identified, offering interesting qualitative information from the outcome of the TF-IDF approach. The

results show that the Bayesian Networks perform in accuracies near 90% of the corresponding statistical approach [33] and the same happens (somewhat inferior performance) for the unsupervised case [34]. The two methods validated the results of the TF-IDF, reaching accuracies within the acceptable limits of the aforementioned proven performances.

### 3.3.1 TF-IDF methodology

The TF-IDF (Term Frequency–Inverse Document Frequency) method is a weighted statistical approach primarily used for mining and textual analysis in large document collections. The method focuses on the statistical importance of a word value emanating by its existence in a document and the frequency of occurrences. In this context, the statistical significance of words grows in proportion to their frequency in the text, but also in inverse proportion to their frequency in the entire corpus of documents. Therefore, if a word or phrase appears with high frequency in an article (TF value is high), but simultaneously seldom appears in other documents (IDF value is low), it is considered a highly candidate word that may represent the article and can be used for classification [35]. In the calculation of TF-IDF, the TF word value is given as:

$$\mathrm{tf}_{i,j} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

In *Eq.* 1, $n_{ij}$ denotes the frequency of the word occurrence of the term $t_i$ in the document $d_j$, and the denominator of the above fraction is the sum of the occurrence frequency of all terms in the document $d_j$. At the same time, the calculation of the IDF value of a term $t_i$, is found by dividing the total number of documents in the corpus by the number of documents containing the term $t_i$, and then obtains the quotient logarithm:

$$\mathrm{idf}_i = \log\left|\frac{|D|}{\{j : t_i \in d_j\}}\right| \tag{2}$$

In Eq. 2, $|D|$ denotes the total number of documents, and the denominator represents the number of documents $j$ which contain the term $t_i$ in that specific document $d_j$. In other words, we consider only the documents where $n_{ij} \neq 0$ in Eq. (1). In the case scenario in which the term $t_i$ does not appear in the corpus, it will cause the dividend to be zero in the above equation, causing the denominator to have the values of $+1$. Using Eqs. (1) and (2), TF-IDF is given by:

$$\mathrm{TF-IDF}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{3}$$

From Eq. 3, it can be assumed that high values of TF-IDF can be produced by high term frequencies in a document, while at the same time, low term frequencies occur in the entire document corpus. For this reason, TF-IDF succeeds in filtering out all high-frequency "common" words, while at the same time, retaining statistically significant words that can be the topic representatives [36]

### 3.3.2 Naïve bayes methodology

This methodology evaluates the performance of TF-IDF by following a purely "machine-learning" oriented approach, which is based on the Bayesian likelihood, i.e., reverse reasoning to discover arbitrary factor occurrences that impact a particular result. These arbitrary factors correspond to the corpus terms and their frequencies within each document and corpus.

The model is a multinomial naive Bayes classifier that utilizes the Scikit-learn, which is a free programming Artificial Intelligence (AI) library for the Python programming language to support: a. training text, b. feature vectors, c. the predictive model, and d. the grouping mechanism.

The results of the TF-IDF method were imported into the Bayesian classifier. More specifically, the entire dataset was first prepared to be inserted by applying noise, stop-words, and punctuation removal. The text was then tokenized into words and phrases. For topic modeling, TF-IDF was used for feature extraction, creating the corresponding vectors as features for classification. In the next step, the Naive Bayes classifier is trained on the pre-processed and feature-extracted data. During training, it learns the likelihood of observing specific words or features given each topic and the prior probabilities of each topic occurring in the training data. Not all data was used for training. The model split the data into a 70% portion used for the unsupervised training, with the remaining 30% to be used for validation. This split ratio is a rule of thumb and not a strict requirement. However, this popular split ratio is to strike a balance between having enough data for training the machine learning model effectively and having enough data for validation or testing to evaluate the model's performance. The split was used within the k-fold cross-validation to assess the performance of the model while mitigating the risk of overfitting. While the dataset is divided into k roughly equal-sized "folds" or subsets, a fixed train-test split is used within each fold. The results from each fold were then averaged to obtain an overall estimate of model performance. This approach has the advantage of assessing the model's performance in a more granular way within each fold, similar to how it is assessed in a traditional train-test split, but at the same time, it provides additional information about how the model generalizes to different subsets of the data within each fold.

In the process of text examination, a cycle of weighting is dynamically updated for cases of term recurrences in the examined documents. The documents still contain only the title, abstract, and keywords for each article. For these cases, the Bayes theorem is used:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{4}$$

where $P(c|x)$ is the posterior probability, $P(x|c)$ is the likelihood, $P(c)$ is the class prior probability, and $P(x)$ is the predictor prior probability with $P(c|x)$ resulting from the individual likelihoods of all documents $P(x_i|c)$, as depicted in Eq. (5):

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c) \tag{5}$$

This model was used as a validation method for the TF-IDF methodology, producing accuracy results reaching values of up to 91%. This value is widely acceptable for most cases of Bayesian classification and is expected to occur since prior classification has been applied [37].

### 3.3.3 Deep learning classification methodology

This is a secondary method of TF-IDF validation, which is based on the bibliometric coupling technique. However, this technique does not use the likelihood probability of the initial classification performed by TF-IDF but rather, this classifier deploys a character-based (i.e., the alphabetical letters composing the articles' text) convolutional deep neural network. Using the letters as basic structural units, a Convolutional Neural Network [38] learns words and discovers features and word occurrences in various documents. The model has been

primarily applied for computer vision and image machine learning techniques [39], but it is easily adapted for textual analysis.

All features previously used features (title, abstract, keywords) were kept and concatenated into a single string of text, which was itself truncated to a maximum length of 4000 characters. The size of the string can be dynamically adapted for each TensorFlow-model [40] according to the GPU performance of the graphics adapter of the hardware used, and basically represents the maximum allowable length for each feature in the analysis. The encoding involved all 26 English characters, 10 digits, all punctuation signs, and the space character. Other extraneous characters were eliminated. Furthermore, keyword information was considered primary in topic classification and encoded into a vector of the proportion of each subfield mentioned in the reference list of all documents/articles used in data. The system was rectified to behave as a linear unit, producing the activation function of the deep neural network between each layer. Only the last layer of the network utilized the SoftMax activation function for the final classification. The model was trained with a stochastic gradient descent as the optimizer and categorical cross-entropy as the loss function producing inferior results when compared with the corresponding Bayesian case as expected.

## 4 Results

### 4.1 Source analysis

To understand the role of diverse academic sources, the leading eleven journals or conference proceedings were identified (Table 1), which have published a minimum of twenty papers between 2012 and 2022 in the field of interest. According to the preliminary data, 995 journals and conference proceedings have contributed to the publication of 2671 papers. Eleven sources have published 553 articles, representing the 2100% of all published papers.

Three sources in the field of computer science have published 239 articles. Five transportation journals and conference proceedings have published 208 papers, while there are

**Table 1** Top eleven journals with twenty or more publications on big data in the transportation sector

| Journals/conference proceedings | No of publications |
|---|---|
| Proceedings- IEEE international conference on big data (2013–2022) | 133 |
| ACM international conference proceeding series | 75 |
| Sustainability | 66 |
| Transportation research part C: emerging technologies | 59 |
| IEEE transactions on intelligent transportation systems | 58 |
| Proceedings of international conference on intelligent transportation, big data and smart city (2015–2022) | 47 |
| Procedia computer science | 31 |
| Journal of transport geography | 23 |
| Transportation research procedia | 21 |
| Cities | 20 |
| Computers, environment and urban systems | 20 |

journals on urban planning and policies (e.g., Cities) that have also significantly contributed to the research field (106 papers).

The research findings indicate the interdisciplinarity of the application of big data in transportation, encompassing not only computer science but also transport and urban planning journals, showing that transport specialists acknowledge the advantages of examining the applications of big data in the transport domain.

### 4.2 Citation analysis

Citation analysis classifies the papers by their citation frequency, aiming to point out their scientific research impact [14] and to identify influential articles within a research area. Table 2 demonstrates the top fifteen studies published between 2012 and 2022 (based on citation count on Scopus). Lv et al. [41] published the most influential paper in this period and received 2284 citations. This study applied a novel deep learning method to predict traffic flow. In this direction, four other articles focused on traffic flow prediction, real-time traffic operation, and transportation management [7, 11, 42, 43].

Other important contributions highlighted the significance of big data generated in cities and analyzed challenges and possible applications in many aspects of the city, such as urban planning, transportation, and the environment [44–47]. Xu et al. [48] focused on the Internet of Vehicles and the generated big data. Finally, among the most influential works, there are papers that investigated big data usage in various subfields of spatial analysis, as well as urban transport planning and modeling, focusing on travel demand prediction, travel behavior analysis, and activity pattern identification [5, 6, 49–51].

### 4.3 Topic model

As previously mentioned, a topic model based on the TF-IDF methodology was used to categorize the papers under different topics. The basic goal was to identify subscientific areas of transportation where big data were implemented. A critical task was to define the appropriate parameters of the model, particularly the number of topics that could provide the most accurate and meaningful results, which would be further processed. To achieve this, a qualitative analysis was conducted taking into account the accuracy results obtained from the validation methodology. Therefore, to obtain a more precise view of the project, the model was implemented under various scenarios, increasing each time the number of topics by one, starting from four topics to fifteen.

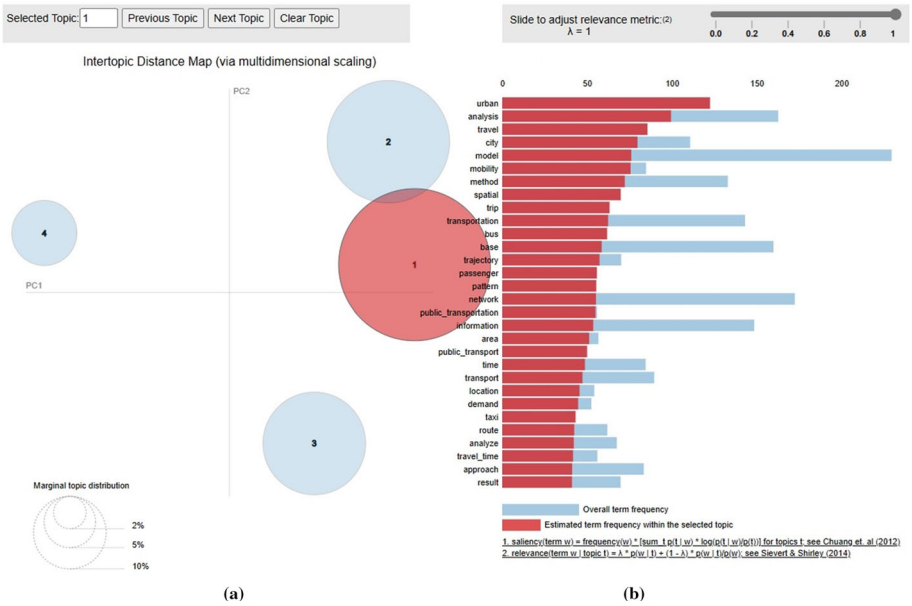
The mapping of the tokenized terms and phrases to each other was used to understand the relationships between words in the context of topics and documents. The technique applied is the Multidimensional Scaling (MDS), which helps to visualize and analyze these relationships in a lower-dimensional space [52, 53]. To highlight the relationships between terms, a matrix of tokens in the corpus was created based on their co-occurrence patterns. Each cell in the matrix represents how often two terms appear together in the same document. The MDS methodology provides the ability to view high-dimensional data in a lower-dimensional space while retaining as much of the pairwise differences or similarities between the terms as possible. More specifically, MDS translates the high-dimensional space of term co-occurrence relationships into a lower-dimensional space, where words that frequently co-occur are located closer to one another, and terms that infrequently co-occur are situated farther apart. When terms are displayed as points on a map or scatterplot via

**Table 2** Top fifteen cited articles

| References | Title | Source | Citations |
|---|---|---|---|
| [41] | Traffic flow prediction with big data: a deep learning approach | IEEE transactions on intelligent transportation systems | 2284 |
| [46] | Urban computing: concepts, methodologies, and applications | ACM transactions on intelligent systems and technology | 758 |
| [50] | Deep multi-view spatial–temporal network for taxi demand prediction | 32nd AAAI Conference on Artificial Intelligence (2018) | 668 |
| [45] | Big data, smart cities and city planning | Dialogues in Human Geography | 658 |
| [44] | Applications of big data to smart cities | Journal of Internet Services and Applications | 569 |
| [42] | A hybrid deep learning-based traffic flow prediction method and its understanding | Transportation Research Part C: Emerging Technologies | 466 |
| [47] | Internet-of-Things-Based Smart Cities: Recent Advances and Challenges | IEEE Communications Magazine | 433 |
| [48] | Internet of vehicles in big data era | IEEE/CAA Journal of Automatica Sinica | 419 |
| [11] | Big data analytics in operations management | Production and operations management | 400 |
| [43] | Predicting short-term traffic flow by long short-term memory recurrent neural network | 2015 IEEE International Conference on Smart City (Proceedings) | 357 |
| [6] | The promises of big data and small data for travel behavior (aka human mobility) analysis | Transportation Research Part C: Emerging Technologies | 352 |
| [49] | Detecting the dynamics of urban structure through spatial network analysis | International Journal of Geographical Information Science | 339 |
| [5] | The path most traveled: Travel demand estimation using big data resources | Transportation Research Part C: Emerging Technologies | 303 |
| [7] | Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways | Transportation Research Part C: Emerging Technologies | 297 |
| [51] | Urban activity pattern classification using topic models from online geo-location data | Transportation Research Part C: Emerging Technologies | 214 |

MDS, points that are close together in the scatterplot are used in the documents in a more connected manner.

As a consequence, the scatterplot that MDS produces can shed light on the structure of the vocabulary in relation to the topic model. It can assist in identifying groups of related terms that may indicate subjects or themes in the corpus. The pyLDAvis library of the Python programming language was used in conjunction with the sklearn library to incorporate the MDS into the clustering visualization process. This was done by superimposing the scatterplot




(a) (b)

**Fig. 1 a** Four topics distance map, **b** Top 30 most relevant terms for Topic 1

with details about the subjects assigned to documents or the probability distributions over different topics.

Initially, the papers were divided into four topics. Figure 1a displays four distinct clusters, each representing a scientific sub-area of interest. Clusters appear as circular areas, while the two principal components (PC1 and PC2) are used to visualize the distance between topics on a two-dimensional plane. Figure 1b illustrates the most relevant terms for Topic 1 in abstracts, titles, and keywords. Table 3 contains the most essential words associated with each sub-area based on the TF-IDF methodology, representing the nature of the four topics. The results of the model, considering, also, the content of the most influential papers in each group, reveal that the biggest cluster is associated with "transport planning and travel behavior analysis". In this topic, most papers have focused on long-term planning, utilizing big data mostly from smart cards, mobile phones, social media, and GPS trajectories to reveal travel and activity patterns or estimate crucial characteristics for transport planning, such as travel demand, number of trips, and trip purposes. The second topic refers to "smart cities and applications", containing papers about how heterogeneous data generated in cities (streamed from sensors, devices, vehicles, or humans) can be used to improve the quality of human life, city operation systems, and the urban environment. Most studies have concentrated on short-term thinking about how innovative services and big data applications can support function and management of cities. In terms of transportation, several studies have managed to integrate Internet of Things (IoT) and big data analytics with intelligent transportation systems, including public transportation service plan, route optimization, parking, rail transportation, and engineering and supply chain management. Two smaller clusters were observed. One cluster is dedicated to "traffic forecasting and management" and includes papers related to traffic flow prediction or accident prediction, while many of them focus on real-time traffic management, city traffic monitoring, and control, mainly using real-time traffic data generated by sensors, cameras,

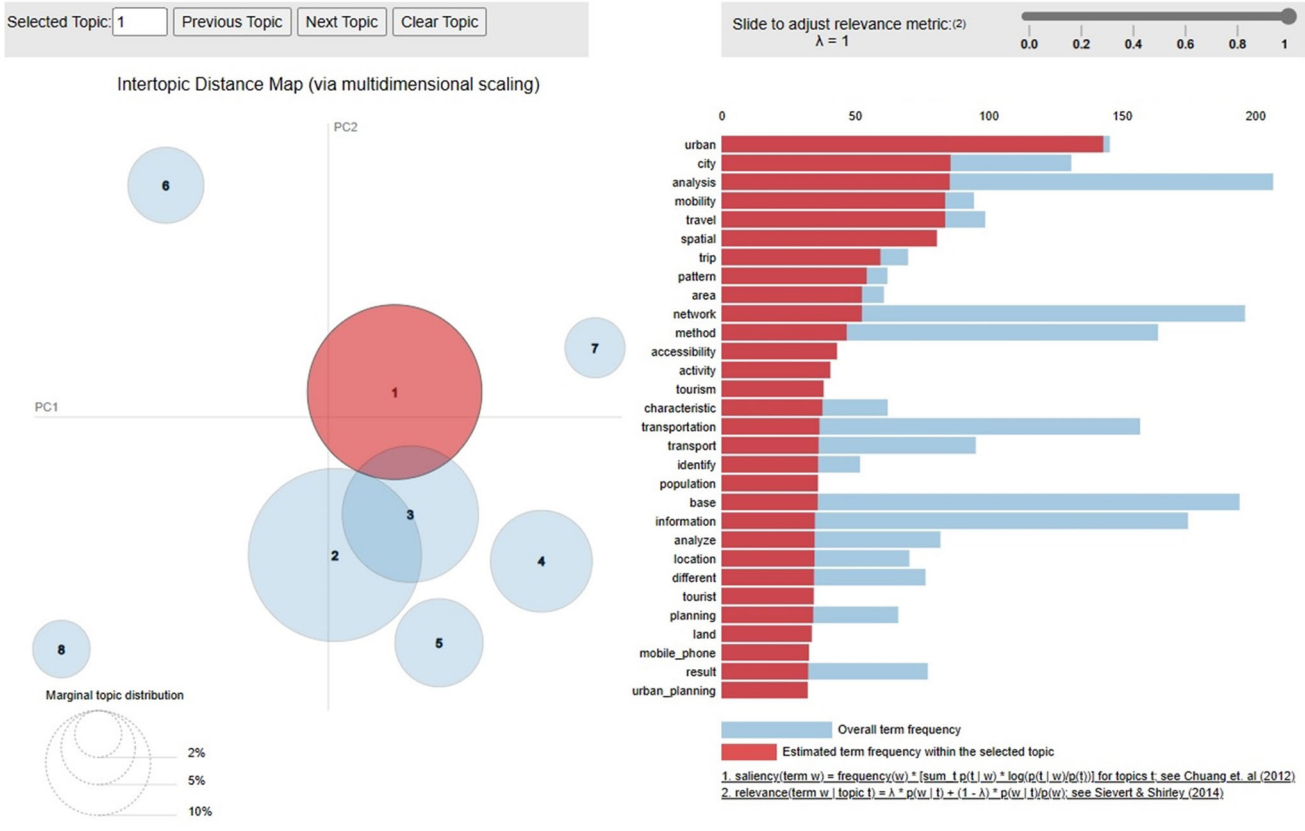
**Table 3** Clustering in four topics and top fifteen words

| Top Words | Topic 1<br>Transport planning and travel behavior analysis | Topic 2<br>Smart cities and applications | Topic 3<br>Traffic forecasting and management | Topic 4<br>Intelligent transportation systems and new technologies |
|---|---|---|---|---|
| 1 | Urban | Smart city | Traffic | Vehicle |
| 2 | Analysis | Smart | Model | Driver |
| 3 | Travel | Application | Traffic flow | Driving |
| 4 | City | Management | Prediction | Road |
| 5 | Model | Technology | Traffic congestion | Communication |
| 6 | Mobility | Transportation | Network | Internet of vehicles (IoV) |
| 7 | Method | Platform | Intelligent transportation system | Detection |
| 8 | Spatial | Services | Traffic flow prediction | Speed |
| 9 | Trip | Information | Deep learning | Vehicular |
| 10 | Transportation | Intelligent | Algorithm | Safety |
| 11 | Bus | Internet of things | Base | Intelligent transportation system |
| 12 | Base | Real time | Forecasting | Control |
| 13 | Trajectory | Logistics | Neural Network | Autonomous vehicle |
| 14 | Passenger | Architecture | Traffic management | Vehicular network |
| 15 | Pattern | Service | Road network | Electric vehicles |

and vehicles. The other is associated with "intelligent transportation systems and new technologies", focusing on topics such as the connected vehicle-infrastructure environment and connected vehicle technologies as a promising direction to enhance the overall performance of the transportation system. Considerable attention has also been given to Internet of Vehicles (IoV) technology, autonomous vehicles, self-driving vehicle technology, and automatic traffic, while many contributions in this topic are related to green transportation and suggest solutions for optimizing emissions in urban areas with a focus on vehicle electrification. The four topics are represented as circular areas. The two principal components (PC1 and PC2) are used to visualize the distance between topics on a two-dimensional plane.

By increasing the number of topics, various categories are becoming more specific, and it can be observed that there is a stronger co-relationship among clusters. The results of the eight topics' categorization are presented in Fig. 2a and b and Table 4. The eight topics are represented as circular areas. The two principal components (PC1 and PC2) are used to visualize the distance between topics on a two-dimensional plane.

According to the top words in each cluster and taking into consideration the content of top papers abstracts, the eight topics were specified as follows:




**(a)** **(b)**

**Fig. 2** **a** Eight topics distance map, **b** Top 30 most relevant terms for Topic 1

**Table 4** Clustering in eight topics and top fifteen words

| Top Words | Topic 1<br>Urban transport planning | Topic 2<br>Smart cities and applications | Topic 3<br>Traffic flow forecasting and modeling | Topic 4<br>Traffic management | Topic 5<br>Intelligent transportation systems | Topic 6<br>Public transportation | Topic 7<br>Railway | Topic 8<br>GPS trajectories |
|---|---|---|---|---|---|---|---|---|
| 1 | Urban | Smart city | Model | Traffic | Vehicle | Bus | Railway | Trajectory |
| 2 | City | Smart | Prediction | Traffic congestion | Driver | Public transportation | Maintenance | Trajectory data |
| 3 | Analysis | Application | Forecasting | Traffic flow | Driving | Passenger | Railroad | Road network |
| 4 | Mobility | Management | Traffic flow | Intelligent transportation system | Road | Public transport | Train | Information |
| 5 | Travel | Technology | Method | Congestion | Communication | Transit | Safety | Road |
| 6 | Spatial | Transportation | Deep Learning | Road network | Internet of vehicle (IoV) | Passenger flow | Rail | GPS trajectory |
| 7 | Trip | Information | Neural network | Traffic management | Detection | Public transit | Monitoring | Clustering |
| 8 | Pattern | Services | Algorithm | Pad traffic | Speed | Service | Monitoring system | Location |
| 9 | Area | Platform | Traffic flow prediction | Network | Vehicular | Passengers | Track | Taxi |
| 10 | Network | Logistics | Base | Analysis | Safety | Route | Inspection | Mining |
| 11 | Method | Real time | Predict | Traffic control | Intelligent transportation system | Scheduling | Analysis | Detection |
| 12 | Accessibility | Intelligent | Feature | Road | Control | Time | Operation | Method |

**Table 4** (continued)

| Top Words | Topic 1<br>Urban transport planning | Topic 2<br>Smart cities and applications | Topic 3<br>Traffic flow forecasting and modeling | Topic 4<br>Traffic management | Topic 5<br>Intelligent transportation systems | Topic 6<br>Public transportation | Topic 7<br>Railway | Topic 8<br>GPS trajectories |
|---|---|---|---|---|---|---|---|---|
| 13 | Activity | Internet of things (IoT) | Prediction model | Urban traffic | Autonomous vehicles | Smart card | Platform | Base |
| 14 | Tourism | Infrastructure | Machine Learning | Road network | Vehicular network | Efficiency | Risk | Clustering algorithm |
| 15 | Transportatio | Transport | Time Series | Intersection | Electric vehicles | Services | Passenger | Ship |

Topic 1 (742 papers), "Urban transport planning": This topic concerns long-term transportation planning within cities, utilizing big and mainly historical data, such as call detail records from mobile phones and large-scale geo-location data from social media in conjunction with geospatial data, census records, and surveys. The emphasis is on analyzing patterns and trends in the urban environment. Most studies aim to investigate the spatial–temporal characteristics of urban movements, detect land use patterns, reveal urban activity patterns, and analyze travel behavior. Moreover, many papers focus on travel demand and origin–destination flow estimation or extract travel attributes, such as trip purpose and activity location.

Topic 2 (723 papers), "Smart cities and applications": This topic remains largely consistent with the previous categorization. As above, the papers aim to take advantage of the various and diverse data generated in cities, analyze new challenges, and propose real-time applications to enhance the daily lives of individuals and city operation systems.

Topic 3 (438 papers), "Traffic flow forecasting and modeling": This area of research involves the use of machine and deep learning techniques to analyze mainly historical data aiming to improve traffic prediction accuracy. The majority of these papers concentrate on short-term traffic flow forecasting, while a significant number of them address passenger flow and traffic accident prediction.

Topic 4 (231 papers), "Traffic management": this topic concentrates on traffic management and real-time traffic control. City traffic monitoring, real-time video, and image processing techniques are gaining significant attention. Numerous studies utilize real-time data to evaluate traffic conditions by image-processing algorithms and provide real-time route selection guidance to users. Most of them manage to identify and resolve traffic congestion problems, as well as to detect anomalies or road traffic events, aiming to improve traffic operation and safety.

Topic 5 (194 papers), "Intelligent transportation systems and new technologies": this topic remains nearly identical to the prior (4-clusters) classification, containing articles on emerging technologies implemented in an intelligent and eco-friendly transport system. Most studies focus on the connected vehicle-infrastructure environment and connected vehicle technologies as a promising direction for improving transportation system performance. Great attention is also given to Internet of Vehicles (IoV) technology and the efficient management of the generated and collected data. Autonomous and self-driving vehicle technologies are also crucial topics. Many papers, also, discuss green transportation and suggest ways to optimize emissions in urban areas, with a particular emphasis on vehicle electrification.

Topic 6 (144 papers), "Public transportation": since public transportation gained special scientific interest in our database, a separate topic was created regarding public transport policy making, service, and management. Most publications focus on urban means of transport, such as buses, metro, and taxis, while a significant proportion refers to airlines. This topic covers studies related to public transportation network planning and optimization, performance evaluation, bus operation scheduling, analysis of passenger transit patterns, maximization of passenger satisfaction levels, and real-time transport applications. Moreover, smart cards and GPS data are extensively used to estimate origin–destination matrices.

Topic 7 (104 papers), "Railway": this topic presents research papers that apply big data to railway transportation systems and engineering, encompassing three areas of interest: railway operations, maintenance, and safety. A significant proportion of studies focus on railway operations, including train delay prediction, timetabling improvement, and demand forecasting. Additionally, numerous researchers employ big data to support maintenance decisions and conduct risk analysis of railway networks, such as train derailments and failure prediction. These papers rely on diverse datasets, including GPS data, passenger travel information, as well as inspection, detectors, and failure data.

Topic 8 (95 papers), "GPS Trajectories": This topic contains papers that take advantage of trajectory data primarily obtained from GPS devices installed in taxis. Most studies forecast the trip purpose of taxi passengers, trip destination, and travel time by analyzing historical GPS data. Additionally, a significant number of these studies focus on real-time analysis to provide passengers with useful applications and enhance the quality of taxi services. Finally, there is research interest in maritime routing and ship trajectory analysis to facilitate maritime traffic planning and service optimization.

In the eight-topic classification, the initial four clusters either remained almost unchanged or were divided into subcategories. For example, the previous cluster "transport planning and travel behavior analysis" is now divided into "urban transport planning" and "public transportation", with "transport management" constituting a separate category. Moreover, several distinct smaller clusters have been identified (e.g., "railway" and "trajectories"). These, along with "public transportation", are highly specialized categories with no correlation to the other clusters. Nevertheless, they constitute a significant proportion of the literature and merit separate analysis.

As the number of topics increased, so did the overlaps among the clusters. Thus, based on this observation and the accuracy results of the validation method, it was assumed that eight clusters were the most appropriate for further analysis.

Based on the results of eight-topic classification, Fig. 3 demonstrates the evolution of the number of published articles per topic and per year. As shown three topics have gained researchers' interest: (1) urban transport planning, (2) smart cities and applications, and (3) traffic forecasting and modeling. Initially, the primary topic was "smart cities", largely based on the computer science sector. Despite a slight decline in publications in 2019, there is an overall upward trend. "Urban transport planning" experienced a steady and notable increase until 2019. A sudden drop was recorded in 2022, but it is not clear whether this is a coincidence or a trend. However, it remains the dominant topic, with most publications over the years. The observed decrease could indicate further specialization and research evolution

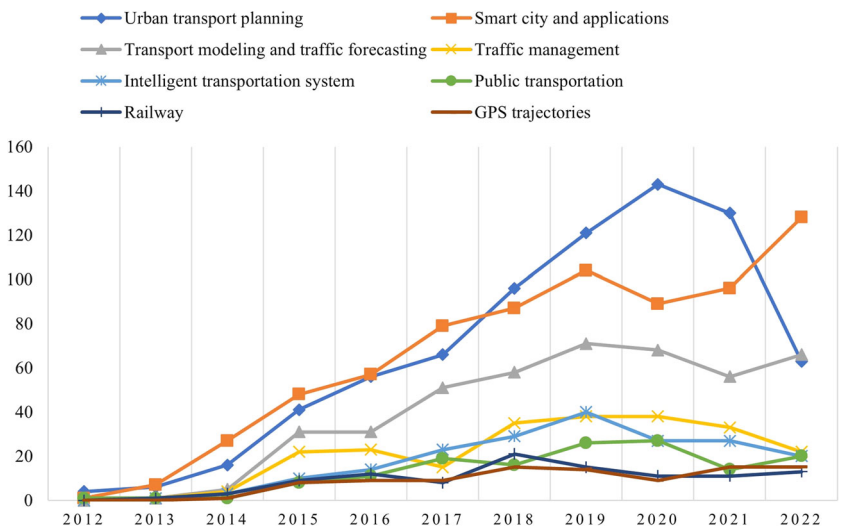



**Fig. 3** Number of papers per topic and year

in the field, given it is also the topic with the most subcategories during the classification process.

## 5 Discussion

### 5.1 Statistical analysis

As shown in the analysis, there is an increasing research interest in big data usage in the transportation sector. It is remarkable that besides computer science journals and conferences, transportation journals have also published relevant articles representing a notable proportion of the research and indicating that transportation researchers acknowledge the significance of big data and its contribution to many aspects of transportation. According to the citation analysis, three research areas emerged among the most influential studies: (1) traffic flow prediction and management (2) new challenges of the cities (smart cities) and new technologies (3) urban transport planning and spatial analysis.

### 5.2 Topic classification

Following the topic model results, eight paper groups are proposed. Most articles (742) fall into the topic of "urban transport planning". Several representative papers in this area attempted to estimate travel demand [5], analyze travel behavior [6], or investigate activity patterns [51] by utilizing big data sourced primarily from mobile phone records, social media, and smart card fare collection systems.

Big data also has significant impacts on "smart cities and applications". Topic 2 is a substantial part of the dataset, which includes 723 papers. They mainly refer to new challenges arising from big data analytics to support various aspects of the city, such as transportation or energy [44] and investigate big data applications in intelligent transportation systems [26] or in the supply chain and logistics [54].

A total of 438 papers were categorized in Topic 3 labeled as "traffic flow forecasting and modeling". The majority applied big data and machine learning techniques to predict traffic flow [41–43]. In risk assessment, Chen et al. [55] proposed a logit model to analyze hourly crash likelihood, considering temporal driving environmental data, whereas Yuan et al. [56] applied a Convolutional Long Short-Term Memory (ConvLSTM) neural network model to forecast traffic accidents.

Among the papers, a special focus is given to different aspects of "traffic management" (231 papers), largely utilizing real-time data. Shi and Abdel-Aty [7] employed random forest and Bayesian inference techniques in real-time crash prediction models to reduce traffic congestion and crash risk. Riswan et al. [57] developed a real-time traffic management system based on IoT devices and sensors to capture real-time traffic information. Meanwhile, He Z. et al. [58] suggested a low-frequency probe vehicle data (PVD)-based method to identify traffic congestion at intersections to solve traffic congestion problems.

Topic 5 includes 194 records on "intelligent transportation systems and new technologies". It covers topics such as Internet of Vehicles [48, 59, 60], connected vehicle-infrastructure environment [61], electric vehicles [62], and the optimization of charging stations location [63], as well as autonomous vehicles (AV) and self-driving vehicle technology [64].

In recent years, three smaller and more specialized topics have gained interest. Within Topic 6, there are 144 papers discussing public transport. Tu et al. [65] examined the use of

smart card data and GPS trajectories to explore multi-modal public ridership. Wang et al. [66] proposed a three-layer management system to support urban mobility with a focus on bus transportation. Tsai et al. [67] applied simulated annealing (SA) along with a deep neural network (DNN) to forecast the number of bus passengers. Liu and Yen [68] applied big data analytics to optimize customer complaint services and enhance management process in the public transportation system.

Topic 7 contains 104 papers on how big data is applied in "railway network", focusing on three sectors of railway transportation and engineering. As mentioned in Ghofrani et al. [24], these sectors are maintenance [69–71], operation [72, 73] and safety [74].

Topic 8 (95 papers) refers mainly to data deriving from "GPS trajectories". Most researchers utilized GPS data from taxis to infer the trip purposes of taxi passengers [75], explore mobility patterns [76], estimate travel time [77], and provide real-time applications for taxi service improvement [78, 79]. Additionally, there are papers included in this topic that investigate ship routes. Zhang et al. [80] utilized ship trajectory data to infer their behavior patterns and applied the Ant Colony Algorithm to deduce an optimal route to the destination, given a starting location, while Gan et al. [81] predicted ship travel trajectories using historical trajectory data and other factors, such as ship speed, with the Artificial Neural Network (ANN) model.

## 6 Conclusions

An extensive overview of the literature on big data and transportation from 2012 to 2022 was conducted using bibliometric techniques and topic model classification. This paper presents a comprehensive methodology for evaluating and selecting the appropriate literature. It identifies eight sub-areas of research and highlights current trends. The limitations of the study are as follows: (1) The dataset came up by using a single bibliographic database (Scopus). (2) Research sources, such as book chapters, were excluded. (3) Expanding the keyword combinations could result in a more comprehensive review. Despite these limitations, it is claimed that the reviewed dataset is representative, leading to accurate findings.

In the process of selecting the suitable literature, various criteria were defined in the Scopus database search, including the language, subject area, and document type. Subsequently, duplicate and non-scientific records were removed. However, the last screening of the titles and abstracts to determine the relevance of the studies to the paper's research interests was conducted manually. This could not be possible for a larger dataset. Additionally, as previously stated, the dataset was divided into eight distinct topics due to multiple overlaps caused by an increase in the number of topics. Nevertheless, the topic of "smart cities and applications" remains broad, even with this division. This makes it challenging to gain in-depth insights into the field and identify specific applications, unlike in "transport planning", where two additional topics were generated by the further classification. Applying the classification model to each topic separately could potentially overcome these constraints by revealing more precise applications and filtering out irrelevant studies.

Despite the above limitations and constraints, the current study provides an effective methodology for mapping the field of interest as a necessary step to define the areas of successful applications and identify open challenges and sub-problems that should be further investigated. It is worth mentioning that there is an intense demand from public authorities for a better understanding of the potential of big data applications in the transport domain towards more sustainable mobility [82]. In this direction, our methodology, along with the

necessary literature review and discussion with relevant experts, can assist policymakers and transport stakeholders in identifying the specific domains in which big data can be applied effectively and planning future transport investments accordingly.

Having defined the critical areas of big data implementation within transportation, trends, and effective applications, the aim is to conduct a thorough literature review in a subarea of interest. This will focus on transport planning and modeling, and public transportation, which appears to be highly promising, based on our findings. A more extensive literature review and content analysis of key studies are crucial to further examine open challenges and subproblems as well as to investigate the applied methodologies for possible revision or further development.

The current study provides a broad overview of the applications of big data in transport areas, which is the initial step in understanding the characteristics and limitations of present challenges and opportunities for further research in the field.

**Acknowledgements** This research is financed by the Research, Innovation and Excellence Program of the University of Thessaly.

**Author contributions** The authors confirm their contribution to the paper as follows: study conception and design done by DT and EN; DT helped in data collection; DT, EN, and KK done analysis and interpretation of results, draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

**Funding** Open access funding provided by HEAL-Link Greece. This research is supported by the Research, Innovation and Excellence Program of the University of Thessaly.

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.



## References

1. European Commission (2021) Sustainable and Smart Mobility Strategy. https://ec.europa.eu/transport/sites/default/files/2021-mobility-strategy-and-action-plan.pdf. Accessed 10 Jul 2021
2. Anda C, Erath A, Fourie PJ (2017) Transport modelling in the age of big data. Int J Urban Sci 21:19–42. https://doi.org/10.1080/12265934.2017.1281150
3. Zhong RY, Huang GQ, Lan S, Dai QY, Chen X, Zhang T (2015) A big data approach for logistics trajectory discovery from RFID-enabled production data. Int J Prod Econ 165:260–272. https://doi.org/10.1016/j.ijpe.2015.02.014
4. Nallaperuma D, Nawaratne R, Bandaragoda T, Adikari A, Nguyen S, Kempitiya T, de Silva D, Alahakoon D, Pothuhera D (2019) Online incremental machine learning platform for big data-driven smart traffic management. IEEE Trans Intell Transp Syst 20:4679–4690. https://doi.org/10.1109/TITS.2019.2924883
5. Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, González MC (2015) The path most traveled: travel demand estimation using big data resources. Transp Res Part C Emerg Technol 58:162–177. https://doi.org/10.1016/j.trc.2015.04.022

6. Chen C, Ma J, Susilo Y, Liu Y, Wang M (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp Res Part C Emerg Technol 68:285–299
7. Shi Q, Abdel-Aty M (2015) Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp Res Part C Emerg Technol 58:380–394. https://doi.org/10.1016/j.trc.2015.02.022
8. Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin-destination matrices using mobile phone call data. Transp Res Part C Emerg Technol 40:63–74. https://doi.org/10.1016/j.trc.2014.01.002
9. Alexander L, Jiang S, Murga M, González MC (2015) Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C Emerg Technol 58:240–250. https://doi.org/10.1016/j.trc.2015.02.018
10. Zannat KE, Choudhury CF (2019) Emerging big data sources for public transport planning: a systematic review on current state of art and future research directions. J Indian Inst Sci 99:601–619
11. Choi TM, Wallace SW, Wang Y (2018) Big data analytics in operations management. Prod Oper Manag 27:1868–1883. https://doi.org/10.1111/poms.12838
12. Chalmeta R, Santos-deLeón NJ (2020) Sustainable supply chain in the era of industry 4.0 and big data: a systematic analysis of literature and research. Sustainability 12:4108. https://doi.org/10.3390/su12104108
13. De Bakker FGA, Groenewegen P, Den Hond F (2005) A bibliometric analysis of 30 years of research and theory on corporate social responsibility and corporate social performance. Bus Soc 44:283–317. https://doi.org/10.1177/0007650305278086
14. Mishra D, Gunasekaran A, Papadopoulos T, Childe SJ (2018) Big data and supply chain management: a review and bibliometric analysis. Ann Oper Res 270:313–336. https://doi.org/10.1007/s10479-016-2236-y
15. Fahimnia B, Sarkis J, Davarzani H (2015) Green supply chain management: a review and bibliometric analysis. Int J Prod Econ 162:101–114. https://doi.org/10.1016/j.ijpe.2015.01.003
16. Okafor O (2020) Automatic Topic classification of research papers using the NLP topic model NMF. https://obianuju-c-okafor.medium.com/automatic-topic-classification-of-research-papers-using-the-nlp-topic-model-nmf-d4365987ec82f. Accessed 10 Jul 2021
17. Iliashenko O, Iliashenko V, Lukyanchenko E (2021) Big data in transport modelling and planning. Transp Res Proced 54:900–908
18. Wang Z, He SY, Leung Y (2018) Applying mobile phone data to travel behaviour research: a literature review. Travel Behav Soc 11:141–155. https://doi.org/10.1016/j.tbs.2017.02.005
19. Huang H, Cheng Y, Weibel R (2019) Transport mode detection based on mobile phone network data: a systematic review. Transp Res Part C Emerg Technol 101:297–312
20. Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C Emerg Technol 19:557–568. https://doi.org/10.1016/j.trc.2010.12.003
21. Lana I, Del Ser J, Velez M, Vlahogianni EI (2018) Road traffic forecasting: recent advances and new challenges. IEEE Intell Transp Syst Mag 10:93–109
22. Miglani A, Kumar N (2019) Deep learning models for traffic flow prediction in autonomous vehicles: a review, solutions, and challenges. Veh Commun 20:100184. https://doi.org/10.1016/j.vehcom.2019.100184
23. Pender B, Currie G, Delbosc A, Shiwakoti N (2014) Social media use during unplanned transit network disruptions: a review of literature. Transp Rev 34:501–521. https://doi.org/10.1080/01441647.2014.915442
24. Ghofrani F, He Q, Goverde RMP, Liu X (2018) Recent applications of big data analytics in railway transportation systems: a survey. Transp Res Part C Emerg Technol 90:226–246. https://doi.org/10.1016/j.trc.2018.03.010
25. Borgi T, Zoghlami N, Abed M (2017). Big data for transport and logistics: a review. In: International conference on advanced systems and electric technologies (IC_ASET), pp 44–49
26. Zhu L, Yu FR, Wang Y, Ning B, Tang T (2019) Big data analytics in intelligent transportation systems: a survey. IEEE Trans Intell Transp Syst 20:383–398
27. Neilson A, Indratmo DB, Tjandra S (2019) Systematic review of the literature on big data in the transportation domain: concepts and applications. Big Data Res 17:35–44
28. Katrakazas C, Antoniou C, Sobrino N, Trochidis I, Arampatzis S (2019). Big data and emerging transportation challenges: findings from the NOESIS project. In: 6th IEEE International conference on models and technologies for intelligent transportation systems (MT-ITS), pp 1–9
29. Pranckutė R (2021) Web of science (WoS) and scopus: the titans of bibliographic information in today's academic world. Publications 9(1):12. https://doi.org/10.3390/publications9010012
30. Elsevier Scopus (2023) Content coverage guide. https://www.elsevier.com/?a=69451. Accessed 27 Sept 2023

31. Jiang Z, Gao B, He Y, Han Y, Doyle P, Zhu Q (2021) Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports. Math Probl Eng. https://doi.org/10.1155/2021/6619088
32. Zhang X, Delpha C, Diallo D (2019) Performance of Jensen Shannon divergence in incipient fault detection and estimation. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2742–2746
33. Ruz GA, Henríquez PA, Mascareño A (2020) Sentiment analysis of twitter data during critical events through Bayesian networks classifiers. Futur Gener Comput Syst 106:92–104. https://doi.org/10.1016/j.future.2020.01.005
34. Kumar A, Srinivasan K, Cheng WH, Zomaya AY (2020) Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Inf Process Manag 57:102–141. https://doi.org/10.1016/j.ipm.2019.102141
35. Pimpalkar AP, Retna Raj RJ (2020) Influence of pre-processing strategies on the performance of ML classifiers exploiting TF-IDF and BOW features. ADCAIJ Adv Distrib Comput Artif Intell J 9:49–68. https://doi.org/10.14201/adcaij2020924968
36. YueTing H, YiJia X, ZiHe C, Xin T (2019) Short text clustering algorithm based on synonyms and k-means. Computer knowledge and technology 15(1).
37. Bracewell DB, Yan J, Ren F, Kuroiwa S (2009) Category classification and topic discovery of japanese and english news articles. Electron Notes Theor Comput Sci 225:51–65. https://doi.org/10.1016/j.entcs.2008.12.066
38. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. In: 52nd Annual meeting of the association for computational linguistics, pp 655–665
39. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: International conference on engineering and technology (ICET), pp 1–6
40. Ertam F, Aydn G (2017) Data classification with deep learning using tensorflow. In: International conference on computer science and engineering (UBMK), pp 755–758
41. Lv Y, Duan Y, Kang W, Li Z, Wang FY (2015) Traffic flow prediction with big data: a deep learning approach. IEEE Trans Intell Transp Syst 16:865–873. https://doi.org/10.1109/TITS.2014.2345663
42. Wu Y, Tan H, Qin L, Ran B, Jiang Z (2018) A hybrid deep learning based traffic flow prediction method and its understanding. Transp Res Part C Emerg Technol 90:166–180. https://doi.org/10.1016/j.trc.2018.03.001
43. Tian Y, Pan L (2015) Predicting short-term traffic flow by long short-term memory recurrent neural network. In: IEEE International conference on smart city/socialcom/sustaincom (SmartCity). IEEE, pp 153–158
44. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J (2015) Applications of big data to smart cities. J Internet Serv Appl 6:1–15. https://doi.org/10.1186/s13174-015-0041-5
45. Batty M (2013) Big data, smart cities and city planning. Dialog Hum Geogr 3:274–279. https://doi.org/10.1177/2043820613513390
46. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol 5(3):1–55. https://doi.org/10.1145/2629592
47. Mehmood Y, Ahmad F, Yaqoob I, Adnane A, Imran M, Guizani S (2017) Internet-of-things-based smart cities: recent advances and challenges. IEEE Commun Mag 55:16–24. https://doi.org/10.1109/MCOM.2017.1600514
48. Xu W, Zhou H, Cheng N, Lyu F, Shi W, Chen J, Shen X (2018) Internet of vehicles in big data era. IEEE/CAA J Autom Sin 5:19–35. https://doi.org/10.1109/JAS.2017.7510736
49. Zhong C, Arisona SM, Huang X, Batty M, Schmitt G (2014) Detecting the dynamics of urban structure through spatial network analysis. Int J Geogr Inf Sci 28:2178–2199. https://doi.org/10.1080/13658816.2014.914521
50. Yao H, Wu F, Ke J, Tang X, Jia Y, Lu S, Gong P, Ye J, Chuxing D, Li Z (2018) Deep multi-view spatial-temporal network for taxi demand prediction. In: AAAI Conference on artificial intelligence. pp 2588–2595
51. Hasan S, Ukkusuri SV (2014) Urban activity pattern classification using topic models from online geo-location data. Transp Res Part C Emerg Technol 44:363–381. https://doi.org/10.1016/j.trc.2014.04.003
52. Saeed N, Nam H, Haq MIU, Saqibm DBM (2018) A survey on multidimensional scaling. ACM Comput Surv (CSUR) 51:1–25
53. Hout MC, Papesh MH, Goldinger SD (2012) Multidimensional scaling. Wiley Interdiscip Rev Cogn Sci 4:93–103
54. Kaur H, Singh SP (2018) Heuristic modeling for sustainable procurement and logistics in a supply chain using big data. Comput Oper Res 98:301–321. https://doi.org/10.1016/j.cor.2017.05.008

55. Chen F, Chen S, Ma X (2018) Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. J Saf Res 65:153–159. https://doi.org/10.1016/j.jsr.2018.02.010
56. Yuan Z, Zhou X, Yang T (2018) Hetero-ConvLSTM: a deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, pp 984–992
57. Riswan P, Suresh K, Babu MR (2016) Real-time smart traffic management system for smart cities by using internet of things and big data. In: ICETT - 2016 : international conference on emerging technological trends in computing, communications and electrical engineering. IEEE, pp 1–7
58. He Z, Qi G, Lu L, Chen Y (2019) Network-wide identification of turn-level intersection congestion using only low-frequency probe vehicle data. Transp Res Part C Emerg Technol 108:320–339. https://doi.org/10.1016/j.trc.2019.10.001
59. Zhou Z, Gao C, Xu C, Zhang Y, Mumtaz S, Rodriguez J (2018) Social big-data-based content dissemination in internet of vehicles. IEEE Trans Ind Inf 14:768–777. https://doi.org/10.1109/TII.2017.2733001
60. Guo L, Dong M, Ota K, Li Q, Ye T, Wu J, Li J (2017) A secure mechanism for big data collection in large scale internet of vehicle. IEEE Internet Things J 4:601–610
61. Sumalee A, Ho HW (2018) Smarter and more connected: future intelligent transportation system. IATSS Res 42:67–71
62. Fetene GM, Kaplan S, Mabit SL, Jensen AF, Prato CG (2017) Harnessing big data for estimating the energy consumption and driving range of electric vehicles. Transp Res D Transp Environ 54:1–11. https://doi.org/10.1016/j.trd.2017.04.013
63. Tu W, Li Q, Fang Z, Shaw S, lung, Zhou B, Chang X, (2016) Optimizing the locations of electric taxi charging stations: a spatial–temporal demand coverage approach. Transp Res Part C Emerg Technol 65:172–189. https://doi.org/10.1016/j.trc.2015.10.004
64. Najada HA, Mahgoub I (2016) Autonomous vehicles safe-optimal trajectory selection based on big data analysis and predefined user preferences. In: EEE 7th annual ubiquitous computing, electronics mobile communication conference (UEMCON). IEEE, pp 1–6
65. Tu W, Cao R, Yue Y, Zhou B, Li Q, Li Q (2018) Spatial variations in urban public ridership derived from GPS trajectories and smart card data. J Transp Geogr 69:45–57. https://doi.org/10.1016/j.jtrangeo.2018.04.013
66. Wang Y, Ram S, Currim F, Dantas E, Sabóia L (2016) A big data approach for smart transportation management on bus network. In: IEEE international smart cities conference (ISC2), pp 1–6
67. Tsai CW, Hsia CH, Yang SJ, Liu SJ, Fang ZY (2020) Optimizing hyperparameters of deep learning in predicting bus passengers based on simulated annealing. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2020.106068
68. Liu WK, Yen CC (2016) Optimizing bus passenger complaint service through big data analysis: systematized analysis for improved public sector management. Sustainability 8:1319. https://doi.org/10.3390/su8121319
69. Li H, Parikh D, He Q, Qian B, Li Z, Fang D, Hampapur A (2014) Improving rail network velocity: a machine learning approach to predictive maintenance. Transp Res Part C Emerg Technol 45:17–26. https://doi.org/10.1016/j.trc.2014.04.013
70. Sharma S, Cui Y, He Q, Mohammadi R, Li Z (2018) Data-driven optimization of railway maintenance for track geometry. Transp Res Part C Emerg Technol 90:34–58. https://doi.org/10.1016/j.trc.2018.02.019
71. Jamshidi A, Hajizadeh S, Su Z, Naeimi M, Núñez A, Dollevoet R, de Schutter B, Li Z (2018) A decision support approach for condition-based maintenance of rails based on big data analysis. Transp Res Part C Emerg Technol 95:185–206. https://doi.org/10.1016/j.trc.2018.07.007
72. Thaduri A, Galar D, Kumar U (2015) Railway assets: a potential domain for big data analytics. Proced Comput Sci 53:457–467. https://doi.org/10.1016/j.procs.2015.07.323
73. Oneto L, Fumeo E, Clerico G, Canepa R, Papa F, Dambra C, Mazzino N, Anguita D (2017) Dynamic delay predictions for large-scale railway networks: deep and shallow extreme learning machines tuned via thresholdout. IEEE Trans Syst Man Cybern Syst 47:2754–2767. https://doi.org/10.1109/TSMC.2017.2693209
74. Sadler J, Griffin D, Gilchrist A, Austin J, Kit O, Heavisides J (2016) GeoSRM: online geospatial safety risk model for the GB rail network. IET Intell Transp Syst 10(1):17–24. https://doi.org/10.1049/iet-its.2015.0038
75. Gong L, Liu X, Wu L, Liu Y (2016) Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartogr Geogr Inf Sci 43:103–114. https://doi.org/10.1080/15230406.2015.1014424
76. Xia F, Wang J, Kong X, Wang Z, Li J, Liu C (2018) Exploring human mobility patterns in urban scenarios: a trajectory data perspective. IEEE Commun Mag 56:142–149. https://doi.org/10.1109/MCOM.2018.1700242

77. Qiu J, Du L, Zhang D, Su S, Tian Z (2020) Nei-TTE: intelligent traffic time estimation based on fine-grained time derivation of road segments for smart city. IEEE Trans Ind Inf 16:2659–2666. https://doi.org/10.1109/TII.2019.2943906
78. Zhou Z, Dou W, Jia G, Hu C, Xu X, Wu X, Pan J (2016) A method for real-time trajectory monitoring to improve taxi service using GPS big data. Inf Manag 53:964–977. https://doi.org/10.1016/j.im.2016.04.004
79. Xu X, Zhou JY, Liu Y, Xu ZZ, Zha XW (2015) Taxi-RS: taxi-hunting recommendation system based on taxi GPS data. IEEE Trans Intell Transp Syst 16:1716–1727. https://doi.org/10.1109/TITS.2014.2371815
80. Zhang SK, Shi GY, Liu ZJ, Zhao ZW, Wu ZL (2018) Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity. Ocean Eng 155:240–250. https://doi.org/10.1016/j.oceaneng.2018.02.060
81. Gan S, Liang S, Li K, Deng J, Cheng T (2016) Ship trajectory prediction for intelligent traffic management using clustering and ANN. In: 2016 UKACC 11th international conference on control (CONTROL), pp 1–6
82. European Union (EU) Horizon 2020 (H2020) (2017) NOESIS: novel decision support tool for evaluating strategic big data investments in transport and intelligent mobility services. https://cordis.europa.eu/programme/id/H2020_MG-8-2-2017/en. Accessed 29 Sep 2023



**Danai Tzika-Kostopoulou** holds an Integrated Masters Diploma from the School of Civil Engineering of the National Technical University of Athens (NTUA) with a specialization in Transport and Traffic Engineering. She received her M.Sc. in Urban and Regional Planning from the School of Architecture of NTUA in 2012. She is currently a Ph.D. student at the Department of Civil Engineering of the University of Thessaly under the supervision of the Professor Eftihia Nathanail. She has received a scholarship from the Center of Research, Innovation and Excellence of the University of Thessaly for her thesis: "Utilizing location-based social networking data to estimate travel demand in urban areas. ". Her current research interests are in transport planning and modelling and in big data analytics in transportation.

**Eftihia Nathanail** is a Professor of Transportation Systems Design and Evaluation, founder and director of the Traffic, Transportation and Logistics Laboratory (TTLog) and Rector of the School of Engineering of the University of Thessaly, visiting professor at the University of Hawaii at Manoa, in USA and Transporta un sakaru institūts, in Latvia. She participates as Dangerous Goods Expert in the Transport Group Inland Surface Transport, NATO and is a Member of the Committee of Transportation Safety Management Systems (ACS10) and the Risk Management Subcommittee (AT040) of the Transportation Research Board and the Committee of Sustainable Urban Mobility in Volos. She holds an Integrated Masters Diploma from the Faculty of Surveying Engineering, Aristotle University of Thessaloniki, an MSc in Transportation Engineering from the Faculty of Civil, ArchtecturalArchitectural and Environmental Engineering, University of Miami and a PhD from the Faculty of Civil Engineering, Aristotle University of Thessaloniki. Her fields of research are sustainable transportation planning, transportation system design, intelligent transportation systems, behavioral modeling, intermodal transportation, logistics, multicriteria evaluation and optimization.

**Konstantinos Kokkinos** was awarded a B.Sc. in Physics, from the Aristotle's University of Thessaloniki, Greece (1989), a M.Sc. and a PhD in Computer Science from C.S. Dept. Western Michigan University, U.S.A. (1995 and 2002 respectively). Dr. Kokkinos is an active member of IEEE Society, an OpenMI-Association member, and an Asst. Professor of Digital Systems Dept. of University of Thessaly Greece. Dr. Kokkinos's current and prior participation in research and innovation projects includes 6 NSF (USA funded), 4 Horizon/3 FP/1 Marie Curie (Europe funded), 4 Erasmus+ and 6 Greek national funded. Konstantinos Kokkinos serves for many years the research community as a reviewer of several Elsevier, IEEE, Springer and MDPI journals. His major research interests include Decision Support Systems, Fuzzy Logic and Systems, Soft Computing, Multi-Criteria Decision Making, Integrated Modelling and Simulation, Artificial Intelligence and Soft Computing Methodologies, Water and Environmental Sustainability, Network Congestion, Load Balancing and Web Services. Furthermore, Dr. Kokkinos has published over 65 journal and conference papers.