SN Computer Science (2024) 5:152
https://doi.org/10.1007/s42979-023-02457-x

ORIGINAL RESEARCH

Literature Review on the Smart City Resources Analysis with Big Data Methodologies

Regina Gubareva[1] · **Rui Pedro Lopes**[1]

Received: 19 March 2023 / Accepted: 27 October 2023

Abstract
This article provides a systematic literature review on applying different algorithms to municipal data processing, aiming to understand how the data were collected, stored, pre-processed, and analyzed, to compare various methods, and to select feasible solutions for further research. Several algorithms and data types are considered, finding that clustering, classification, correlation, anomaly detection, and prediction algorithms are frequently used. As expected, the data is of several types, ranging from sensor data to images. It is a considerable challenge, although several algorithms work very well, such as Long Short-Term Memory (LSTM) for timeseries prediction and classification.

Keywords Big data · Smart city · Resources consumption

Introduction

The smart city concept is popular in scientific literature, characterizing a healthy environment that improves the quality of life and well-being of citizens [1]. Nowadays, most operations are controlled via comprehensive information and communication technologies. Due to the diversity of services, resources, and projects, smart cities manage huge amounts of data, typically within the Big Data concept.

Over the past ten years, the number of sensors and metering devices has been increasing geometrically. The intention to control and understand everything surrounding us became a significant step in the development of technologies of environmental sensors: smart houses, smart cities, smart devices, Internet of Things (IoT), and many others. Legacy information is also laying around, in spreadsheets or databases, which can be valuable if correctly accessed and integrated. Citizens and institutions also make use of social networks to convey opinions, criticism, or information about resources, services, or events. The essential questions are how to use this data and how to extract practical and meaningful information from all these measurements.

This work is developed within the project "PandIA - Management of Pandemic Social Isolation Based on City and Social Intelligence", which focus on providing detailed information, such as resource consumption trends, estimation of people in each area or household, a heat map of suspected outbreaks, and others, to health, municipal authorities and to emergency personal. For that, it uses information from several sources, including pathogen characteristics, infection statistics, municipal information, social networks, and hospital information and statistics.

The work described in this paper uses a systematic literature review to understand the nature and purpose of the data generated and collected in the context of a city. It aims to understand what types of data are usually considered, how they were collected, what algorithms are used, and for what purposes. We look for evidence and best practices for using city information in Big Data settings, impact, and results [2, 3]. Broadly speaking, the purpose of this work is to systematize the basic principles of digital data handling in the formation and development of smart cities. The review

Regina Gubareva and Rui Pedro Lopes have contributed equally to this work.

This article is part of the topical collection "Computer Sciences and Artificial Intelligence for Smart Cities" guest edited by Sergio Nesmachnow, Luis Hernández Callejo and Pedro Moreno.

✉ Rui Pedro Lopes
rlopes@ipb.pt

Regina Gubareva
regina.gubareva@ipb.pt

[1] Research Center on Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança, Portugal

Published online: 06 January 2024

summarizes the research being done in the last five years. The literature is categorized according to the algorithms used, the approach to handling data, the data's nature, and the data processing results.

The paper is structured in four sections, starting with this introduction. Section 2 describes the methodology followed in this study. The Result and Analysis follows, with the results and associated discussion and it finishes in Sect. 4 with some conclusions.

Methodology

The main objective of this literature review is to try to understand the data structure and nature, the sources, processes for collecting and storing data, the algorithms and tasks used to analyze the data, and, finally, the purposes or intentions of the analysis process. This literature review follows the approach suggested by Materla, Cudney, and Antony [4] and by Subhash and Cudney [5], including three phases: planning, operation, and dissemination (Fig. 1).

To guide the search, a set of research questions was considered:

1. What is the format, structure, and context of the data each paper uses?
2. What algorithms are used for urban data handling?
3. What is the complexity of each algorithm?
4. What are the results achieved after the analysis and what do they mean?

The papers were searched in Scopus and IEEEXplore. These databases were selected because they provide a wide set of areas and the key terms provide an initial focus on the main objective of this work. A total of 230 papers were identified in the first run (Table 1).

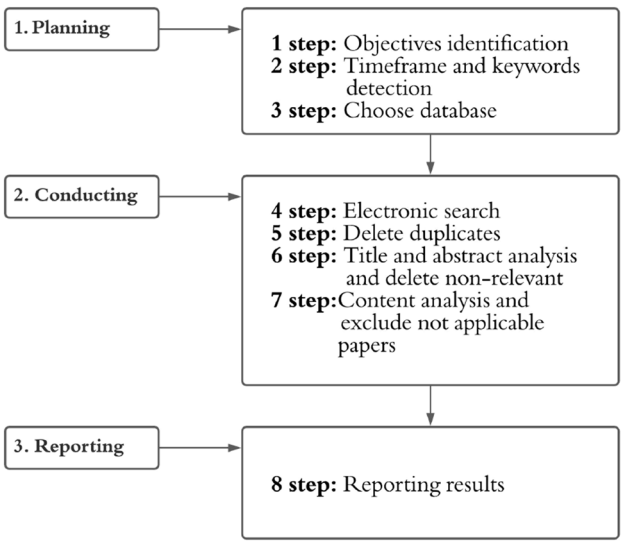

Fig. 1 Phases of the systematic literature review

Only the papers retrieved from Scopus and IEEEXplore published between January 1, 2010, and December 31, 2022, whose text was available in the institutional repositories were considered. Moreover, papers without a peer review process and written in a different language than English were also excluded. After removing the duplicate entries, the total number of papers was 208.

Some guidelines were defined for the title, abstract, and text analysis. After a primary assessment, a detailed analysis of conformity to a chosen theme was made. The title's meaningfulness, associated with the abstract description helped with this. Next, the text was skimmed to assess if all the information needed could be found. So, in summary, the papers were analyzed according to the following steps:

1. Searched articles were limited to the predefined time frame (2010–2022)
2. Papers with non-relevant titles, abstracts, and keywords were excluded
3. Papers with text that did not mention the required subjects was excluded

After analysis of the title, abstract, and text, 196 more papers were excluded for being out of the scope of this work. A total of 20 papers remained for the analysis (Table 2).

Analysis and Discussion

The analysis process started with the characterization of the selected references. Content analysis followed, to assess the context and definition of self-study and the purpose of the work described in the paper. Before delving

Table 1 Search terms and the number of papers retrieved

Database	Search term	Results
Scopus	TITLE ("big data" AND ("urban data" OR "smart city" OR "geo data" OR "social network" OR "predictive maintenance" OR "algorithms"))	123
IEEEXplore	(("Document Title": "big data") AND ("Document Title": "urban data" OR "Document Title": "smart city" OR "Document Title": "geo data" OR "Document Title": "social network" OR "Document Title": "predictive maintenance" OR "Document Title": "algorithms"))	107
	Total	230

Table 2 The results of the search by journals

Phase	Scopus	IEEEXplore	Total
Search	123	107	230
Del. duplicates	114	94	208
Title	42	36	78
Abstract	18	15	33
Content	12	8	20

into the literature analysis, some background is added for completeness.

Background

The concept of smart city emerges withing two main dimensions. The first is related to provide citizens with better services in terms of comfort, mobility, energy and health [6]. The second is related to better resources management. In this context, a smart and sustainable city uses Information and Communications Technologies (ICT) to reduce the city's environmental footprint and improve citizens' quality of life. The use of ICT allows building "smart" applications as well as making cities more sustainable and more pleasant to live in. This is a challenge, particularly when cities bring together a constantly growing population, expand and become denser, with the problems that these imply. On the basis of this lies the collection of a huge amount and diversity of data from several sources and areas [7, 8].

With increasing amounts of data, the uncovering of patterns and extraction of useful information becomes an increasing challenge. Manual analysis and interpretation becomes impossible or infeasible in useful time, due to the difficulty in correlating several sources and nature of information. The analysis of this amount of data has to rely on statistical approaches or on algorithms that allow extracting hidden patterns within data. The diversity of techniques is considerable. Descriptive statistics, correlation analysis, neural networks, or deep learning algorithms are becoming essential tools for these tasks [9].

To be able to extract and understand the information within the data, it is necessary to understand their format and to select the best algorithms to use. Descriptive statistics can be used in structured data, such as spreadsheets, or databases, correlation analysis allow inferring interdependency between different columns or data, neural networks allow predicting, estimating and classifying unstructured data, such as time series, text or sound. Finally, deep learning algorithms can be particularly useful for image classification, video interpretation, among many others. Thus, this work tries to approach these questions in the following sections, as a support for future work in the area of smart cities and water consumption estimation and analysis.

Characterization

In total, papers from 15 different countries were found. The distribution by countries is illustrated in Fig. 2. China has the highest amount of articles. USA and Spain have two papers each, and one paper listed for the remaining for countries: Morocco, Italy, England, Greece, Malaysia, Taiwan, Iceland, India, South Korea, Brunei, Cyprus, and Switzerland.

Only articles published no earlier than 2010 were selected for consideration. The distribution by year reveals a peak number of papers in 2018, with 8 papers. That indicates the increasing interest in this issue. Although in 2019 and 2021,

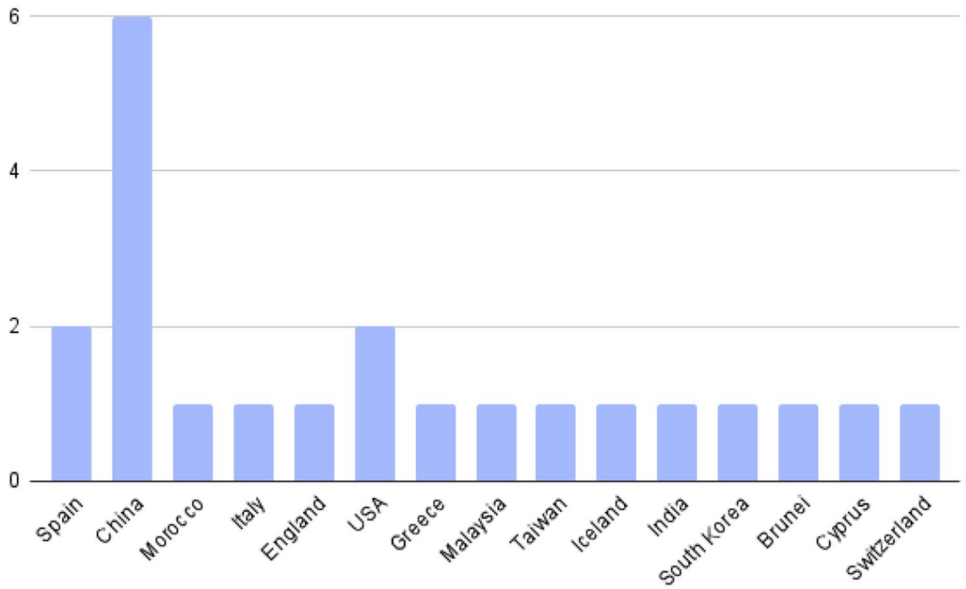

Fig. 2 Papers distribution by countries

there is only one paper and no papers in 2020 which reflects the slump in activity. However, by the number of publications over the last two years, we can observe the growth of attention, which is illustrated in Fig. 3. Despite considerable attention from the scientific community to the issue of the impact of such a resource as digital data on the development of modern socio-economic systems, including cities, this area is only beginning to develop, and the understanding of the use of data as a tool for the development of smart cities remains limited in the scientific literature.

In general, it can be noted that research is increasingly focused on the use of digital data as a new socio-economic phenomenon, and attempts are made to conceptualize, classify and evaluate the role of different types of data in socio-economic processes. In most cases, such studies are related to the use of big data in certain areas of the urban environment, such as transportation, public safety, and environmental protection. At the same time, the literature lacks studies of a general systemic nature on the use of big data for smart cities regardless of the field of application.

Data types and sources

The smart city concept implies integrating multiple information and communication technologies for city infrastructure management: transport, education, health, systems of housing and utilities, safety, etc. Municipal governments collect numerous heterogeneous information, and an "urban data" term can mean various datasets: data from video surveillance cameras, traffic, air quality, energy, and water consumption, and images for smart recognition. Therefore for this study, the essential is to recognize and classify different datasets utilized in considered resources.

Trilles et al. describe a methodology of (big) data process produced by sensors in real time [10]. It assumes that it works with different sensor data sources with different formats and connection interfaces. Wireless Sensor Networks (WSN) are used for monitoring the physical state of the environment: air pollution, forest fire, landslide, and water quality. Although the system proposed by the authors is designed to process all data types, the WSNs mainly produce numerical data like water level, and the gas concentration in the air, mainly classified as quantitative information. An efficient method to derive spatio-temporal analysis of the data, using correlations was proposed by [11]. The authors use data from Bluetooth sensors installed in light poles. The data was collected from the road sensors in the city of Aarhus in Denmark. The measurements are taken every 5 min and the dataset includes a timestamp, location information, average speed, and a total of automobiles at the time of commit. The data were classified as numerical as there are no text, images, sound, or video information.

Bordogna et al. used in their paper big mobile social data, which included users-generated, geo-referenced and timestamped contents [12]. The content means text data that users post in modern emerging social systems like Twitter, Facebook, Instagram, and so forth. Hereby, the dataset can be classified as heterogeneous by way of containing the text of social network posts and numerical data of location and time. Wang et al. considered another approach to analysis and evaluated the effectiveness of deep neural networks [13]. The aim of their paper was the monitoring and control of local HIV epidemics. The collection includes statistics on the number of morbidities, mortality, and mortality by region, age, sex, and occupation. The type of data is categorized as text and numerical.

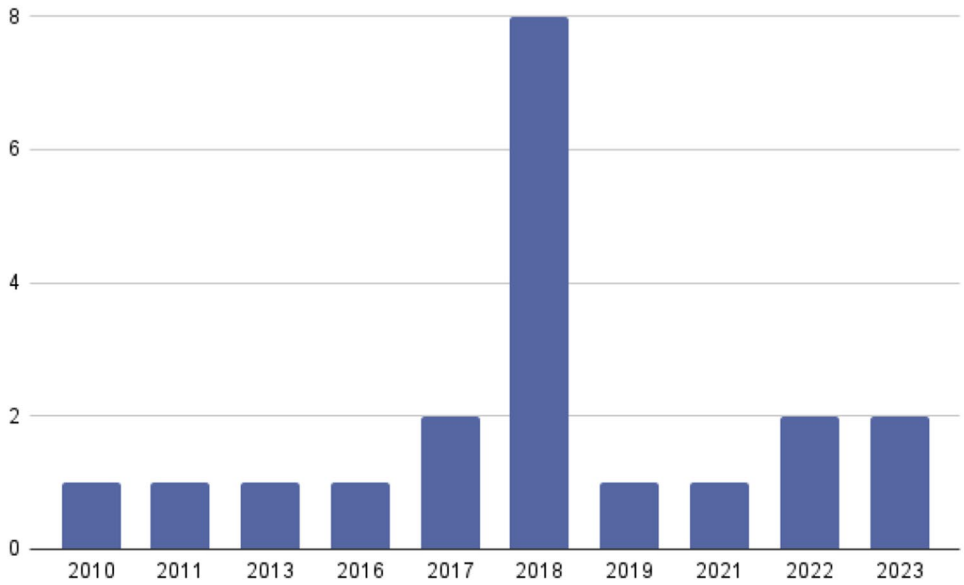

Fig. 3 Papers distribution by years

The work by Pérez-Chacón et al. proposed a methodology to extract electric energy consumption patterns in big data time series [14]. The study used the big data time series of electricity consumption of several Pablo de Olavide University buildings, extracted using smart meters over six years. Karyotis et al. presented a novel data clustering framework for big sensory data produced by Internet of Things (IoT) applications [15]. The dataset was collected from an operational smart-city/building IoT infrastructure provided by the Federated Interoperable Semantic IoT/cloud Testbeds and Applications (FIESTA-IoT) testbed federation. The array is heterogeneous and represents measurements of different types: temperature, humidity, battery level, soil moisture, etc.

Azri et al. presented a technique of three-dimensional data analytics using a dendrogram clustering approach [16]. It is assumed that the algorithm can be applied to large heterogeneous datasets gathered from sensors, social media, and legacy data sources. Alshami et al. tested the performance of two partition algorithms K-Means and Fuzzy c-Mean for clustering big urban datasets [17]. Compared techniques can be applied to huge heterogeneous datasets in various areas like medicine, business, biology, etc. In the paper, the authors utilized urban data from various data sources, such as the Internet of Things, LIDAR data, local weather stations, and mobile phone sensors.

Chang et al. developed a new iterative algorithm, called the K-sets+ algorithm for clustering data points in a semi-metric space, where the distance measure does not necessarily satisfy the triangular inequality [18]. The algorithm is designed for clustering data points in semi-metric space. To understand what semi-metric space is, it is necessary to briefly consider the concept of metrics in space. The metric is the mapping for some set $d : X \times X \rightarrow R$, for which the axioms of non-degeneracy and symmetry have to be satisfied but not necessarily the triangle inequality. If the distance between different points can be zero, the metric is semi-metric. The method was evaluated with two experiments: community detection of signed networks and clustering of real networks. The dataset included 216 servers in different locations, and the latency (measured by the round trip time) between any two servers of these 216 servers is recorded in real-time.

Chae et al. have compared the performance of the deep neural network (DNN), long-short-term memory (LSTM), and the auto-regressive integrated moving average (ARIMA) in predicting three infectious diseases [19]. The study uses four kinds of data to predict infectious diseases, including search query data, social media big data, temperature, and humidity. Data related to malaria, chickenpox, and scarlet fever, for 576 days, were considered. As a result, the data is partly numerical and partly text. The research of Chen et al. focuses on multi-source urban data analysis [20]. The points of interest are geographical, street view, road map, and real-estate data. The record comprises the road network of the city, longitude, latitude, name, and functionality of a structure in the urban environment, and imagery of locations. Obviously, the dataset is ranked as heterogeneous.

Simhachalam and Ganesan presented a multidimensional mining approach in a successive way by finding groups (clusters) of communities with the same multi-dynamic characteristics [21]. The data refers to the statistics of population, migration, tax capacity, dwellings, employment, and commuters.

The majority of the studies assume heterogeneous nature data. There are two research papers with only numerical data and one of the papers investigates image data processing. Text and numerical data are dominant and they are collected from multiple sources (Table 3).

Algorithms

Mainly the goal of this research is to identify the application of modern techniques to the consumption of resources, such as energy and water and, globally speaking, there is an inclusive question: "What useful information can be extracted from the expenditures statistics?" Considering that there are not many investigations in connection with the primary living resources consumption, the authors review various data samples and techniques to identify suitable ones for future reference.

In general, 20 different approaches to big urban data processing were considered. The methods can be divided into groups depending on the manner of information handling: clustering, classification, correlation, deep neural network, frameworks, and community detection. Figure 4 illustrates the proportion between different techniques. The most popular approach is clustering, which is considered in 6 reviewed papers. The same number of times frameworks is offered, which implies the architectures for automatic data processing, inside frameworks, different methods can be applied depending on the purpose, for this reason, authors separated this category. Deep neural networks were used for prediction and occurred only twice. But the studies present the evaluation characteristics for several deep neural network techniques which gives a fairly broad picture. In the category "Classification", three papers were considered. Correlation methods are presented in 3 papers. The classification and the correlation here are the machine learning problem types. Classification combines a whole set of methods and algorithms designed to divide a set of objects, each described by a set of variables, into some number of homogeneous, in a certain sense, classes. These classes may be related to some extent. Correlation is a measure of the relationship of variables to each other. In Machine Learning, it is often

Table 3 Data types and sources

Paper	Data	Category
[10]	Data from different sensors	Heterogeneous
[11]	Traffic data collected from the road sensors in the city: geographical location, time-stamp, average speed, and a total of automobile	Numerical
[12]	Social networks posts, timestamp, geo-location	Heterogeneous
[13]	10-year historical HIV incidence data: the number of morbidities, morbidity, mortality and mortality by region, age, sex, occupation	Heterogeneous
[14]	Electricity consumption for 6 years for several buildings	Numerical
[15]	Big sensory data, measurements of different types: temperature, humidity, battery level, soil moisture	Heterogeneous
[16]	smart city data	Heterogeneous
[17]	Data from the Internet of Things, LIDAR data, local weather stations, mobile phones sensors	Heterogeneous
[18]	Locations and the latency (measured by the round trip time) between any two data points	Heterogeneous
[19]	Search query data, social media big data, temperature, and humidity	Heterogeneous
[20]	Geographical data, points of interests data(longitude, latitude, name, and functionality of a structure in the urban environment), street view data, real estate data, mobile phone location data, social network data, micro-blog data, taxi GPS trajectory data, taxi profile data	Heterogeneous
[21]	The measurements of the blood tests as the corpuscular volume of test substances and the number of half-pint equivalents of alcoholic beverages drunk per day	Numerical
[22]	Multi-source geospatial big data: points of interests, bus smart card transactions, taxi trajectories	Heterogeneous
[23]	Social media, web forums, images, videos, RDBMS, OLAP, Data warehousing, XML formatted data, CSV, HTML, RDF	Heterogeneous
[24]	Hotel reservations data: time, sample size, orders number, user number, old user number	Numeric
[25]	Water and electricity consumption data with timestamp	Numeric
[26]	Water consumption	Numeric
[27]	Henan's gross domestic product, coal consumption, crude oil consumption, natural gas consumption, hydro-power consumption	Numeric

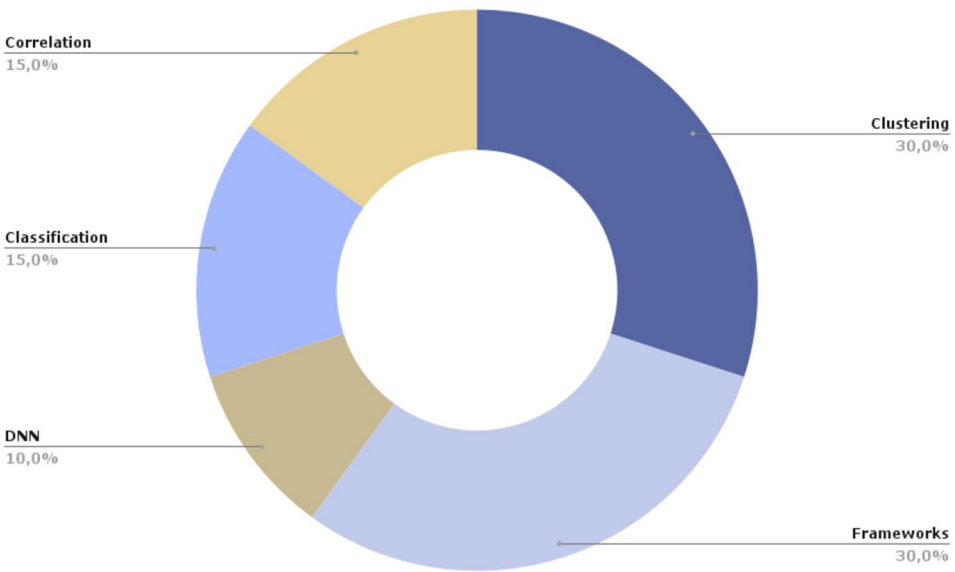

Fig. 4 Methods of urban data analysis

implied as a relationship between a predictor variable and a target variable.

All reviewed papers have different research goals and the given categorization is conditional. Our final objective is to identify trends.

The chapter is organized according to the studied techniques, each subsection is a discussion of the reviewed papers in the category.

Clustering

Clustering is a partition of the data set into groups based on similar features. It is used in a variety of data processing tasks, including pattern recognition, machine learning, automatic classification, development of control strategies, and others.

So far, no universal algorithm has been found that is effective on various kinds of data. Mainly iterative methods of clustering are used, which are based on a priori setting the number of clusters and some choice of initial partitioning. In this case, the result of their application depends significantly on the correctness of the estimation of the number of clusters.

Clustering is one of the most popular approaches in applications for a considerable amount of problems with big data. It can already be considered a base step for further investigation, as it gives a clearer picture of patterns in any set of data. There are 6 papers with different clustering algorithms utilized for diverse types of input samples.

The community detection algorithm Girvan–Newman GN [28] was modified for big data clustering of IoT sensors by [15]. Their method organizes complex data in blocks, called communities or modules, according to certain roles and functions, organized in a multi-graph. The problem is to find in a given multi-graph a partition of vertices where the objective function is minimized. To achieve this, the graph edges are deleted iterative, depending on the value of the metric. The Edge Betweenness Centrality (EBC) is the most common metric used, but the computation for this is time-consuming. The authors suggested a new measure approximating EBC, which capitalizes on hyperbolic network embedding and can be considered as the "hyperbolic" analog of EBC. This measure is denoted as Hyperbolic Edge Betweenness Centrality (HEBC), and it is computed by utilizing the hyperbolic node coordinates assigned to the embedded nodes. The novel metric enhances the performance without harming accuracy.

The other data organizing and processing technique proposed by [16] implies 3D data analytics using a dendrogram (hierarchical) clustering approach. 3D data represents a structure of information that combines, simultaneously, the classification and clustering tasks. The organized data is mapped to a tree structure and retrieved by tree traversal algorithms. Dendrogram clustering is a method of merging objects into bunches. In the study, the bottom-up algorithm of clustering is utilized, which means that each item in a class is assigned to a single cluster, then combine the groups until all objects are merged together. An important parameter is a distance between objects in a class. The metric shows a quantitative assessment of the items' similarity ratio according to different criteria. The given research does not provide a selection of the specific parameter, although the choice of metric occurs in the second step of the method. The ability to retrieve information and the efficiency of the structure were measured. In general, the technique demonstrates a good characteristic of information extraction but not the most attractive performance parameters.

Other clustering algorithms, Fuzzy c-Mean (FCM) and k-Means were tested by [17]. The k-Means algorithm is one of the simplest methods but at the same time the most inaccurate. The main idea is that at each iteration, the center of mass is recalculated for each cluster obtained in the previous step, then results are partitioned into clusters again under new centers. The algorithm ends when the cluster is not changed in iteration. The fuzzy c-Mean method allows for obtaining "fuzzy" clustering of large sets of numerical data and makes it possible to correctly identify objects at the boundaries of clusters. However, the execution of this algorithm requires serious computational resources and the initial setting of the number of clusters. In addition, ambiguity may arise with objects remote from the centers of all clusters.

A new approach for clustering data points was designed by [18]. In essence, the method is an extension of the K-set clustering algorithm for semi-metric space. The problem with the K-sets approach is that the triangle distance is not non-negative. Thus, the K-sets algorithm may not converge at all and there is no guarantee that the output of the K-sets algorithm is clustering. For solving this difficulty, the definition of triangle distance was adjusted, so that the non-negativity requirement could be lifted. The experimental results confirm the proficiency of the method for the geographic distance matrix and the latency matrix.

One more clustering method is used by [21]. Fuzzy c-Means (FCM), k-means (KM), and Gustafson–Kessel (GK) clustering algorithms are implemented. According to the paper, the most accurate and effective algorithm is k-Means clustering, but the other methods have their own advantages and show higher correctness in certain cases.

Geospatial big data were clustered with a "spectral clustering algorithm" [22]. In the research, the authors presented a newly developed method with adaptive graphs to constrain multi-view subspace clustering of multi-source geospatial big data, which are taxi GPS trajectories, points of interest, and bus smart card transactions. The data were gathered and processed according to traffic analysis zones. Autoencoder networks were used to map high-dimensional and noisy original geospatial big data into a latent representation. The latent representation of each type of geospatial big data was used to construct the low-dimensional subspace. A shared nearest

neighbor method was applied to construct adaptive graphs for high-dimensional, non-uniform, and noisy geospatial big data. Finally, a spectral clustering algorithm was used to obtain the clustering results with the similarity matrix. The clusters reveal urban function zones in a city in line with human activities. The accuracy of the clustering was 68,11% [22]. The proposed method is interesting from the socio-economical point and can be extrapolated for the bigger regions and consumption statistics can be utilized instead of the geo data. Nonetheless, the accuracy is still considerably low and the method requires further refinement. The complexity evaluation was not presented.

The K-sets+ algorithm yields the highest performance of all cluster algorithms. The time complexity is linear $O((Kn+m)I)$, where I is the number of iterations. The other method with linear time is Fuzzy c-mean with $O(nCI)$, where C - number of clusters, I - number of iterations. If we compare the exponent for these two approaches, the apparent fact is that the K-sets+ gives a little advantage. Considering the accuracy, K-sets+ has 95% as the worst result. The Fuzzy c-Means algorithm gives the complexity on average 81,97%. It is noteworthy that the Girwan–Newman modification provides 100% accuracy for most datasets and only 50% in the case of outliers. It could be used for a dataset with low sparseness if high accuracy is required. The dendrogram clustering method is slower than the others but can produce a hierarchical tree structure for data. K-means clustering is the simplest method but has a quadratic complexity and an accuracy of not more than 88% for different input data.

One of the contributions, which is to use a clustering approach to group the consumption load curves of water consumption was presented by [25]. The clustering aims to categorize daily water usage patterns according to intensity, such as normal/abnormal, days/weekends, and vacations. The hierarchical agglomerative clustering method was used, which complexity is quadratic. Three different approaches of clustering were considered: Euclidean distance, Dynamic Time Warping, and using Self-Organizing Map. The clustering method with Euclidean distance was the most efficient from the view of accuracy and complexity. The given method is not the most effective in all clustering methods, but worthy of consideration in consumption data categorization.

In general, clustering methods are effective and allow identifying masses of regularities helpful for analysis. However, it still strongly depends on the type of data, on pre-processing, and on a case-by-case basis in addition to medium accuracy and resource-intensive. In short, the analysis can be terminated with clustering, it is positive with supplemental techniques.

Frameworks

Big data processing is a challenging task and it is often convenient to create a chain or pipeline to automate the analysis. A framework is usually developed for this purpose, consisting in a pipeline for big data processing. There are 6 articles that presented different framework implementations.

A visual analysis framework for exploring and understanding heterogeneous urban data was presented by Chen et al. [20]. A visually assisted query model (Visual Analysis Approach for Exploring Spatio-Temporal Urban Data - VAUD) is introduced as a foundation for interactive exploration coupled with simple, yet powerful, structural abstractions and reasoning functionalities. VAUD presents the visualization of a variety of urban data, composed of human mobility information, mobile phone calls, traffic, and many others. The approach discussed by the authors is based on queries to the database. In this sense, time complexity could not be estimated. Data is thus retrieved, related and visualized through a custom made query language. The accuracy on average for queries is 76%.

The approach followed by Trilles et al. [10] includes three layers: content layer, services layer, and application layer. The content layer includes sensor network data sources, and the services layer provides database connection, transformations of data, and communications protocols for real-time data handling and processing. The application layer comprises the user application. The service layer implements the Cumulative SUM (CUSUM) algorithm of anomaly detection. The method considers the set of observations following a normal distribution. For each collection of measurements, the cumulative sum is calculated. When the score overcomes the threshold, the algorithm detects anomalies. If the parameter exceeds the threshold, the anomaly will be due to the increase (up-event), and if the sum is greater than the threshold, it will be due to the decrease (down-event). Different data types from multiple sources are processed by a special wrapper and transformed into standard form. Transformed observation is encoded in line according to Open Geospatial Consortium (OGC) standard for Observations and Measurements.

The patterns in electricity consumption were searched by Pérez-Chacón et al. [14]. The methodology describes all stages of data processing: data collection, cleaning, transformation, index analysis, clustering, and results. The first stage aims to pre-process the data so that they can be clustered. The second phase consists of obtaining the optimal number of clusters for the dataset by analyzing and interpreting various cluster validation indices. Next, k-means is used for clustering and, finally, retrieves the centroids for each cluster. The processing is done in Apache Spark and the algorithms include big data clustering validity indices (BD-CVIs) and k-means.

The multi-agent approach was presented by Sassite et al. [23]. The proposed architecture is composed of three layers: data acquisition and storage, data management and processing, and the service layer, based on a multi-agent system to automate big data analytics. The system is composed of multiple agents, with different roles:

- Receiver agent: handles data of multiple types and from different sources, such as sensors, external databases, web services, and others
- Storage Agent: this agent is in charge of all operations related to storage management and data reading from the Hadoop cluster
- Service Agent: interacts with applications and services and provides the processing results
- Offline Analysis Agent: handles the requests for data analysis using voluminous historical data
- Stream Analysis Agent: processes data streams in real time. As was mentioned, the input data are heterogeneous from different sources.

The authors use a classification algorithm for the data processing. The Random Forest, Naive Bayes, and Multilayer perceptron were compared. The highest accuracy was shown by Random Forest - 85%, Naive Bayes demonstrated the lowest accuracy - 61% and 73% was the accuracy of the Multilayer perceptron. The extracted group was used for real-time prediction. The proposed approach is utilized in urban management and planning.

Deep Neural Networks

Neural networks are mathematical models and their software implementation is inspired in the structure of the human nervous system. The main feature of neural networks is the ability to learn from a potentially large set of examples. Recently, neural networks are used almost in all scientific areas, of which the smart city is no exception. There are only two papers that reviewed neural networks, but 6 methods were examined and evaluated.

The Long Short-Term Memory (LSTM) neural network models, Auto-regressive Integrated Moving Average (ARIMA) models, Generalized Regression Neural Network (GRNN) models, and Exponential Smoothing (ES) models to estimate HIV incidence in Guangxi, China, and explore which model is the best and most precise for local HIV incidence prediction were used by Wang et al. [13]. ARIMA is the model used for time series forecasting. LSTM is a recurrent neural network, characterized by the ability to learn long-term dependencies. In this study, several models were built. The model with the lowest mean square error (MSE) was considered the optimal model. GRNN is a feed-forward neural network, which estimates values for continuous dependent variables. The principal advantages of GRNN are fast learning and convergence to the optimal regression surface as the number of samples becomes very large. GRNN is particularly advantageous with sparse data in a real-time environment because the regression surface is instantly defined everywhere, even with just one sample. The method is usually used for functions' approximation, so it can provide very high accuracy, but for huge samples is computationally expensive. ES model is one of the simplest and most widespread practices of series alignment. The method can be presented as a filter that receives the original series members as the input, and the output forms the current values of the exponential average.

Classification

The classification problem is characterized by the fact that the set of admissible answers is finite, containing the class labels. A class is a set of all the objects that share the same label. Classification is used to determine which category a sample of data belongs to. There are 3 papers that explore classification methods.

A general approach for the transformation of the original problem into a traditional point-valued classification problem with a sampling-based method was proposed by Huang et al. [24]. The "general approach" means that the technique can be used for any distribution-based dataset. For the testing purpose, real data about hotels reservation were used. The main purpose of the classification is the recommendation system. The authors proposed their own system of accuracy estimation based on the Top-K hit rate. It means that the system recommends K products to a user once, and if the user's final choice is in the K recommended products, we say it is an accurate hit. The rate is rising steadily with an increase in the number of recommendations [24]. The time complexity is $O(TR)$ where T is the number of training tuples each containing R attributes. With the rough estimate, we can say that the complexity is quadratic ($O(n^2)$).

The real-time classification of water consumption patterns and electrical devices was presented by Bashir et al. [25]. This involves training machine learning algorithms to identify normal and abnormal water consumption patterns and differentiate between different types of electrical devices. The signatures of electrical devices are classified using three ML algorithms: multi-layer perceptron (MLP), k nearest neighbor (KNN), and decision tree (DT). The input dataset sample is an occurrence of a unit of consumption, such as one liter of water or one watt-hour (Wh) of electricity, and each event is timestamped. From the experiments conducted, the authors concluded that One Against All (OAA) - MLP and Error-Correcting Output Codes (ECOC)-MLP P perform better than MLPs, DT, and KNN. Also, it can be highlighted that OAA-MLP outperformed all other

approaches in terms of performance and accuracy, but it is more complex than the ECOC-MLP.

The last classification approach considers water consumption. The platform proposed by Charalampous et al. [26] incorporates advanced signal processing methodologies combined with supervised machine learning classifiers to classify water flows, thus identifying residential water appliances with high accuracy. Models achieve an accuracy of 91% in classifying the four most used household water appliances. The input data set is the water flow with timestamps ranged by area and appliances. For example, offices, houses, and factories were considered as well as different appliances: kitchen faucets, showers, and others. The suggested system is aimed to expand water consumption monitoring and to rise resource savings.

Correlation

Another widespread statistical method applied to big data is correlation. In the listed papers, one algorithm considers the correlation applied to smart city data. The study compared two types of methods: Pearson correlation and Mutual information. The time complexity for both is a cube, but Pearson correlation can discover the linear distribution of data, and mutual information can find dependencies in more general data distribution cases. However, if an application prioritizes real-time response over accuracy, Pearson correlation will be suitable as it will only give a few false negatives. In other scenarios with different types of data streams (temperature, pollution, etc.), it is better to use mutual information without *a priori* knowledge of the potential correlations because we do not know the percentage of cases where Pearson correlation will fail to detect the correlations [11].

The method used by Bermudez et al. [11] tried to apply correlation methods to urban data analysis. They suggested an efficient method to derive spatio-temporal analysis of the data, using correlations, with Pearson and Entropy-based methods and compares the results of both algorithms. Pearson's coefficient characterizes the presence of linear dependence between two values. The weakness of Pearson correlation is poor accuracy when variables are not distributed normally. Mutual information is the statistical function of two random variables, which describes the quantity of information of one random value in another. The constraint of mutual information is that it has a higher processing complexity than Pearson correlation. The technique continuously calculates the average correlation for sensory road data divided into two sectors until the data runs out. Different types of correlation were tested.

Research about the correlation between energy consumption and economic growth was presented by Cao et al. [27]. To analyze the correlation between economic development and energy consumption structure deeply, authors select the main resource to get a grey relevancy correlation. The analysis was conducted in the Chinese province of Henan. There are the parameters of input data: Henan's GDP, units of billion; coal consumption, units of tons of standard coal; Crude Oil consumption, unit of tons of standard coal; Natural Gas consumption, a unit of tons of standard coal; Hydro-power consumption, a unit of tons of standard coal. As the method, the Grey relevancy analysis model was applied. The examination basic is the distances between reference points and compares points. By analyzing the differences and the proximities of factors, we could quantify uncertainties. The result correlation between energy consumption and economic efficiency is 72%, which can be considered meaningful. The authors did not provide accuracy and complexity inspection.

Processing Outcomes and Purpose

However, remote sensing techniques are hard to capture the socioeconomic attributes and human dynamics that are highly related to urban land use [22]

The assessment of time and accuracy of all proposed algorithms demonstrates that if the purpose is prediction, the best variant is deep neural networks, like LSTM. For effective clustering, the K-sets+ or Fuzzy c-Means algorithms are the most powerful. If it is necessary to obtain additional analysis, it is possible to find the correlation. Considering the context of municipal data the frameworks are beneficial, as they assume all stages of data processing from storage to visualization.

The CUSUM algorithm is time linear complexity. The solution is straightforward and fast but has limitations that must be taken into account, such as the consideration that all the series must follow a normal distribution and a series of observations cannot have trends [10].

The algorithms described above have characteristics of performance and scalability that should be understood. Table 4 gives a comprehensive description of the complexity and accuracy of the considered algorithms. For many, it was not easy to evaluate the complexity since the time depends on the characteristics of the machine. Therefore, we provide only rough estimates, and all presented assessments are for worst-case values.

The anomaly detection by the CUSUM algorithm [10] creates the warning message for the client side in the case of rare events. Each event contains a sensor identifier (sender field) and the identifier of the particular observation that has caused the event (identifier field). An event dashboard visualizes this data. The panel shows all sensing nodes of a network on a map using markers. Inside each marker, the amount of events that have been detected for this particular sensing node appears. If this node triggers an event, the marker turns red, if not the marker remains blue.

Table 4 Algorithms assessment

Algorithm CUSUM algorithm	Purpose anomaly detection	Complexity O(n)	Accuracy —	Ref [10]
Mutual information and Pearson correlation	Find correlation between sensory data	$O(n^3)$	+	[11]
LSTM	Predict diseases	$O(w)$	85%+	[13]
ARIMA	Predict diseases	–	80%	[13]
GRNN	Predict diseases	–	76%	[13]
ES	Predict diseases	–	74%	[13]
K-means clustering	To extract electric energy consumption patterns	$O(n^2)$	78%	[14]
modification of Girvan-Newman algorithm	Community detection	$O(n^2)$	65–100%	[15]
Dendrogram clustering	Produce hierarchical tree structure for data retrieval and analytics	$O(n^3)$	–	[16]
K-Means	Clustering	$O(n^2)$	87,94%	[14, 17]
Fuzzy c-Mean	Clustering	$O(nCI)$	81,91%	[17]
K-sets+	Clustering in metric space	$O(Kn + m)$	95%	[18]
DNN	Predicting infectious diseases	$O(wnk)$	77%	[19]
VAUD	Spatio-temporal data visualization	–	78.6%+	[20]
K-Means	Clustering		60.41–87.94%	[21]
Fuzzy c-Means (FCM)	Clustering		56.25–81,91%	[21]
Gustafson-Kessel (GK)	Clustering		66.19–95.83%	[21]
Similarity-Matrix-based Clustering	Trip clustering	$O(n^3)$	–	[12]
Multi-View Subspace Clustering	Clustering		68.11%	[22]
Naive Bayes	Classification		61%	[23]
Multilayer perceptron	Classification		73%	[23]
Random forest	Classification		85%	[23]
Classification of distribution-based data	Classification	$O(n^2)$	90% - best case, 44% - worst-case	[24]
Classifying Electrical Devices	Classification	400 s	99,93%	[25]
Water Consumption Classification	Classification		91%	[26]
Hierarchical Agglomerative Clustering With Euclidean Distance	Clustering	$O(n^2)$	72%	[25]

The analyses based on the correlation and mutual information were used to monitor the traffic of the city. Three sets of experiments have been performed. In the first one, the performance of Pearson correlation and mutual information was compared [11]. The results were visualized on Google Maps. It can be concluded that the Pearson correlation is effective for the linear distribution of data, and mutual information is vital for nonlinear dependencies but requires more time.

The results obtained by Wang et al. [13] are predictions of HIV disease for two years. Each compared algorithm has its metrics. For example, ARIMA includes a moving average process, an auto-regressive moving average process, an auto-regressive moving average process, and an ARIMA process according to the different parts of the regression and whether the original data are stable. To evaluate data accuracy, they compared with original information about HIV cases for 2015 and 2016 years. The same type of outcomes data demonstrates the [19]. They compared the same parameters for LSTM, DNN, and ARIMA to evaluate infectious disease prediction correctness. All cluster algorithms give the same result as a count of clusters and their accuracy.

The electricity consumption data were clustered into 4 and then into 8 groups in [14]. The outcomes are presented as diagrams. The clusters are categorized depending on buildings, seasons of the year, and days of the week.

The modification of the Girvan–Newman method with a novel metric provided by [15] was applied to multidimensional data obtained from an operational smart-city/building IoT infrastructure. The authors presented an accuracy evaluation, modularity, and time comparison of HGN and GN, comparing the execution time of GN and HGN algorithms for graphs with known communities and modularity comparison for 5-, 10-, 20-, 30-, and 60-minute sampling. Given that statistics demonstrate the computational efficiency and that algorithm can give accurate outcomes.

The cluster visualization into dendrograms, as tested on the information about 1000000 buildings was presented by Azri et al. [16]. Response time analysis was provided as

well, which exhibits that response time for the proposed method is 50–60% faster than non-constellated data.

Conclusion

The aim of the article was to figure out what is the trend in the city's infrastructure data processing. The authors were interested in consumption data of electricity, water, heat, data on city traffic, and the methods for creating predicting models, clustering, and classifying. Increasingly, big data are seen as a key resource for the development of the urban environment, which presents opportunities for the optimization of economic processes, the creation of innovations in the social sphere, formation of new management models. The literature review serves as a foundation for future work in resource expenditures data analysis and urban management system creation.

The article presents a detailed analysis of the twelve papers from the last five years. The authors considered the techniques of urban data processing. The input and output data, assessments of algorithms' effectiveness, and methods description are provided. The inspection gives the following results: 1 algorithm of correlation, 2 algorithms of classification, 1 method of anomaly detection, 2 approaches for data visualization, 5 algorithms of predicting, and 6 methods for data clustering. A disproportion between the number of reviewed articles and the number of techniques dues to the fact that more than one method in each research was provided.

The input and the output data vary depending on the method and purposes of the research. Predominantly, heterogeneous data sources are considered. In one case, the images are exploited, and in two cases, the numerical information is leveraged. The heterogeneous data mean that the information of different bases is used: images, text, and numerical. The video or sound data is not used in reviewed papers. The major part of the investigation offers a clustering model in the capacity of output results. The second place in prevalence is frameworks. The remaining outcomes can be divided between deep neural networks, classification, and correlation models.

The most interesting approach is the leverage of LSTM. Based on surveyed articles, LSTM gives the highest accuracy of prediction and is the fastest solution in comparison with similar solutions. The forecasting of social phenomena based on city data is the most desirable result. Although, in the context of the modern situation is a still challenging task, as the whole pipeline of the assembly, processing, and analysis is important. As the given review demonstrates, different data are necessary for various problems, different algorithms give diverse findings. The dilemma of the practical benefits and standardization in smart city data is still open.

Acknowledgements This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/AI/0088/2020.

Funding Open access funding provided by FCT|FCCN (b-on).

Data availability Data sets generated during the current study are available from the corresponding author on reasonable request. The water consumption data are available from Bragança's municipality but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

References

1. Kwon O, Kim YS, Lee N, Jung Y. When collective knowledge meets crowd knowledge in a smart city: a prediction method combining open data keyword analysis and case-based reasoning. J Healthc Eng 2018;2018. https://doi.org/10.1155/2018/7391793
2. Dong XL, Srivastava D. Big data integration. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), 1245–1248. 2013. https://doi.org/10.1109/ICDE.2013.6544914
3. Naeem M, Jamal T, Diaz-Martinez J, Butt SA, Montesano N, Tariq MI, De-la-Hoz-Franco E, De-La-Hoz-Valdiris E. Trends and future perspective challenges in big data. In: Pan J-S, Balas VE, Chen C-M, editors. Advances in intelligent data analysis and applications. Singapore: Springer; 2022. p. 309–25.
4. Materla T, Cudney EA, Antony J. The application of Kano model in the healthcare industry: a systematic literature review. Total Qual Manag Bus Excell. 2017;1–22. https://doi.org/10.1080/14783363.2017.1328980. (**Accessed 2019-01-06**).
5. Subhash S, Cudney EA. Gamified learning in higher education: a systematic review of the literature. Comput Hum Behav. 2018;87:192–206. https://doi.org/10.1016/j.chb.2018.05.028. (**Accessed 2019-01-06**).
6. Chehri A, Fofana I, Yang X. Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence. Sustainability. 2021;13(6):3196. https://doi.org/10.3390/su13063196. (**Accessed 2023-06-24**).
7. Wang X, Zou Z. Open data based urban for-profit music venues spatial layout pattern discovery. Sustainability. 2021;13(11):6226. https://doi.org/10.3390/su13116226. (**Accessed 2023-06-24**).
8. Chang J, Nimer Kadry S, Krishnamoorthy S. Review and synthesis of big data analytics and computing for smart sustainable cities. IET Intell Transp Syst. 2020;14(11):1363–70. https://doi.org/10.1049/iet-its.2020.0006. (**Accessed 2023-06-24**).
9. Gutman AJ, Goldmeier J. Becoming a data head: how to think, speak, and understand data science and machine learning, indianapolis: statistics. 2021.
10. Trilles S, Belmonte O, Schade S, Huerta J. A domain-independent methodology to analyze IoT data streams in real-time. A proof of

concept implementation for anomaly detection from environmental data. Int J Digit Earth. 2017;10(1):103–20. https://doi.org/10.1080/17538947.2016.1209583.
11. Bermudez-Edo M, Barnaghi P, Moessner K. Analysing real world data streams with spatio-temporal correlations: entropy vs. pearson correlation. Autom Construct. 2018;88:87–100. https://doi.org/10.1016/j.autcon.2017.12.036.
12. Bordogna G, Cuzzocrea A, Frigerio L, Psaila G. An effective and efficient similarity-matrix-based algorithm for clustering big mobile social data. 2017;514–521. https://doi.org/10.1109/ICMLA.2016.188.
13. Wang G, Wei W, Jiang J, Ning C, Chen H, Huang J, Liang B, Zang N, Liao Y, Chen R, Lai J, Zhou O, Han J, Liang H, Ye L. Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. Epidemiol Infect. 2019;147:194. https://doi.org/10.1017/S095026881900075X. (**Accessed 2020-05-27**).
14. Pérez-Chacón R, Luna-Romera JM, Troncoso A, Martínez-Alvarez F, Riquelme JC. Big data analytics for discovering electricity consumption patterns in smart cities. Energies 2018;11 (3). https://doi.org/10.3390/en11030683.
15. Karyotis V, Tsitseklis K, Sotiropoulos K, Papavassiliou S. Big data clustering via community detection and hyperbolic network embedding in IoT applications. Sens (Switzerl) 2018;18(4). https://doi.org/10.3390/s18041205.
16. Azri S, Ujang U, Abdul Rahman A. Dendrogram clustering for 3D data analytics in smart city. Issue: 4/W9. 2018;42:247–53. https://doi.org/10.5194/isprs-archives-XLII-4-W9-247-2018.
17. AlShami A, Guo W, Pogrebna G. Fuzzy partition technique for clustering Big Urban dataset. In: 2016 SAI Computing Conference (SAI), 2016; 212–216. https://doi.org/10.1109/SAI.2016.7555984.
18. Chang C-S, Chang C-T, Lee D-S, Liou L-H. K-sets+: a linear-time clustering algorithm for data points with a sparse similarity measure. 2018;1–8. https://doi.org/10.1109/UIC-ATC.2017.8397636.
19. Chae S, Kwon S, Lee D. Predicting infectious disease using deep learning and big data. Int J Environ Res Public Health. 2018;15(8):1596. https://doi.org/10.3390/ijerph15081596. (**Accessed 2020-05-27**).
20. Chen W, Huang Z, Wu F, Zhu M, Guan H, Maciejewski R. VAUD: a visual analysis approach for exploring spatio-temporal urban data. IEEE Trans Visual Comput Graph. 2018;24(9):2636–48. https://doi.org/10.1109/TVCG.2017.2758362. (**Conference Name: IEEE Transactions on Visualization and Computer Graphics**).
21. Simhachalam B, Ganesan G. Performance comparison of fuzzy and non-fuzzy classification methods. Egyp Inform J. 2016;17(2):183–8. https://doi.org/10.1016/j.eij.2015.10.004. (**Accessed 2021-05-14**).
22. Liu Q, Huan W, Deng M. A method with adaptive graphs to constrain multi-view subspace clustering of geospatial big data from multiple sources. Remote Sens. 2022;14(17):4394. https://doi.org/10.3390/rs14174394. (**Accessed 2023-06-24**).
23. Sassite F, Addou M, Barramou F. A machine learning and multi-agent model to automate big data analytics in smart cities. Int J Adv Comput Sci Appl. 2022;13(7):441–51. https://doi.org/10.14569/IJACSA.2022.0130754.
24. Huang J, Zhu L, Liang Q, Fan B, Li S. Efficient classification of distribution-based data for internet of things. IEEE Access. 2018;6:69279–87. https://doi.org/10.1109/ACCESS.2018.2879652. (**Conference Name: IEEE Access**).
25. Bashir S. Real-time water and electricity consumption monitoring using machine learning techniques. IEEE Access. 2023;11:11511–28. https://doi.org/10.1109/ACCESS.2023.3241489. (**Conference Name: IEEE Access**).
26. Charalampous A, Papadopoulos A, Hadjiyiannis S, Philimis P. Towards hydro-informatics modernization with real-time water consumption classification. In: 2021 IEEE AFRICON, 1– 6 2021. ISSN: 2153-0033. https://doi.org/10.1109/AFRICON51333.2021.9570909.
27. Cao W, Zhang H, Li J. A Grey relevancy analysis on the relationship between energy consumption and economic growth in Henan Province. In: 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC). 27– 30. 2011. https://doi.org/10.1109/AIMSEC.2011.6010196
28. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99(12):7821–6. https://doi.org/10.1073/pnas.122653799. (**Accessed 2021-11-28**).