Machine learning for spatial analyses in urban areas
a scoping review

Casali, Ylenia; Aydin, Nazli Yonca; Comes, Tina

DOI
10.1016/j.scs.2022.104050

Publication date
2022

Document Version
Final published version

Published in
Sustainable Cities and Society

Citation (APA)
Casali, Y., Aydin, N. Y., & Comes, T. (2022). Machine learning for spatial analyses in urban areas: a scoping review. *Sustainable Cities and Society*, *85*, Article 104050. https://doi.org/10.1016/j.scs.2022.104050

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Machine learning for spatial analyses in urban areas: a scoping review

Ylenia Casali [*], Nazli Yonca Aydin, Tina Comes

TU Delft, Faculty of Technology, Policy and Management, Building 31, Jaffalaan 5, 2628BX, Delft, The Netherlands

ARTICLE INFO

Keywords:
Machine learning
Urban areas
Review
Spatial analyses
Geospatial data

ABSTRACT

The challenges for sustainable cities to protect the environment, ensure economic growth, and maintain social justice have been widely recognized. Along with the digitization, availability of large datasets, Machine Learning (ML) and Artificial Intelligence (AI) are promising to revolutionize the way we analyze and plan urban areas, opening new opportunities for the sustainable city agenda. Especially urban spatial planning problems can benefit from ML approaches, leading to an increasing number of ML publications across different domains. What is missing is an overview of the most prominent domains in spatial urban ML along with a mapping of specific applied approaches. This paper aims to address this gap and guide researchers in the field of urban science and spatial data analysis to the most used methods and unexplored research gaps. We present a scoping review of ML studies that used geospatial data to analyze urban areas. Our review focuses on revealing the most prominent topics, data sources, ML methods and approaches to parameter selection. Furthermore, we determine the most prominent patterns and challenges in the use of ML. Through our analysis, we identify knowledge gaps in ML methods for spatial data science and data specifications to guide future research.

1. Introduction

Cities are facing tremendous environmental, infrastructural and social challenges that are unprecedented in scale, scope, and complexity (Meerow & Newell, 2019). To become sustainable, cities need to accommodate a growing population, meet greenhouse gas targets, adapt to a changing climate, and ensure fair and equal living conditions for all. To address these challenges and to improve urban efficiency, justice and quality of life, sustainable smart cities use information and communication technology (Colding et al., 2020). The associated rise of sensors, crowd sourcing and real-time monitoring has tremendously increased the availability of large spatial datasets. Advances in urban geographic information sciences and spatial data analytics have opened new avenues to analyze and visualize spatial data (Goodchild & Haining, 2004). Leveraging these advances and benefiting from the increasing of digital innovations of our cities has been identified as one of the key transformations needed for achieving the Sustainable Development Goals (Sachs et al., 2019).

Today, most prominently Artificial Intelligence (AI) and Machine Learning (ML) provide new opportunities to better monitor, understand, and predict the (sustainable) development of urban areas. As such, urban analytics and modeling have become increasingly prominent to deal with the complex sustainability challenges that cities grapple with (Batty, 2008). Studies such as Nosratabadi et al., and Aram (2020)) and Vinuesa et al. (2020) have used machine learning to improve sustainability and achieve the sustainable development goals. Here, we follow the vision of Elmqvist et al. (2019) in defining a sustainable city via the "integration of all sub-systems in an urban region in ways that guarantee the wellbeing of current and future generations'' (Elmqvist et al., 2019). As such, we review subsystems that relate to the social, economic and environmental aspects of sustainable cities, as well as the infrastructural systems that shape the interactions between the different elements (see also Section 3.2 for an overview of categories).

Machine learning (ML) has gained popularity in many research fields. The foundations of ML were already laid in 1959, when Arthur Samuel, a pioneer in AI, coined the term (Samuel, 1959). In a nutshell, ML is a method to train algorithms to understand patterns inherent in data and predict outcomes based on statistical analysis. ML methods are data-driven: they extract meaningful information from data, instead of a priori modeling causal links. The 'learning' aspect herein implies that the better an algorithm performs in a specific task, the better it learned from that experience (Mitchell, 1997).

ML algorithms are divided into two main groups: supervised and unsupervised learning. Supervised learning uses a training set of examples with correct responses (targets) (Hastie et al., 2009; Marsland, 2014) In contrast, in unsupervised learning, correct responses are not

[*] Corresponding author.
E-mail address: ylenia.casali@gmail.com (Y. Casali).

https://doi.org/10.1016/j.scs.2022.104050
Received 26 April 2022; Received in revised form 6 July 2022; Accepted 6 July 2022
Available online 12 July 2022

provided. Instead, the algorithms aim to identify similarities between inputs and group them (Celebi & Aydin, 2016). Moreover, natural language processing (NLP) developed techniques that aim to extract a fuller meaning representation from free text (Kao & Poteet, 2007). Studies can combine algorithms from supervised, unsupervised and NLP methods.

In the 1990s, Openshaw and Openshaw (1997) published one of the first books about ML applications in geography. Since then, ML has contributed to the fields of geography and spatial analyzes generally, and urban systems more specifically. Spatial ML uses primarily geospatial data, which refers to data containing a geographic component that identifies locations (e.g., coordinates, addresses, and postcodes) or indicates geographically referenced features and conditions, such as the population of a district, seasonal weather of a region, number of vehicles passing a highway intersection, and geo-tagged social media data (Boulos et al., 2019). Moreover, urban spatial ML analyzes different aspects of the urban system, consisting of multiple tangible (e.g., infrastructures, land use) and intangible aspects (e.g., social equality, gentrification).

Recently, GeoAI was proposed as a framework for analyzing data-driven problems in geographic information science (Janowicz et al., 2020; Li, 2020). GeoAI aims to integrate artificial intelligence, in particular deep learning techniques, with geospatial big data and high-performance computing to investigate geospatial problems. In GeoAI, spatially explicit models are viewed as a significant research direction. Those models fulfil at least one of these four requirements: the results are not invariant under the relocation of studied phenomena (invariance test), the models contain a spatial representation of the studied phenomena (representation test), the models make use of spatial concepts in their implementation (formulation test) and the spatial forms of input and outcomes differ (outcome test) (Goodchild, 2001; Janowicz et al., 2020). Clear steps to build spatially explicit models shifting from general ML models to designing more complex ones are not yet well-evaluated.

While spatial data collection has been accelerated through technological innovations (such as social and remote sensing), the availability of the data is not equally distributed throughout the world (Guigoz et al., 2017; Leyk et al., 2019). At locations where data is available, local statistical data are related to different areas of a municipality, which can vary among organizations and time. Because of the heterogeneous nature of data sources and availability, spatial analyses need to integrate data from different sources and spatial granularities to establish a comprehensive understanding (Cheng et al., 2006). Due to the intense data collection and processing requirements, the reuse of spatial data has become a new norm (Janowicz et al., 2020). Lack of standards and unclear data collection procedures become a potential risk in the development of reliable datasets.

Table 1
Previously published reviews on machine learning applications for geography and urban analysis.

Authors (year)	Field of study
Biljecki and Ito (2021)	Street view imagery
Chaturvedi and de Vries (2021)	Urban land use planning
Grekousis (2019)	ANN and deep learning in urban geography
Hegde and Rokseth (2020)	Engineering risk assessment
Ibrahim et al. (2020)	Computer vision
Lary et al. (2015))	Remote sensing
Ma et al. (2019)	Remote sensing
Maxwell et al. (2018)	Remote sensing
Milojevic-Dupont and Creutzig (2021)	Climate change mitigation
Nikparvar and Thill (2021)	Spatial data
Toch et al., and Ben-Gal (2019)	Mobility data
Zhu et al. (2017)	Remote sensing

Reflecting the increasing popularity of ML methods, several reviews were published in the fields of geography and urban analysis (see Table 1[1]). While there are publications that focus on specific areas of application or ML algorithms, there is no comprehensive overview across urban domains that allows researchers to compare and choose the most adequate methods for their topic, neither it is possible to understand the potential overlaps and synergies, or to leverage the insights from one field for another. Moreover, a discussion about the types of spatial data used for urban ML analyzes, or methods for choosing parameters is missing. We address this gap by conducting a scoping review of the fields and domains in urban analysis, which have priority in ML research, along with a mapping of the specific approaches, algorithms, or data sets and their fit to specific applications.

As indicated in Table 1, there are already numerous reviews on remote sensing (for example see Lary et al., 2015; Ma et al., 2019; Maxwell et al., 2018; Zhu, Tuia et al., 2017). These reviews show that support vector machines (SVM), random forests (RF), and boosted decision trees (DTs) have been shown to be very powerful methods for classification of remotely sensed data. However, all remote sensing studies aim to detect and monitor the physical surface of the world by using remotely sensed images. What is missing though, is the relation of the physical features of a city to its functions and sustainability. Therefore, in our scoping review, we focus on studies that primarily use geospatial data for urban sustainability. We explain the eligibility criteria in depth in section 2.1. The remainder of this paper is organized as follows. Section 2 explains the material and methods used for this scoping review. Subsequently, the paper provides insights into (i) the main themes and domains of applications of ML in urban analytics (Section 3.2), (ii) the data sources used (Section 3.3), (iii) the ML algorithms applied (Section 3.4) and (iv) the approaches for parameter selection (Section 3.5). The paper continues with a discussion of the main gaps and presents a research agenda to address these gaps. We conclude with the main findings.

2. Material and methods

This section describes the process and methods that have been followed in this review. As our objective here is to scope the field and its many applications for sustainable cities, we opted for a scoping review. Scoping reviews have been developed as a methodology to develop a mapping of study domains, data sources, approaches, and methods (Peters et al., 2015). While scoping reviews are still relatively new as compared to systematic reviews, they have been described as an ideal tool to determine the scope or coverage of an (emergent) body of literature on a given topic and provide an overview of its focus (Munn et al., 2018). A scoping review is especially suitable because the number of publications on ML applications for urban analyses has grown rapidly in the past years. Therefore, it is impossible to conduct a rigorous systematic review without excluding aspects of the field. Moreover, systematic reviews are not immune to exclusions of relevant papers (Biljecki & Ito, 2021).

Methodologically, our review process falls into the conventional three steps of a scoping review (Peters et al., 2015): (i) planning the review by developing eligibility criteria; (ii) identifying relevant literature through a database search, screening and selection; (iii) conducting the review and charting the results.

[1] In addition, Kamel Boulos, Peng, and Vopham (2019) presented works in GeoAI for healthcare topics, which might have applications for urban sustainability. However, this is not a formal review, so not included here.

2.1. *Planning the review: Eligibility Criteria*

ML and urban analysis for sustainability are broad fields. We used the following criteria to select papers that are relevant for our analysis on ML applications that use spatial information in urban areas:

1) Papers mainly used ML algorithms to solve urban problems. We included supervised, unsupervised ML and neural linguistic programming methods. We excluded papers that used solely linear regression or that discussed the theory of ML.
2) Papers primarily focused on urban scales ranging from neighborhoods to counties. The broad scoping allows us to include applications on smaller areas that could scale to cities or metropolitan regions.
3) Papers used geospatial datasets, i.e., data series, vector, or raster datasets when they are used in conjunction with a geographic location stored by coordinates or by indexes. For example, we included papers that used satellite images in combination with feature datasets as references for land use and land cover.
4) We excluded papers that solely focused on remote sensing, detection of geospatial objects and features from remote sensing images, image processing, image classification, computer vision, urban street images.
5) Papers are published in journals or peer-reviewed conferences and written in English.

2.2. *Database search and screening*

To identify an initial pool of literature for this study, the Web of Science was used to ensure the highest academic standards and validity of the articles, and for its broad and multi-disciplinary coverage. The web of Science (WoS) is the oldest, most widely used, and authoritative database of publications (Birkle et al., 2020), and in a recent comparative study has been shown to guarantee reproducible results (Gusenbauer & Haddaway, 2020).

In the literature, the terms urban areas, cities and urban environment are often used interchangeably. Therefore, we included each term in the search. Moreover, we included the keywords ‘urban spatial analysis’ and ‘land use change’ to aim for papers with a spatial analytical component. As a result, we used as keywords ‘urban area’, ‘cities’, ‘urban environment’, ‘urban spatial analysis’, ‘land use change’ and ‘machine learning’ for our database search (search string: ((‘urban area’ OR ‘urban spatial analysis’ OR ‘land use change’ OR ‘cities’ OR ‘urban environment’) AND ‘machine learning’)). We screened the literature by following three approaches to lower the risk of bias. First, we looked for papers that included the keywords in the abstract and that were highly cited according to Web of Science statistics to ensure inclusion of publications with high impact. Second, we looked for papers that were published in 2021 to ensure that the most recent trends and developments are covered. Third, we identified additional papers by snowballing with Google Scholar. In total, we screened 245 papers and selected 162 papers that met all eligibility criteria. We collected this set of papers on December 2021.

2.3. *Review and analysis*

After selecting the articles based on the eligibility criteria, we analyzed the body of literature. We selected key information in the papers: title, authors, year of publication, the purpose of the study, place of the case study, the method used, data reported, training-testing information, and hyperparameter or parameter information. For our mapping of themes and methods, we collected the information in tables by analyzing each paper. If a paper did not provide any information about a specific detail, we reported it as missing.

Our analysis covered five perspectives.

1) We investigated the **spatial and temporal distribution of papers**. For the spatial analyzes, we use the locations of case studies, and grouped them into seven regions: Africa, Asia, Europe, North America, Central and South America, Middle East, and Oceania.
2) We mapped out the **topics** studied in papers to identify priority research areas and gaps. We developed four categories of studies that represented specific urban sub-systems: land use and urban form, socioeconomic, environment and infrastructures.
3) To identify patterns of data (sources), we investigated the **type of data** that papers used to develop their models. We distinguished data stored in tabular form (e.g., csv) and spatial data in vector and raster forms as well as remote sensing data.
4) To map out the most prominent ML **methods** in each category of study, we analyzed the **methods** used. We distinguish ML methods based on supervised, unsupervised, a mix of unsupervised and supervised and natural language processing algorithms.
5) We analyzed the **training-testing** and the **hyperparameter-parameter** information reported to study how authors implemented their analyzes and reported the associated information.

3. Results

3.1. *Spatial and temporal distribution*

159 papers reported the location of the case studies. Fig. 1 shows the distribution of case studies by country and over time, clearly highlighting the discrepancy between regions. Most cases are located in China and the US, followed by the UK. Overall, 31% of the case studies were in Europe, 29% in Asia, and 27% in North America. 7 papers include multiple case studies in different countries and 4 of them on different continents. If a paper covered multiple case studies, we counted each case study separately and assigned it to the respective continent. In Fig. 1, the bars show the year of publication of papers. 84% of the papers were published between 2014 and 2021 indicating the increasing popularity of the field.

3.2. *Categories*

To derive topical categories, we built on the conceptual model of a sustainable urban system developed by Meerow and Newell (2019). As we focused on spatial attributes, we omitted the governance layer, and included those categories that characterize the urban system in terms of people, physical features, and services (i) land use and urban form, (ii) socioeconomic, (iii) infrastructures. To also capture the importance of environmental factors and hazards on the urban system, we put forward a fourth category on *environment*.

We found that 34% of the studies (55 papers) are dedicated to infrastructure, 24% to socioeconomic topics (39 papers), 23% to land use and urban form (38 papers), and 18% to environmental topics (30 papers). In the following sections, we present a summary of selected papers based on each category. We will discuss these categories and findings in section 4.

3.2.1. *Land use and urban form*

Fig. 2 shows the main topics studied in these categories. We distinguish papers dedicated to the (A) land use (29/38 papers) from studies related to the (B) urban form (9/38 papers).

A Studies on the **land use** focus on (A.1) land use detection (15 papers) and (A.2) land use change (14 papers).

A.1 **Land use detection** characterizes areas in cities spatially. We identified four categories (see Fig. 2): land use detection from social sensing, functional areas, urban identities, and informal settlements. Of these, the identification of functional areas was the most prominent topic of study (6/15 papers).

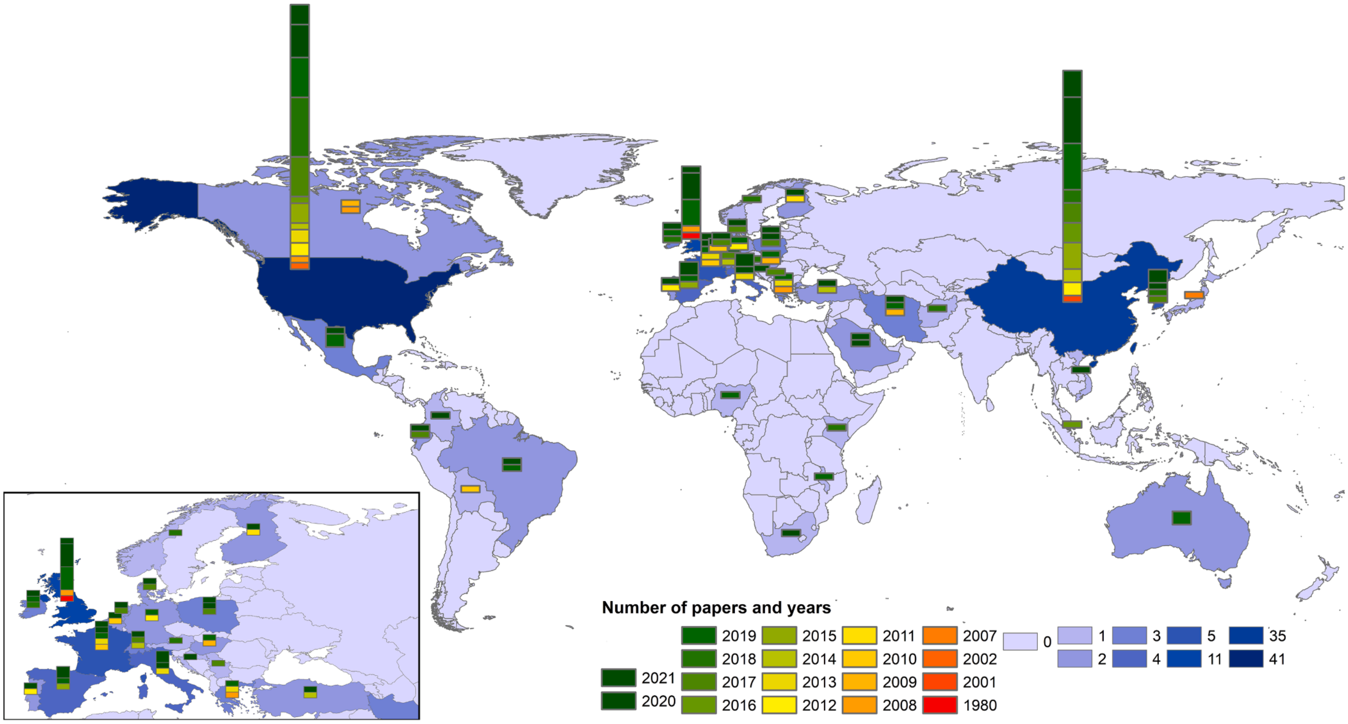

Fig. 1. Spatial and temporal distribution of papers. The map shows the distribution of case studies per countries, the bars show the years of publication.

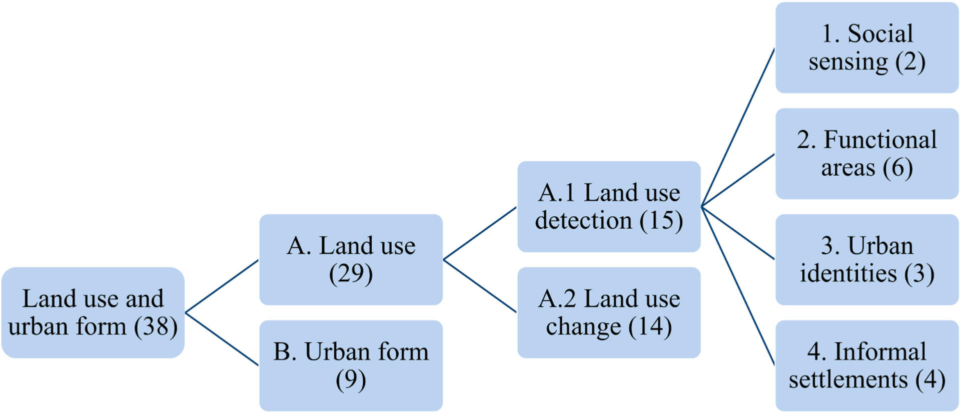

Fig. 2. Topics studied in the land use and urban form category. The tree-plot shows the main research themes. We reported the number of papers per category in brackets.

1 **Social sensing** based on mobile phone or social media data is used in two papers to map urban areas based on the activity patterns (Cranshaw et al., 2012; Toole et al., 2012).

2 **Functional areas** were identified by specific activities and mobility patterns (Hu et al., 2020; Yan et al., 2017; Yao et al., 2017; Yuan et al., 2012; Zhai et al., 2019; Zhang et al., 2018).

3 **Urban identities** were studied by exploring the physical or spatial characteristics of public spaces (Chang et al., 2017; Chang et al., 2018), or the characteristics of attributes of urban blocks (Laskari et al., 2008).

4 **Informal settlements** studies focus on mapping (Fallatah et al., 2020; Jochem et al., 2018), detecting (Mahabir et al., 2020) or understanding the growth (Badmos et al., 2019) of urban informal settlements and slums. Because these studies are conventionally situated in the Global South, the papers also present approaches to address the problem of limited geographic data.

A.2 **Land use change** is used to analyze urban evolution over time. Predicted changes can be the transition from non-urbanized to urbanized (Huang et al., 2009; Pijanowski et al., 2002, 2014) or between land use classes (Chan et al., 2001; Petrović et al., 2017; Sangermano et al., 2010; Zubair et al., 2017). ML helps to detect urban change with cellular automata models (Feng et al., 2016; Moghaddam & Samadzadegan, 2009). Studies then focus on evaluating densification potentials in neighborhoods (Eggimann et al., 2021), analysing the dynamics of urban change from building alteration activities (Lai & Kontokosta, 2019), investigating the land use intensity from new masterplans (Gong et al., 2014), looking at the abandonment of residential areas (Xu et al., 2019), studying the evolution of the urbanization level in a metropolitan area (Grekousis et al., 2013).

Continue2A For the **urban form**, studies analyze the spatial structure of cities. Urban morphology was investigated in architectural scales (Gil et al., 2012; Hanna, 2007; Li et al., 2020;

Thomas et al., 2010). Urban areas were delineated by their vertical extensions (Arribas-Bel et al., 2019) or by their land cover extension (Liu et al., 2019). Other publications addressed the problem to predict the height of buildings (Biljecki et al., 2017) or derived them in time (Farella et al., 2021). Lee et al., and Yu (2017) studied map generalization tasks of cartography.

3.2.2. *Socioeconomic*

The papers covering socioeconomic aspects in urban areas were categorized into (A) socioeconomic attributes, (B) land economy, and (C) social issues (see Fig. 3). The large majority considered social issues (26/39 papers), with only a small number of socio-economic attributes (3/39 papers).

A The earliest paper aiming to detect **socioeconomic attributes** by Grove and Roberts (1980) studies the social and economic variation of British towns. Then, socioeconomic attributes were predicted in neighbourhoods (Dong et al., 2019) and GDP was investigated in relation to geographic predictors (Chen et al., 2020).

B **Land economy** falls almost equally into the prediction of retail attributes and real estate prices. When looking at the **retail attributes**, some studies predicted locations of stores (Satman & Altunbey, 2014; Xu et al., 2016). Other publications analyzed success indicators of retail store locations (Karamshuk et al., 2013) and hotels (Yang et al., 2015). For **real estate prices**, publications predicted market values of houses (Kauko, 2009; Xue et al., 2020), rent prices of residential units (Santibanez et al., 2015), or the real estate prices in different cities from the same country (Tchuente & Nyawa, 2021). Two publications studied the factors or amenities that drive prices for green building projects (Ma & Cheng, 2017) or land prices (Gao & Asami, 2007).

C **Social issues** included (a) social inequality, (b) gentrification, (c) social vulnerability and (d) crime prediction (see Fig. 3). Of these, crime prediction was the most prominent topic (11/26 papers).

1 **Social inequality** covers a broad range of topics. Most papers that aim to detect inequalities by studying unequal access from urban services as different as health (Lalloué et al. 2013; Mayaud et al., 2019); leisure space (Wang & Zhang, 2017); or digital services (Singleton, Alexiou, and Savani 2020). Others analyse differences in welfare (Wójcik & Andruszek, 2021) or aim to predict future income distribution (Auerbach et al., 2017). Only one paper investigates the impact of inequality by studying isolation (Wang et al., 2018).

2 **Gentrification** is addressed by studying socioeconomic characteristics in urban neighbourhoods (Alejandro & Palafox, 2019; Palafox & Ortiz-Monasterio, 2020; Reades et al., 2019; Walks & Maaranen, 2008).

3 The **social vulnerability** papers analyze how urban communities will respond or adapt to hazards. As such, this work complements the analysis of hazards in the environment category (see Section 3.2.3). All papers here focus on identifying vulnerable groups, either before a hazard occurs (Cutter & Finch, 2008; Dong et al., 2020), or investigate and monitor the impact of a hazard on welfare (Knippenberg et al., 2019; Allen et al., 2016).

4 Recently, there has been increasing interest in the topics of **safety and crime prediction**. Patterns of crime were identified by Mohammed and Baiee (2020) and a comparison of safety level in

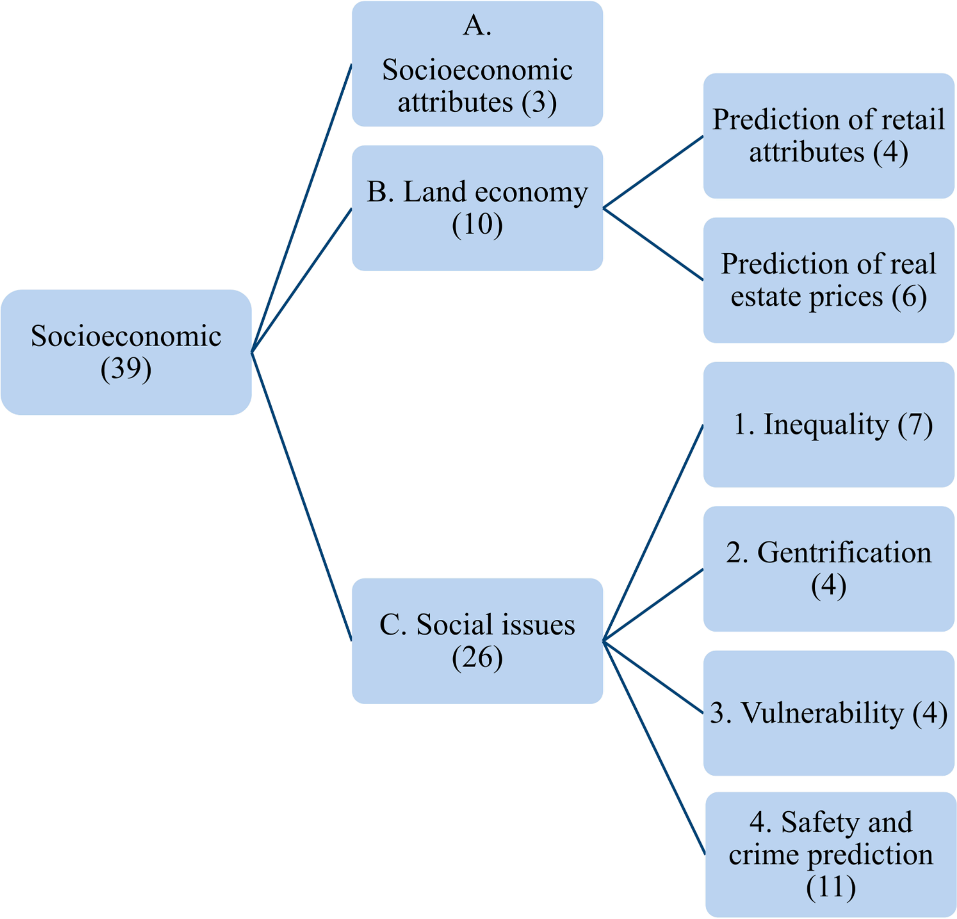
Socioeconomic (39)
A. Socioeconomic attributes (3)
B. Land economy (10)
C. Social issues (26)
Prediction of retail attributes (4)
Prediction of real estate prices (6)
1. Inequality (7)
2. Gentrification (4)
3. Vulnerability (4)
4. Safety and crime prediction (11)

Fig. 3. Topics studied in the socioeconomic category. The tree-plot shows the main research issues investigated in papers of the socioeconomic category. We showed the number of papers in brackets.

cities was studied by Kourtit et al., and Traian Pele (2021). Other authors predicted crime (Bappee et al., 2020; Cichosz, 2020; Dash et al., 2019; De Nadai et al., 2020; Kim et al., 2021; Lin et al., 2018; Redfern et al., 2020; Yang et al., 2018) or identified spatial and temporal factors for crime prediction (Yi et al., 2018) .

3.2.3. Environment

For the environment, we distinguished studies on (A) physical system (6/30 papers) from (B) hazard and risk (see Fig. 4).

A Studies on the environment as a **physical system** are categorized into weather (5/9 papers) and ecological aspects. For the **weather**, studies analyzed mostly temperatures in cities by studying land surface temperature in relation to land use attributes (Osborne & Alvares-Sanches, 2019; Sun et al., 2019), assessing heatwave thresholds (Park & Kim, 2018) or mapping Local Climate Zones (Bechtel et al., 2019). A method to predict turbulent air flows in the urban environment was developed by Xiao et al. (2019). In **ecology**, studies investigated the occurrence of ravens (Baltensperger et al. (2013), the ecological footprint of urban areas (Yao, 2012), the carbon storage of urban trees (Strohbach & Haase, 2012), and socio-ecological indicators of urban soil (Bonilla-Bedoya et al., 2021).

B For analyzing **hazards and associated risks** in urban systems, we distinguish flood risk prediction and detection of pollution. A variety of studies analyzed **flooding** mainly for classification and prediction purposes. The studied topics were the classification of the severity of flood events based on rainfall intensities (Ke et al., 2020), susceptibility maps (Tehrany et al., 2019; Zhao et al., 2019) and flood risk maps of cities (Darabi et al., 2019; Eini et al., 2020; Motta et al., 2021). For **pollution** in urban areas, most papers focus on predicting air pollution. Studies predicted PM2.5 based on meteorological data (Banga et al., 2021; Deters et al., 2017), CO2 emissions from metereological and socioeconomic variables (Li & Sun, 2021), carbon emission from urban blocks (Zhang et al., 2021), air pollution on roads by using traffic and meteorological data (Arnaudo et al., 2020; Suleiman et al., 2019). Studies investigated the relations between land use and air quality in urban areas (Brokamp et al., 2017; Champendal et al., 2014; Liu et al., 2015). Related to COVID-19 disease, studies analyzed the relationship between pollution levels and COVID-19 spread (Magazzino et al., 2020); Mirri et al., 2021) or analyzed the changes in the air quality from lockdowns (Shi et al., 2021). For other kinds of pollution, studies investigated chemical pollution from industrial areas in air and water (Shi & Zeng, 2014) and noise pollution (Hernandez-Jayo & Goñi, 2021; Torija & Ruiz, 2015).

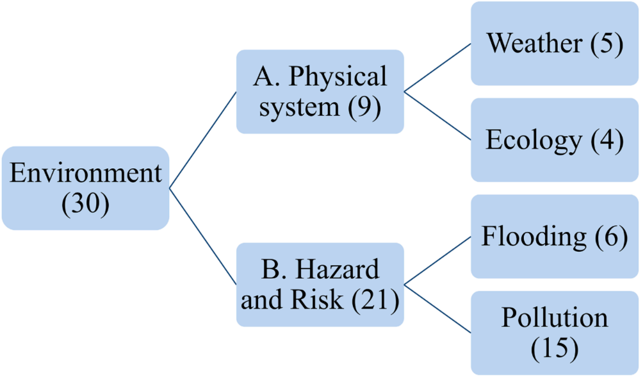

Fig. 4. Topics studied in the environmental category. The tree-plot shows the main research issues investigated in papers of the environmental category. We show the number of papers per topic in brackets.

3.2.4. Infrastructure

The investigated infrastructures were predominantly (A) transport (33/55 papers), followed by (B) energy, (C) water and sewer system, and (D) waste. Gas was only considered by one publication. Other networked infrastructures, such as information and communication technologies have not been considered in the urban ML literature thus far (see Fig. 5).

A Studies on **transportation** infrastructures mainly focus on (A.1) mobility and behavior (31 out of 33 papers), while 2 out of 33 papers analyze (A.2) physical infrastructure (see Fig. 5).

A.1 From the **mobility and behavior** perspective, studies detect transportation system properties.

1 Studies detected **transportation modes** (Aschwanden et al., 2019; Badii et al., 2021; Bjerre-Nielsen et al., 2020; Tang et al., 2018; Zhu et al., 2016), analyzed driving modes in a framework to estimate vehicle emission (Lehmann & Gross, 2017), and predicted travel mode and destinations (Truong et al., 2021)
2 **Clusters of mobility** data were studied to characterize the spatial-temporal properties of urban areas (Jiang et al., 2012; Kim, 2020; Wang et al., 2021; Xie et al., 2018). Wang et al., and Wang (2020) proposed a clustering method to analyze traffic data.
3 For **bike-sharing** systems, studies examined public opinion (Taleqani et al., 2019), identified suitable locations to place bike stations (Chen et al., 2015), predicted the number of available bikes and free bike slots (Collini et al., 2021)
4 When looking at **electric vehicles**, studies predicted locations of charging pools (Straka et al., 2020) and investigated the charging behavior by predicting the departure time and energy needs (Shahriar et al., 2021).
5 More broadly, Oke et al. (2019) studied **urban typologies** based on different urban dimensions to investigate the relationships between mobility and environmental sustainability.
6 Most studies (13/31 papers) analyzed **traffic characteristics** for predicting traffic speed (Ma et al., 2017; Magalhaes et al., 2021), traffic congestion spots (Awan et al., 2021; Majumdar et al., 2021; Qin et al., 2020; Saldana-Perez et al., 2019), traffic flows (Moretti et al., 2015) and traffic flow in relation to air vehicle emissions (Alam et al., 2018; Nyhan et al., 2016), commuting patterns between cities (Spadon et al., 2019), and driving distance in relation to the built environment and demographic (Ding et al., and Næss (2018)). When studying road accidents and events, studies looked at how to predict short-term car crashes (Bao et al., 2019) or studied a way to detect traffic-related events (Alomari et al., 2021).

A.2. From the **physical infrastructure** perspective, two papers look at the structural characteristics of transportation networks. The topology of road network was compared in different cities (Strano et al., 2013) and the road network vulnerability was analyzed against river flooding (Abdulla & Birgisson, 2020).

A In the **energy** sector, papers predicted the energy use. They built bottom-up approaches that estimated the energy demand or consumption in buildings at different scales and types (Abbasabadi & Azari, 2019; Ali et al., 2020; Carrera et al., 2021; Kontokosta & Tull, 2017; Ma & Cheng, 2016; Nutkiewicz et al., 2018; Rahman et al., 2018; Robinson et al., 2017; Zekić-Sušac et al., 2021; W. Zhang et al., 2018).

B For the **water** sector, some studies developed predictions of the water use. They modelled residential water consumption and clustered residences accordingly (Aksela & Aksela, 2011), studied how to predict the water usage in urban areas from mobility data (Smolak et al., 2020), analyzed outdoor residential water demand (Gage &

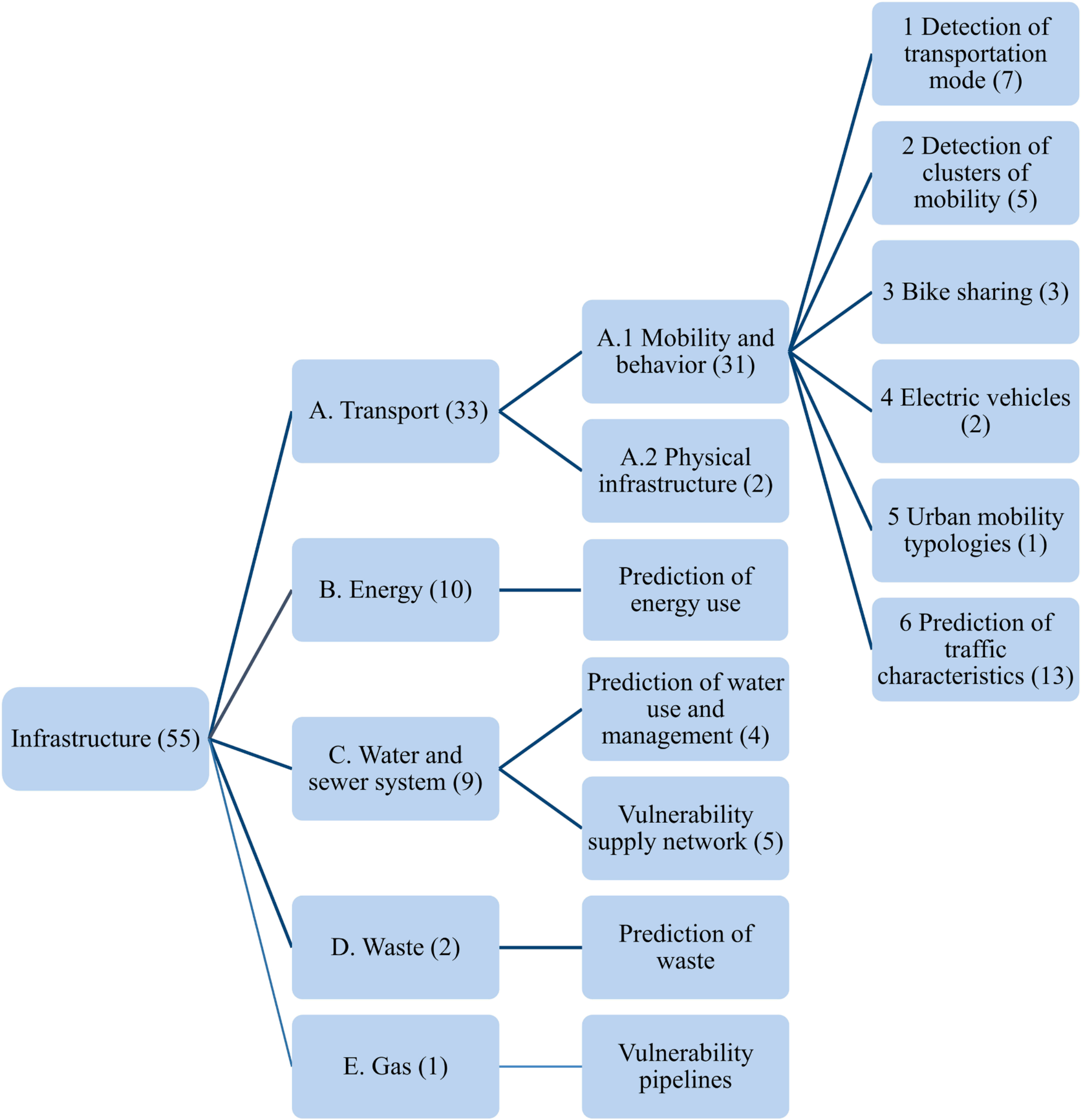
Infrastructure (55)
A. Transport (33)
B. Energy (10)
C. Water and sewer system (9)
D. Waste (2)
E. Gas (1)
A.1 Mobility and behavior (31)
A.2 Physical infrastructure (2)
Prediction of energy use
Prediction of water use and management (4)
Vulnerability supply network (5)
Prediction of waste
Vulnerability pipelines
1 Detection of transportation mode (7)
2 Detection of clusters of mobility (5)
3 Bike sharing (3)
4 Electric vehicles (2)
5 Urban mobility typologies (1)
6 Prediction of traffic characteristics (13)

Fig. 5. . Topics studied in the infrastructure category. The tree-plot shows the main research issues investigated in papers of the infrastructure category. We showed the number of papers for topics in brackets.

Cooper, 2015), modelled water demand to optimize water distribution (Rozos, 2019). Other authors investigated the vulnerabilities of the water supply infrastructure. They investigated leakages scenarios in the urban water supply system (Candelieri et al., 2013) and predicted pipe failure and breakage (Konstantinou & Stoianov, 2020; Kutyłowska, 2017; Winkler et al., 2018). The only study that looked at the sewer system was the one of Liu et al., and Prigiobbe (2021), who predicted groundwater infiltration into the sewer network.

C Two studies studied how to predict **waste** in cities. They investigated the amount of solid waste (Ayeleru et al., 2021) and looked at how much municipal waste could be used to produce energy (Kaya et al., 2021).

D Li et al., and Wang (2019) predicted vulnerabilities of the underground **gas** pipeline network in a city.

3.3. Data

Machine learning studies reveal critical and hidden information in datasets. For our analysis on the underlying data, we start with an overview of the frequency, at which different types of data were used. We distinguish numeric (e.g., csv) data, vector data, remote sensing and raster maps. Fig. 6 shows a heatmap of the numeric and vector data across the different topical categories. We listed data used at least in two papers in alphabetical order. We calculated the percentages of the total number of papers that used each data type over the total number of reviewed papers.

The most popular data was Demographic data (29%), which describe the size of the population in an area. Points of Interest (POI) data (28%), which report locations and labels of public services and private businesses in the city, were primarily used for land use and socioeconomic analyzes, but also in environment and infrastructure (Fig. 6). Not surprisingly, socioeconomic data (22%; incl. GDP per capita, ownership, income, education, employment, subsidy, and tax assessment characteristics) was most prominent in the socioeconomic category. Road data (24%; information about road network structure) was most often used for land use, and less frequently for infrastructure studies.

Although social media and telecommunication datasets are gaining prominence, especially when it comes to privacy (de Montjoye et al., 2018), we found that such data is still not frequently used: 4% of the papers used Twitter and 3% Foursquare. Twitter, Foursquare and Flickr

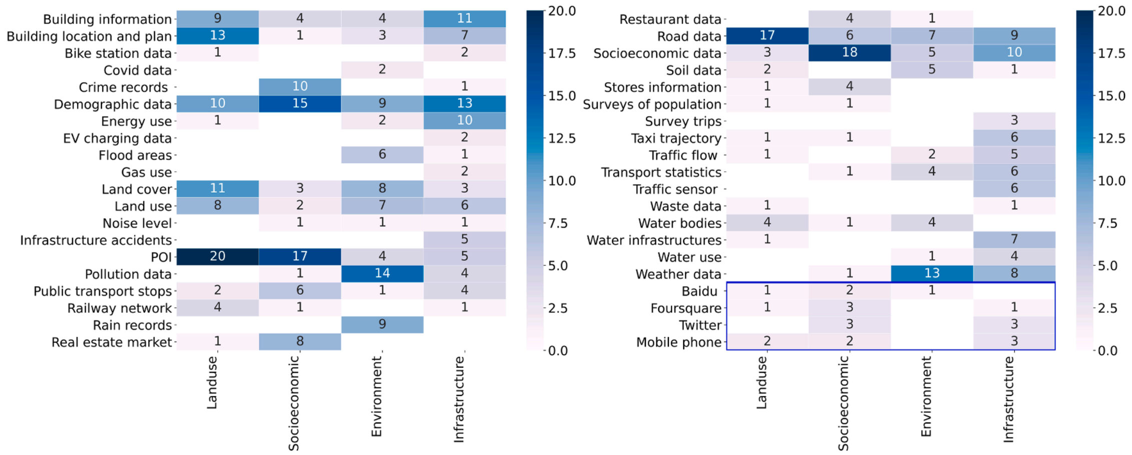

Fig. 6. Heatmap of numeric tables and vector datasets, in alphabetical order. We reported the type of data used and the number of papers that reported their use by each category. We highlighted the social media and telecommunication data.

(1%) were the only social media platforms in this review. 2% of the papers used data from the search engine Baidu, 4% used mobile phone data. Only one publication used Wifi connections (Xu et al., 2016).

Fig. 7 shows remote sensing data and raster data, which were used in 13% of the reviewed papers. Generally, remote sensing and raster data were used most in land use and environment, and less in the infrastructure category. The most common datasets in land use and environment studies used digital elevation models (DEM) (7%). Satellite imagery of the land surface (7%) and topographic maps (4%) were important for land use; night light data (4%) and vegetation indexes (2%) had applications in different categories. Other data accounted for 1% to 2% of papers.

Next, we compared the underlying data in each topic. In all topics, we found that there is a large heterogeneity in datasets used even for similar problems. In addition, papers often chose data without reporting a systematic methodology that guided the selection processes. However, we can still identify one or two common datasets in most topics. For example, in detection of functional areas, papers rely mainly on POI data, to which researchers have been adding different types of mobility data. Taxi trajectories were used by Yuan et al. (2012)), bicycle stations and their rental records were used by Zhang et al., and Du (2018), while mobile phone data and origin-destination (OD) data trucks were used by Zhai et al. (2019). Similarly, for land use change detection, research is conventionally built on land use or land cover data as it represents the real distribution of land use at a certain time, and then complemented by datasets such as DEM, road and population data, or demographic information.

3.4. ML methods

In this section, we analyze the methods adopted per topic. We group the studies into four categories: 1) supervised, 2) unsupervised, 3) a mix of unsupervised and supervised or 4) natural language processing methods. Fig. 8 shows the distribution of methods per category. We calculated the ratio over the total number of methods. We omitted the paper of Spadon et al. (2019)) as an outlier in the count, who used 34 supervised algorithms to avoid distortions.

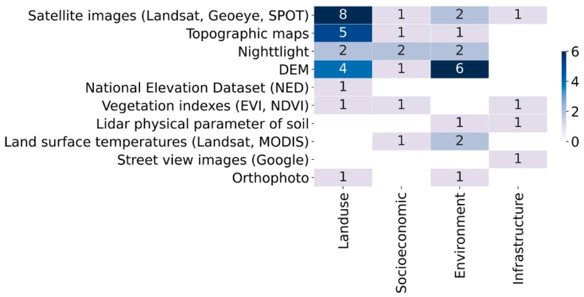

Fig. 7. Heatmap of remote sensing and raster data. We reported the type of data used and the number of papers that reported their use by each category.

We find that supervised methods dominate across topics, with infrastructure most prominently, followed by environment, and equally, by land use and environment. Unsupervised methods were mainly adopted by socioeconomic topics, followed by land use, infrastructure, and environment topics. A mix of unsupervised and supervised methods is mainly used for socioeconomic topics, followed by infrastructure, environment and land use. Natural language processing was used mostly by the land use category, whereas socioeconomic and environmental problems do not use it.

Because of their prominence, we provide a closer analysis of the most popular algorithms for supervised and unsupervised ML. Fig. 9 (left) shows the number of times papers used specific supervised algorithms. For supervised learning, we listed algorithms used in at least two papers. Despite the wide range of algorithms, papers tended to use mainly a few. Neural networks (NN), random forests (RF), support vector machines (SVM), gradient boosting decision trees (GBDT), decision trees (DT), K-nearest Neighbour (KNN) and logistic regression were the most frequently used supervised algorithms. Less frequently papers combined supervised ML algorithm with Cellular Automata analyzes. Studies that adopted only unsupervised ML algorithms (Fig. 9, right) used mostly PCA for data selection and k-means for clustering purposes.

Further, we analyzed the link between topics and methods. Table 3

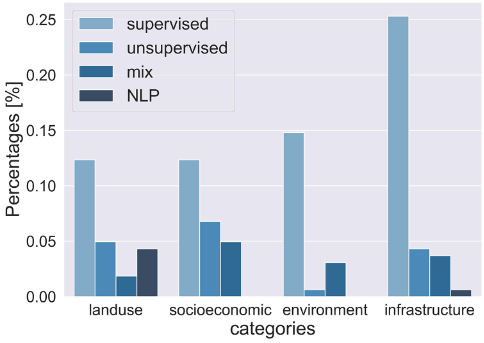

Fig. 8. Machine learning methods per topic. The histogram shows the ratio of papers that used supervised, unsupervised, a mix of unsupervised and supervised algorithms or neural language processing (NLP) methods per category.

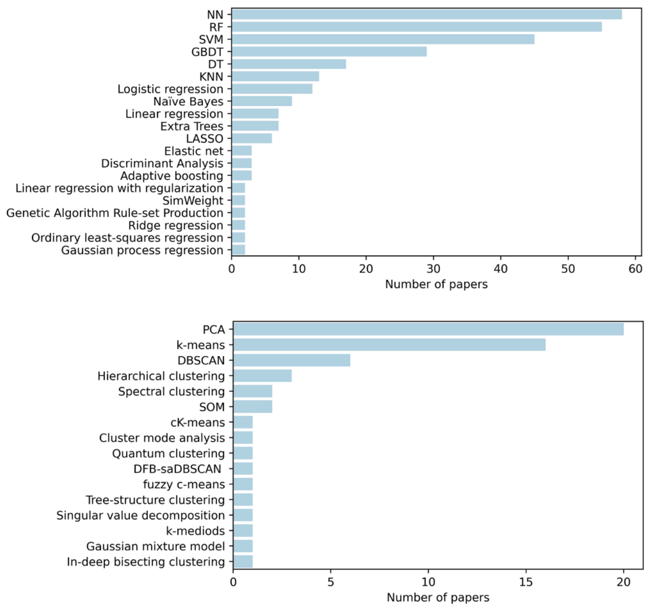

Fig. 9. Histograms of supervised (left) and unsupervised (right) applications of machine learning algorithms.

shows our results. As for the datasets, we see a broad variety and heterogeneity of methods. Few algorithms were used several times per topic, with neural networks (NN) in energy use prediction as the top algorithm (7 papers in the topic), followed by NN in land use change and air pollution modeling (5 papers). The greatest heterogeneity in selection of supervised algorithms was found within the infrastructure category. Of the studies that used a mix of supervised and unsupervised ML, principal component analysis (PCA) was the most used unsupervised algorithm for feature selection as input to supervised ML methods, followed by the k-means algorithm. Natural language processing was prominently used when studies investigated functional areas in the land use category. These papers use Word2Vec, Place2Vec, DMR, TF-IDR, and LDA topic models to discover the thematic structure of spatial data.

3.5. Patterns in parameter selection

In this section, we analyze patterns in parameter selection in supervised and unsupervised ML.

In supervised learning, we investigated the training and testing information reported. Although the parameter choices have an important impact on the results, we found that often authors did not report any information about the selection of training and testing parameters. The ones that did report mostly divided the data into two datasets: the training dataset comprised usually 70%-80% of data, and the testing dataset was in between 30%-20% of the total data. Some papers instead divided the training and testing data by years because they were developing temporal analyzes (Dash et al., 2018, Yang et al., 2018). Limited papers used three datasets for training, validation, and testing purposes (Arnaudo et al., 2020; Kutyłowska, 2017; Lee et al., 2017).

Few papers systematically report selection of hyperparameters, and thus far there is no common standard. For example, Xu et al. (2016) listed the hyperparameters for different algorithms in bullet points. Satman and Altunbey (2014), Grekousis et al. (2015), Kontokosta et al. (2017), Ma and Cheng (2016) and Li et al. (2019) reported hyperparameters or information about the architecture of NN in tables. However, the vast majority of papers failed to report hyperparameters and details about the model architecture. If authors reported hyperparameters, those mainly appeared in the body text rather than in a detailed and systematic fashion via tables or figures, making the information not easily readable. This finding was in line with earlier criticism on the lack of reporting of hyperparameters in artificial neural networks (Grekousis et al., 2019).

For **unsupervised ML,** we analyze how clustering and PCA were used. For **clustering algorithms,** we distinguish two approaches: (i) the optimal number of clusters is determined by using algorithms systematically, or (ii) the number of clusters is determined by users and algorithms assign data to each cluster. For the first category, for instance, optimal number of clusters were selected by using the partition coefficient and classification entropy by Grekousis et al. (2013)). Clustergram was used to identified clusters for k-means by Singleton et al. (2020), which plots a series of potential k values. The optimal number of k-means clusters was evaluated by using the silhouette coefficient by Shi and Zheng (2014). In the second category, authors typically selected the number of clusters based on empirical evaluations of the case studies (e.g., Chang et al., 2018, Aksela & Aksela, 2011; Lehmann & Gross, 2017).

For **principal component analysis** (PCA), we found that different methods were used when selecting the number of principal components (PCs). For example, Cutter and Finch (2008), Wang and Zhang (2017) and Gao and Asami (2007) used the Kaiser criterion. Champendal et al. (2014)) used Kaiser, Joliffe and Catell criteria. Other papers selected the numbers of PCs that captured the majority of the total variance without fixing a priori a number or following any specific criteria (Lalloue' et al., 2013, Dong et al., 2020 and Ke et al., 2020, Reades et al., 2019, Suleiman et al., 2019).

4. Discussion

In this section, we discuss the implications of the results from the review, identify research gaps and objectives for future research across the dimensions of our scoping review.

4.1. Spatial distribution

We found that case studies were mainly located in Asia, North America, and Europe (Fig. 1). The lack of studies in other regions or comparative studies may be driven by limited access and availability of data, which strongly affects computational studies. This gap presents an opportunity to develop methods for data sparse environments (Brajard et al., 2020; Nikparvar & Thill, 2021), and comparative studies that identify the impact of different data sets and granularities of the results.

Some of the most promising research avenues in data sparse contexts are the creation of satellite Earth observations (EO). Related approaches have proven successful to monitor agri-food systems (Nakalembe et al., 2021) or to understand urban sprawl and land-use change (Sankhala & Singh, 2014). Other methodologies improve spatial data collection in some regions. For the social dimension of sustainable urban development, the DesInventar methodology is an example, with a focus on disaster losses (Panwar & Sen, 2020). However, this methodology reports still important limitations regarding the level of urban disaggregated data and consistent coverage (Osuteye et al., 2017).

Digital technologies can be beneficial for sustainable development goals (Nosratabadi et al., 2020; Sachs et al., 2019; Vinuesa et al., 2020). However, there are risks and downsides that countries must identify and tackle through integrated strategies and a focus on the leave-no-one-behind principle (Sachs et al., 2019). Some of these risks concern ethical issues, for example, the loss of jobs for lower-skilled workers, the theft of digital identities, invasion of privacy by governments or businesses, and discrimination based on personal data. Therefore, responsible implementations and use of AI methods should address these topics and principles. Furthermore, model interpretability of AI algorithms must be addressed jointly with requirements and constraints related to data privacy, model confidentiality, fairness, and accountability (Barredo et al., 2020).

As a result, there is a need to conduct research that focuses on:

Table 2
List of data used in similar studies. The table shows the data information adopted by papers grouped in by topic of study.

TOPIC	DATA	PAPER
Detection of functional areas	POI. POI, taxi trajectory. POI, bicycle stations and rental records. POI, mobile phone data and OD data of trucks.	Hu et al., 2020. Yao et al., 2016. Yan et al., 2017 Yuan et al., 2012, Zhai et al., 2019, Zhang et al., 2018
Detection of land use change	Land use, roads, DEM. Land use, roads, population, DEM. Land use, roads, population. Landsat - land use, roads, population, POI. Land cover. SPOT - land cover, roads, population, tax-break development areas.	Pijanowski et al., 2014, Pijanowski et al., 2002. Huang et al., 2009, Petrović et al., 2017, Sangermano et al., 2010, Zubair et al., 2017
Prediction of store location	POI, retail stores data. POI, user data from Baidu and Wifi connection data.	Satman & Altunbey, 2014, Xu et al., 2016
Model land surface temperature (LST)	Landsat-LST data, Night-time light, building information, road, POI, water surface ratio and vegetation index. Landsat - LST data, land cover	Osborne & Alvares-Sanches, 2019, Sun et al., 2019
Predict flood risk	Inundated areas, DEM, land use, rain data, curve number, slope, urban density and texture, quality of buildings, road, POI, distance to river, distance to channel, population, socioeconomic data (household; age; women population; education; immigrant; tenant; women head of household; employment) Inundated areas, DEM, rain data, land use/land cover, slope percent, curve number, distance to river, distance to channel, and depth to groundwater, urban density, quality of buildings, age of buildings, population, socioeconomic conditions divided in levels. Inundated areas, weather data.	Eini et al., 2020 Darabi et al., 2019, Motta et al., 2021
Assessment of air quality related to land use	Pollution data in points, land use, distance to roads, airports, hydrographic networks, population, maximal power of heating systems, traffic. Pollution data in points, land use, expressway and major roads, weather variables. Pollution data in points, land cover, traffic intensity, DEM, population density, greenspace, emission point sources.	Champendal et al., 2014 Brokamp et al., 2017, Liu et al., 2015
Detection of transportation mode	GPS trajectory, road network, bus stops, subways lines, real time road condition. Socioeconomic data and transport statistics from travel survey in households. Georeferenced trips from survey, satellite images. Wi-Fi and Bluetooth traces, railway network, public transport information. Mobility data of trajectories (speed, acceleration, location, etc.) from mobile data, transport facilities, green areas	Zhu et al., 2016 Tang et al., 2018 Aschwanden et al., 2019, Badii et al., 2021, Bjerre-Nielsen et al., 2020

- Developing ML applications for prominent urban or rapidly urbanizing regions in Southeast Asia, India, or Africa. This entails the development of the appropriate data fusion, assimilation, or sampling techniques to generate databases that are fit for machine learning applications in data sparse contexts.
- Comparative studies across continents, validating findings from the different areas of studies while acknowledging the diversity of underlying data sets (see also data).

4.2. Categories

Although a broad range of topics has been investigated by using ML (see Figs. 2-5), some topics are underrepresented. In the socioeconomic category, studies are missing that investigate social attributes, for example, in urban cultures and the labour market. Social justice issues were partially treated with respect to inequalities and gentrification studies. Given the prominence of discussion related to accessible and affordable housing in fast-growing areas (Kramer, 2018; Rodríguez-Pose & Storper, 2019), and equality and equity in access to public urban services and infrastructures (Martínez et al., 2018; Modai-Snir & van Ham, 2018), more research in these domains is urgently needed. In the environmental field, which only marks 18% of the publications, more contributions are needed in evaluating more kinds of disaster risks (e.g. sea-level rise), natural resources consumption and ecosystem services. In the infrastructure category, crucial subjects such as waste management, logistics, renewable energy systems were underrepresented, which is surprising given the growing interest in circularity for cities (Sachs et al., 2019).

Further, there is a lack of cross domain publications. Two prominent areas that require research across urban systems are sustainability and resilience. Cities will be exposed to an increasing number of extreme events and will have to balance long-term sustainable and green development with resilience to hazards (Elmqvist et al., 2019). For both areas, infrastructural, social and environmental aspects need to be combined. Therefore, ML approaches are particularly promising. Sustainable urbanism is increasingly becoming smart and data-driven (Bibri 2020). However, our findings on a lack of cross-cutting publications in sustainability and resilience confirmed Milojevic-Dupont and Creutzig (2021), who found that despite their potential, ML tools were not common in climate change research communities. Similarly, despite urban resilience becoming a research and policy priority (Krishnan et al., 2021; Meerow & Newell, 2019), we identified one paper that used ML to evaluate resilience (Knippenberg et al., 2019), while resilience was discussed by Cutter and Finch (2008); Dong et al. (2020)); Motta et al. (2021)).

In sum, the most pressing research needs are related to:

- Developing machine learning applications targeting new areas of research such as (i) labour market and cultural attributes of cities; (ii) accessible and affordable housing in fast growing areas and equity or fairness in access to urban services; (iii) a circular urban economy, including waste management and logistics; and (iv) climate change and related extreme events.
- Designing applications that focus on the interplay between the different urban environmental infrastructural and socioeconomic settings with land use, especially in the areas of sustainability and urban resilience.

4.3. Data

Despite the popularity of data sources such as demographic data and POI, our findings showed a significant heterogeneity in the datasets

Table 3

Machine learning algorithms used in urban analyzes. For each topic, we report the algorithm used in the analyzes. Number of publications in bracket for algorithms with more than one application in the topic. If a paper uses more than one method, all are reported under the same topic. Abbreviations: Decision Trees (DT), Density-based spatial clustering of applications with noise (DBSCAN), Dirichlet Multinomial Regression (DMR), Gradient Boosting Decision Tree (GBDT), k-nearest neighbors (KNN), Latent Dirichlet Allocation (LDA), Least absolute shrinkage and selection operator (LASSO), Neural Network (NN), Principal Component Analysis (PCA), Support Vector Machines (SVM), Term frequency–inverse document frequency (TF-IDF), Topical Word Embeddings (TWE).

Pattern in methods	Topic of study	Algorithms
Supervised		
	Land use and form	
	Land use change	NN (5), SVM (3), DT (1), SimWeight (2).
	Land use from social data	RF.
	Land use intensity	NN, Linear regression.
	Abandonment of rural areas	RF, SVM, Naïve Bayes.
	Identify informal settlements	RF (2), Logistic regression (2), DT, Discriminant Analysis.
	Urban extension	NN, RF, GBDT, and their ensemble models.
	Map generalization	DT, SVM, KNN, Naïve Bayes.
	Predict building heights	RF (2), Ordinary least-squares regression, GBDT, SVM, NN.
	Socioeconomic	
	Mapping GDP	RF.
	Real estate price	GBDT (4), RF (3), NN (2), SVM (2), KNN, Cubist, Partial Least Squares, Adaptive boosting.
	Tax assessment	RF.
	Hazards	LASSO, RF.
	Site location of stores	NN, LASSO, SVM, KRR, RF, GBDT, Learning to rank models.
	Predict popularity of stores	SVM, DT, Linear regression with regularization.
	Predict popularity of hotels	NN, SVM, Boosted regression, Linear regression.
	Crime prediction	RF (4), SVM (3), NN (2), LASSO, DT, KNN, GBDT, Polynomial regression, Logistic regression, Naïve Bayes.
	Gentrification	NN, RF.
	Predict well-being	SVM, RF, GBDT, LASSO.
	Influenza monitoring	SVM.
	Environment	
	Temperatures and heatwaves	RF (2), Multivariate Adaptive Regression Splines, Ordinary least-squares regression, GBDT.
	Flooding	SVM (2), Genetic Algorithm Rule-Set Production (2), DT(2), NN (2), SVM (2), RF, KNN, Maximum Entropy, Logistic regression, Naïve-Bayes.
	Air pollution	NN (5), RF (5), GBDT (3), Adaptive Boosting (2), SVM (2), Linear regression with Bayesian Ridge Regularization, Extra tree, DT.
	Air pollution and COVID	NN (2), SVM, KNN, Extra tree.
	Noise pollution	Logistic regression, LDA, KNN, DT, SVM, Naïve Bayes.
	Ecology-animal detection	RF.
	Ecological footprint	NN.
	Carbon storage maps	RF.
	Urban soil indicators	RF.
	Infrastructure	
	Crash risk prediction	NN.
	Detect road traffic events	SVM, Naïve Bayes, Logistic regression.
	Estimate transportation mode	NN (3), RF (3), SVM (2), DT, Logistic regression, Bayesian network, Extra trees, GBDT.
	Prediction travel behaviour	SVM, Multinomial logit.
	Prediction traffic speed	NN, KNN, RF.
	Prediction traffic congestion	NN (2), SVM.
	Prediction driving distance	GBDT.
	Prediction traffic flow	NN (4), Log-linear regression model.
	Prediction intercity flow	34 different algorithms (see Spadon et al., 2020).
	Predict electric vehicles sessions	RF, SVM, GBDT, NN.
	Popular location for charging pools	Logistic regression, RF, GBDT.
	Site location bike stations	NN, Linear Regression-and-Ranking (LRR).
	Prediction of energy use	NN (7), SVM (5), RF (5), Linear regression (4), GBDT (4), Elastic net (3), Extra Trees (2), Ridge regressor (2), KNN (2), DT(2), LASSO, Naïve Bayes, Generalized Linear Model, Logistic regression, Bagging regressor.
	Gas pipeline vulnerability	NN, SVM.
	Road network vulnerability	NN, KNN, RF, Logistic regression, and Naïve Bayes.
	Water network vulnerability	RF (2), NN (2), SVM, DT, Naïve Bayes, Poisson GLM, Logistic Poisson GLM, GBDT, Probabilistic Random Forest, Discriminant Analysis, Time Linear Model, Time Exponential Model.
	Infiltration in water sewer	Logistic regression.
	Water use prediction	RF (2), SVM, Extra trees.
	Water supply management	NN.
	Predict solid waste	NN (2), SVM (2), RF, Extra trees, GBDT.
Unsupervised	**Land use and form**	
	Urban form	k-means (3), DBSCAN.
	Land use from social data	Spectral clustering.
	Urban identities	PCA (3), k-means (2)
	Socioeconomic	
	Socioeconomic attributes	PCA, Cluster mode analysis.
	Real estate price	PCA, SOM.
	Hazard vulnerability	PCA (2)
	Public leisure space	PCA.
	Neighborhood isolation	DBSCAN.
	Health	PCA, Hierarchical clustering, SOM.

(*continued on next page*)

Table 3 (*continued*)

Pattern in methods	Topic of study	Algorithms
	Crime	DBSCAN.
	Gentrification	PCA.
	Environment	
	Air pollution	k-means.
	Air and water pollution	k-means.
	Infrastructure	
	Cluster mobility data	k-means (2), Spectral clustering, Hierarchical clustering, Quantum clustering, DBSCAN, DFB-saDBSCAN algorithm, cK-means.
	Urban mobility typologies	Hierarchical clustering.
	Road network topology	PCA.
Unsupervised + Supervised	**Land use and form**	
	Model land use change	NN + fuzzy c-means.
	Identify densification potential	SVM, RF + PCA.
	Architectural style	SVM + PCA.
	Socioeconomic	
	Socioeconomic attributes	LASSO, GBDT + DBSCAN.
	Green building market	GBDT + DBSCAN.
	Digital inequalities	GBDT, KNN + k-means.
	Crime prediction	Logistic regression, SVM, DT, RF + PCA. CCRF model + Tree-structure clustering. Bayesian negative binomial + PCA.
	Safety level	SVM, GBDT + k-means.
	Gentrification	RF + PCA.
	Environment	
	Flooding	DT, Discriminant Analysis, SVM, KNN, Ensemble models + PCA.
	Air pollution	NN, GBDT, SVM + PCA. NN, RF + PCA.
	Noise pollution	NN, Gaussian Process Regression (GPR) + PCA.
	Air flow	Gaussian Process Regression (GPR) + Singular value decomposition.
	Infrastructure	
	Cluster vehicle prediction	KNN + K-means, Gaussian Mixture Model, K-mediods
	Speed prediction	RF, GBDT, Multivariate regression + k-means.
	Predict the number of available bikes	GBDT, RF, NN + k-means.
	Energy	Multiple linear regression, Nonlinear regression, RF, DT, KNN, NN + k-means.
	Water use prediction	Regression models + k-means.
	Water network vulnerability	Linear regression + In-deep bisecting clustering algorithm.
Natural language Processing	**Land use and form**	
	Functional areas	Word2Vec, TWE + HDBSCAN. Word2Vec, RF + k-means. Place2Vec + k-means. DMR, TF-IDF + k-means (2) Place2Vec.
	Urban change and construction activities	LDA.
	Infrastructure	
	Public opinion on dockless bike-sharing systems	TF-IDF + Naïve Bayes, Logistic regression, SVM.

used, even within the same topical category (see Table 2). Further, the rationale for the selection of datasets was often unexplained. We assume that many case studies followed a pragmatic approach built on the availability of data. With the significant differences in data sets both in scale and content, the results of ML algorithms are not comparable, hampering the development of urban analytics and evidence-based urban planning. While data availability is an issue, we argue that standards and explicit methods to select data for the different topical areas will help producing more generalizable and comparable results.

The first promising way ahead is the integration of human sensing data. Surprisingly, urban ML still does not significantly rely on data from mobile phones or social media, confirming findings of Grekousis (2019) on ANNs. While such data is promising to understand interaction patterns or the use of urban services, there are also challenges related to privacy and data protection that need to be addressed.

Another promising research avenue is investigating approaches to integrate data from various sources. This issue has been recently raised within the data fusion and Big Data research domain (Favaretto et al., 2020; Kar & Dwivedi, 2020; Yang et al., 2020). For the geographical domain, most Big Data are produced with space and time stamps. These are samples of sequential observations from various remote, in-situ, mobile and human sensing systems or simulations, which lead to an increased need for cross-scale data fusion, including integration across various sources and interpolation across spatiotemporal domains (Yang et al., 2020). The use of a large amount of data without solid reasoning can lead to misinterpretation of findings.

In sum, research on data is needed that that focuses on:

- Analyzing the impact of the different dataset choices for urban ML problems within and across the different topical categories. Based on this, standards and explicit methods can be developed to select data for the various topical areas.
- Integrating sensing data such as from mobile phones or social media, while respecting privacy and data protection.

- Investigating new methods that merge data from different sources to derive meaningful results, especially in the context of data sparse environments (see also spatial distribution).

4.4. Methods

In this scoping review, we distinguish four categories of methods (see Section 4.4). Natural language processing was primarily used in land use topics, even though it may also have promising applications in other domains. While we found that most papers lean on supervised ML (Table 3), there is a broad variety of supervised, unsupervised, or mixed methods. As the method selection depends on the scientific problem to analyze, future research should compare different methods within specific topics. Like for the data category, systematic comparisons are beneficial to understanding the significance and output variety of using specific algorithms.

We found that NN, RF, SVM, gradient boosting DT, DTs, KNN and logistic regression were the most popular supervised ML algorithms, while PCA and k-means were the most popular unsupervised ML algorithms. In terms of the rationale for methodology selection, different reasons are put forward. Overall, some algorithms performed well in predicting or classifying data for specific problems, therefore they grew a strong reputation over time. Supervised ML methods are often chosen based on their complexity, overfitting properties, parameter requirements, data requirements, and interpretability of results. Looking at the complexity of methods, RF and SVM are more complex compared to logistic regression, ordinary least square regressions or LASSO because they account for non-linear relationships (Cichosz, 2020; Knippenberg et al., 2019; Kontokosta & Tull, 2017). Whereas, NN are efficient predictors because they have higher computational complexity embedded in their network topology (Grekousis, 2019; Ma et al., 2017). When looking at overfitting properties, RF is often chosen because it avoids the overfitting of data (Chen et al., 2020; Jochem et al., 2018; Xu et al., 2019). For the parameter requirements, KNN does not require parameters in input for classification problems (Lee et al., 2017; Ma et al., 2017). For the interpretability, DT is often selected for the easy interpretability of the results (Lee et al., 2017). When looking at the data requirements, NN was not used because of the insufficient number of data (Knippenberg et al., 2019), or a NN based on an ensemble method was used to deal with data scarcity (Zhang et al., 2021). When studying unsupervised ML methods, PCA is often selected because it helps to synthesize datasets in a few sets of principal components (dimensionality reduction) and still preserves interpretability by loadings (Cutter & Finch, 2008; Laskari et al., 2008; Wang & Zhang, 2017).

New methods are needed that link the research on ML algorithms to urban science. Kitchin (2014) already discussed the challenges of new epistemologies and paradigm shifts that the use of big data and data-driven analytics might bring, highlighting the need for critical reflection on the epistemological implications of the data and analytical revolution. Falco (2015) demanded a human-centered approach. We argue that studies in ML and urban science should be aware of these challenges when developing appropriate methods that connect analytical frameworks with the broader urban science and policy. Especially knowledge transfer is a promising concept that fosters and supports collaborations between research organizations, business entities, and public sector (Heinimann & Hatfield, 2017).

Therefore, there is a need for:

- In depth explorations of using natural language processing for other fields than land use topics.
- Comparing different machine learning algorithms within specific topics to study potential differences in results and their relationship.
- Develop new methodological frameworks that go beyond the mere application of ML, but rather establish novel ways to explain, translate and transfer the results from ML to urban sciences, practice, and policy.

4.5. Patterns in parameter selection

We found that most papers tended to train and test the models for supervised learning, while only some authors included validation in their research. Often, papers did not report the parameters selected to build the models and the information about training-testing phases, leading to a lack of reproducibility. Although tables and figures are beneficial to the reader, only few studies presented parameter information in these formats. These gaps were already identified by Grekousis (2019) for ANNs. We argue that a consistent way of reporting parameters is vital to increase reproducibility and advance the state-of-the-art.

Furthermore, we found that papers select number of clusters and principal components for unsupervised learning by using different approaches. Often, the approach that authors used to define the number of clusters is not appropriately justified.

Benchmarking analyses might help to prepare better standards. This need is confirmed also by other reviews about the use of deep learning applications (Grekousis, 2019; Ma et al., 2019), which means it is a recurrent need in the field. For benchmark analyses, there should be benchmark datasets accessible to everyone. Therefore, the problem of transparency of pattern selection is linked to the accessibility of data.

Thus, there is an urgent need for research on:

- Studying protocols for reporting parameters in publications for supervised and unsupervised algorithms
- Analyzing the impact of the ML algorithm results across different topical categories to define joint standards and increase reproducibility

5. Conclusions

In this paper, we set out to review the state-of-the-art in Machine learning based on spatial data for sustainable cities. Since this is an emerging and highly dynamic research field, we conducted a scoping review to (i) map out the most prominent topics, data sources, ML algorithms, and approaches to parameter selection, (ii) determine the most prominent patterns and challenges in the use of ML, (iii) identify knowledge gaps to guide future research. We reviewed papers covering different ML algorithms across all aspects of sustainable urban systems, which are divided into the categories of land use and urban form, socioeconomic, environment, and infrastructure. Overall, the analyses helped to create a classification of ML approaches according to topics, methods, and data sources.

There are three main takeaways from this study. First, there are still ample opportunities to evolve this research field. This can be achieved by investigating missing topics or by working on cross-domain or comparative case studies (see Section 4). As ML and AI are gaining momentum, we expected that applications of these technologies will serve to solve contingent problems around pressing issues pertaining to sustainability, such as circularity and resilience.

Second, there is still a need to standardize the selection of data, algorithms, and parameters. Systematic comparisons of the data and algorithms selection can help in exploring the significance of these methods and the impacts of the results, while systematically reporting parameters increase the reproducibility of works and the transparency of the analytical process (see Section 4). Grekousis (2019) confirmed partially these findings for ANNs. This lack of transparency and systematic comparison of ML methods also hinders application. Sustainable urban planning decisions and policies that influence the urban environment require concrete reasoning and clarity.

Third, spatial ML will benefit in shifting attention to the creation and types of datasets. There are limitations in developing spatial ML studies in data-sparse areas (such as the Global South). Moreover, studies use often heterogeneous data from different data sources. Studying how to integrate and merge data will help spatial data-driven analyses to become more meaningful. Along with these themes, ethical studies about the role of technology and its possible risks for communities and individuals (e.g. the loss of jobs for lower-skilled workers, the theft of digital identities, invasion of privacy by governments or businesses, discrimination based on personal data) should be addressed to develop a more conscientious use of spatially-explicit technology.

This scoping review has some limitations related to the scope and the focus on spatial urban data. Although studies related to street view images have been on rise recently, we did not include papers that adopted images as the sole source of information to develop urban analyzes. We refer the reader to literature reviews in computer vision for urban analytics (Biljecki & Ito, 2021; Ibrahim et al., 2020). For ML applications in the field of remote sensing, we refer to Lary et al., 2015, Ma et al., 2019, Maxwell et al., 2018 and Zhu, Tuia et al., 2017. Another limitation is that we developed an initial mapping of an emerging field, while a systematic review would include all published papers in literature by following a protocol (e.g. PRISMA protocol by Moher et al. (2009)).

The scientific community can use this review as a guideline to understand which approaches and data sets have been used for which type of urban problem. Further, our analyses help shape a comprehensive understanding of the use between ML and geospatial data. Moreover, we identified several promising areas for future research in all domains, ranging from the need for more comparative studies and to an improved understanding of the impact of the selection of data sets, algorithms, or parameters. We especially stress the need to foster explainable machine learning approaches and invest in knowledge transfer to create impact and help equip cities with the tools they need to address the many challenges they are facing.

Funding

This work was supported by the Delft University of Technology (TU Delft).

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

References

Abbasabadi, N., & Azari, R. (2019). A data-driven framework for urban building operational energy use modeling. *Simulation Series, 51*(8), 71–77.

Abdulla, B., & Birgisson, B. (2020). Predicting road network vulnerability to fluvial flooding using machine learning classifiers: case study of houston during hurricane harvey. In *Construction Research Congress 2020: Computer Applications* (pp. 38–47).

Aksela, K., & Aksela, M. (2011). Demand estimation with automated meter reading in a distribution network. *Journal of Water Resources Planning and Management, 137*(5), 456–467. https://doi.org/10.1061/(asce)wr.1943-5452.0000131

Alam, M. S., Duffy, P., Hyde, B., & McNabola, A. (2018). Downscaling national road transport emission to street level: A Case study in dublin, Ireland. *Journal of Cleaner Production, 183*, 797–809. https://doi.org/10.1016/j.jclepro.2018.02.206

Alejandro, Y., & Palafox, L. (2019). Gentrification prediction using machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11835*(LNAI), 187–199. https://doi.org/10.1007/978-3-030-33749-0_16

Ali, U., Shamsi, M. H., Bohacek, M., Purcell, K., Hoare, C., Mangina, E., & O'Donnell, J. (2020). A data-driven approach for multi-scale gis-based building energy modeling for analysis, planning and support decision making. *Applied Energy, 279*, Article 115834. https://doi.org/10.1016/j.apenergy.2020.115834

Alomari, E., Katib, I., Albeshri, A., Yigitcanlar, T., & Mehmood, R. (2021). Iktishaf+: a big data tool with automatic labeling for road traffic social sensing and event detection using distributed machine learning. *Sensors, 21*(9).

Arnaudo, E., Farasin, A., & Rossi, C. (2020). A comparative analysis for air quality estimation from traffic and meteorological data. *Applied Sciences (Switzerland), 10* (13), 1–20. https://doi.org/10.3390/app10134587

Arribas-Bel, Daniel, M., Garcia-López, & Elisabet, V.-M. (2019). Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, (September 2018), Article 103217. https://doi.org/10.1016/j.jue.2019.103217

Aschwanden, G. D. P. A., Wijnands, J. S., Thompson, J., Nice, K. A., Zhao, H., & Stevenson, M. (2019). Learning to walk: Modeling transportation mode choice distribution through neural networks. *Environment and Planning B: Urban Analytics and City Science, 48*(1), 186–199. https://doi.org/10.1177/2399808319862571

Auerbach, J., Chaganti, V., Blackburn, C., Barton, H., Ghai, B., Zegura, E., Blunt, T., Meng, A., & Flores, P. (2017). Using data science as a community advocacy tool to promote equity in urban renewal programs: An analysis of atlanta's anti-displacement tax fund. *ArXiv*.

Awan, F. M., Minerva, R., & Crespi, N. (2021). Using noise pollution data for traffic prediction in smart cities: experiments based on LSTM recurrent neural networks. *IEEE Sensors Journal, 21*(18), 20722–20729.

Ayeleru, O. O., Fajimi, L. I., Oboirien, B. O., & Olubambi, P. A. (2021). Forecasting municipal solid waste quantity using artificial neural network and supported vector machine techniques: A case study of Johannesburg, South Africa. *Journal of Cleaner Production*, 289.

Badii, C., Difino, A., Nesi, P., Paoli, I., & Paolucci, M. (2021). Classification of users' transportation modalities from mobiles in real operating conditions. *Multimedia Tools and Applications*.

Badmos, O. S., Rienow, A., Callo-Concha, D., Greve, K., & Jürgens, C. (2019). Simulating slum growth in Lagos: An integration of rule based and empirical based model. *Computers, Environment and Urban Systems, 77*(July), Article 101369. https://doi.org/10.1016/j.compenvurbsys.2019.101369

Baltensperger, A. P., Mullet, T. C., Schmid, M. S., Humphries, G. R. W., Kövér, L., & Huettmann, F. (2013). Seasonal observations and machine-learning-based spatial model predictions for the common raven (Corvus Corax) in the urban, sub-arctic environment of Fairbanks, Alaska. *Polar Biology, 36*(11), 1587–1599. https://doi.org/10.1007/s00300-013-1376-7

Banga, A., Ahuja, R., & Sharma, S. C. (2021). Performance analysis of regression algorithms and feature selection techniques to Predict PM2.5 in smart cities. *International Journal of Systems Assurance Engineering and Management*. https://doi.org/10.1007/s13198-020-01049-9

Bao, J., Pan, L., & Satish, V. U. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis and Prevention, 122*, 239–254. https://doi.org/10.1016/j.aap.2018.10.015

Bappee, F. K., Petry, L. M., Soares, A., & Matwin, S. (2020). Analyzing the impact of foursquare and streetlight data with human demographics on future crime prediction. *ArXiv*, 1–15.

Barredo, A., Alejandro, Natalia, D.-R., Javier, D. S., Adrien, B., Tabik, S., Barbado, A., Salvador, G., Sergio, G.-L., Daniel, M., Richard, B., Raja, C., & Francisco, H. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*(December 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Batty, M.. (2008). The size, scale, and shape of cities. *Science, 319*(5864), 769–771.

Bechtel, B., Paul, J. A., Christoph, B., Jürgen, B., Oscar, B., Jason, C., Matthias, D., Cidália, F., Tamás, G., Julia, H., Peter, H., Ariane, M., Gerald, M., Chao, R., Linda, S., Panagiotis, S., Marie, L. V., Guang, X., & Yong, X. (2019). Generating WUDAPT level 0 data – current status of production and evaluation. *Urban Climate, 27*(August 2018), 24–45. https://doi.org/10.1016/j.uclim.2018.10.001

Bibri, & Simon, E. (2020). Compact urbanism and the synergic potential of its integration with data-driven smart urbanism : An extensive interdisciplinary literature review. *Land Use Policy, 97*(September 2019), Article 104703. https://doi.org/10.1016/j.landusepol.2020.104703

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*(August), Article 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Biljecki, F., Ledoux, H., & Stoter, J. (2017). Generating 3D city models without elevation data. *Computers, Environment and Urban Systems, 64*, 1–18. https://doi.org/10.1016/j.compenvurbsys.2017.01.001

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1* (1), 363–376. https://doi.org/10.1162/qss_a_00018

Bjerre-Nielsen, A., Minor, K., Sapieżyński, P., Lehmann, S., & Dreyer Lassen, D. (2020). Inferring transportation mode from smartphone sensors: Evaluating the potential of Wi-Fi and bluetooth. *PLoS ONE, 15*(7), 1–24. https://doi.org/10.1371/journal.pone.0234003

Bonilla-Bedoya, S., López-Ulloa, M., Mora-Garcés, A., Macedo-Pezzopane, J. E., Laura, S., & Miguel, Á. H. (2021). Urban soils as a spatial indicator of quality for urban socio-ecological systems. *Journal of Environmental Management, 300*(September). https://doi.org/10.1016/j.jenvman.2021.113556

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science, 44*, Article 101171.

Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment, 151*, 1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066

Candelieri, A., Archetti, F., & Messina, E. (2013). Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation.

WIT Transactions on Ecology and the Environment, 171, 107–117. https://doi.org/10.2495/WRM130101

Carrera, B., Peyrard, S., & Kim, K. (2021). Meta-regression framework for energy consumption prediction in a smart city: A case study of Songdo in South Korea. *Sustainable Cities and Society, 72*.

Celebi, M.E., & Aydin, K.T.A.T.T. (2016). "Unsupervised learning algorithms.".

Champendal, A., Kanevski, M., & Huguenot, P. E. (2014). Air pollution mapping using nonlinear land use regression models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8581 LNCS(PART 3)*, 682–690. https://doi.org/10.1007/978-3-319-09150-1_50

Chan, J. C. W., Chan, K. P., & Yeh, A. G. O. (2001). Detecting the nature of change in an urban environment: A comparison of machine learning algorithms. *Photogrammetric Engineering and Remote Sensing, 67*(2), 213–225.

Chang, M.-C., Buš, P., Tartar, A., Chirkin, A., & Schmitt, G. (2018). Big-data informed citizen participatory urban identity. In *Computing for a Better Tomorrow: The 36th International Conference on Education and Research in Computer Aided Architectural Design in Europe (ECAADe 2018) 2(June 2019)* (pp. 669–678).

Chang, M. C., Bus, P., & Schmitt, G. (2017). Feature extraction and K-means clustering approach to explore important features of urban identity. In *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017 2017-Decem:1139–44*. https://doi.org/10.1109/ICMLA.2017.00015

Chaturvedi, V., & de Vries, W. T. (2021). Machine learning algorithms for urban land use planning: A review. *Urban Science, 5*(3), 68.

Chen, L., Zhang, D., Pan, G., Ma, X., Yang, D., Kushlev, K., Zhang, W., & Li, S. (2015). Bike sharing station placement leveraging heterogeneous urban open data. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 571–575). https://doi.org/10.1145/2750858.2804291

Chen, Q., Ye, T., Zhao, N., Ding, M., Ouyang, Z., Jia, P., Yue, W., & Yang, X. (2020). Mapping China's regional economic activity by integrating points-of-interest and remote sensing data with random forest. *Environment and Planning B: Urban Analytics and City Science, 0*(0), 1–19. https://doi.org/10.1177/2399808320951580

Cheng, J., Turkstra, J., Peng, M., Du, N., & Ho, P. (2006). Urban land administration and planning in China: opportunities and constraints of spatial data models. *Land Use Policy, 23*(4), 604–616. https://doi.org/10.1016/j.landusepol.2005.05.010

Cichosz, P.. (2020). Urban crime risk prediction using point of interest data. *ISPRS International Journal of Geo-Information, 9*(7). https://doi.org/10.3390/ijgi9070459

Colding, J., Colding, M., & Barthel, S. (2020). The smart city model: a new panacea for urban sustainability or unmanageable complexity? *Environment and Planning B: Urban Analytics and City Science, 47*(1), 179–187.

Collini, E., Nesi, P., & Pantaleo, G. (2021). Deep learning for short-term prediction of available bikes on bike-sharing stations. *IEEE Access, 9*, 124337–124347. https://doi.org/10.1109/ACCESS.2021.3110794

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Understanding collective activity patterns of a city from social media. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 58–65).

Cutter, S. L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the National Academy of Sciences of the United States of America, 105*(7), 2301–2306. https://doi.org/10.1073/pnas.0710375105

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., & Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. *Journal of Hydrology, 569*(November 2018), 142–154. https://doi.org/10.1016/j.jhydrol.2018.12.002

Dash, S. K., Safro, I., & Sakrepatna Srinivasamurthy, R. (2019). Spatio-temporal prediction of crimes using network analytic approach. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* (pp. 1912–1917). https://doi.org/10.1109/BigData.2018.8622041

Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering, 2017*. https://doi.org/10.1155/2017/5106045

Ding, C., Cao, X.(J.), & Næss, P. (2018). Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transportation Research Part A: Policy and Practice, 110*(March), 107–117. https://doi.org/10.1016/j.tra.2018.02.009

Dong, L., Ratti, C., & Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences of the United States of America, 116*(31), 15447–15452. https://doi.org/10.1073/pnas.1903064116

Dong, S., Esmalian, A., Farahmand, H., & Mostafavi, A. (2020). An integrated physical-social analysis of disrupted access to critical facilities and community service-loss tolerance in urban flooding. *Computers, Environment and Urban Systems, 80* (November 2019), Article 101443. https://doi.org/10.1016/j.compenvurbsys.2019.101443

Eggimann, S., Wagner, M., Ho, Y. N., Züger, M., Schneider, U., & Orehounig, K. (2021). Geospatial simulation of urban neighbourhood densification potentials. *Sustainable Cities and Society, 72*.

Eini, M., Seyed Kaboli, H., Rashidian, M., & Hedayat, H. (2020). Hazard and vulnerability in urban flood risk mapping: Machine learning techniques and considering the role of urban districts. *International Journal of Disaster Risk Reduction, 50*(May). https://doi.org/10.1016/j.ijdrr.2020.101687

Elmqvist, T., Andersson, E., Frantzeskaki, N., McPhearson, T., Olsson, P., Gaffney, O., Takeuchi, K., & Folke, C. (2019). Sustainability and resilience for transformation in the urban century. *Nature Sustainability, 2*(4), 267–273. https://doi.org/10.1038/s41893-019-0250-1

Falco, G. J. (2015). City resilience through data analytics: A human-centric approach. *Procedia Engineering, 118*, 1008–1014. https://doi.org/10.1016/j.proeng.2015.08.542

Fallatah, A., Jones, S., & Mitchell, D. (2020). Object-based random forest classification for informal settlements identification in the Middle East: Jeddah a case study. *International Journal of Remote Sensing, 41*(11), 4421–4445. https://doi.org/10.1080/01431161.2020.1718237

Farella, E., Mariarosaria, E.Ö., & Remondino, F. (2021). 4D building reconstruction with machine learning and historical maps. *Applied Sciences (Switzerland), 11*(4), 1–24.

Favaretto, M., de Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of big data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE, 15*(2).

Feng, Y., Liu, Y., & Batty, M. (2016). Modeling urban growth with GIS based cellular automata and least squares SVM rules: A case study in Qingpu–Songjiang area of Shanghai, China. *Stochastic Environmental Research and Risk Assessment, 30*(5), 1387–1400. https://doi.org/10.1007/s00477-015-1128-z

Gage, E., & Cooper, D. J. (2015). The influence of land cover, vertical structure, and socioeconomic factors on outdoor water use in a Western US city. *Water Resources Management, 29*(10), 3877–3890. https://doi.org/10.1007/s11269-015-1034-7

Gao, X., & Asami, Y. (2007). Effect of Urban Landscapes on Land Prices in Two Japanese Cities. *Landscape and Urban Planning, 81*(1–2), 155–166. https://doi.org/10.1016/j.landurbplan.2006.11.007

Gil, J., Nuno Beirão, J., Montenegro, N., & Duarte, J. P. (2012). On the discovery of urban typologies: Data mining the many dimensions of urban form. *Urban Morphology, 16*(1), 27–40.

Gong, J., Chen, W., Liu, Y., & Wang, J. (2014). The intensity change of urban development land: Implications for the city master plan of Guangzhou, China. *Land Use Policy, 40*, 91–100. https://doi.org/10.1016/j.landusepol.2013.05.001

Goodchild, M. (2001). Issues in spatially explicit modeling. special workshop on agent-based models of land use/land cover change. *Center for Spatially Integrated Social Science, Irvine*.

Goodchild, M. F., & Haining, R. P. (2004). GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science, 83*(1), 363–385.

Grekousis, G.. (2019). Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems, 74*(September 2018), 244–256. https://doi.org/10.1016/j.compenvurbsys.2018.10.008

Grekousis, G., Manetos, P., & Photis, Y. N. (2013). Modeling urban evolution using neural networks, fuzzy logic and GIS: The case of the athens metropolitan area. *Cities, 30*(1), 193–203. https://doi.org/10.1016/j.cities.2012.03.006

Grove, D. M., & Roberts, C. A. (1980). Principal components and cluster analysis of 185 large towns in England and Wales. *Urban Studies, 17*(1), 77–82. https://doi.org/10.1080/00420988020080091

Guigoz, Y., Giuliani, G., Nonguierma, A., Lehmann, A., Mlisa, A., & Ray, N. (2017). Spatial data infrastructures in Africa: A gap analysis. *Journal of Environmental Informatics, 30*(1).

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of google scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217.

Hanna, S.. (2007). Automated representation of style by feature space archetypes: Distinguishing spatial styles from generative rules. *International Journal of Architectural Computing, 5*(1), 1–23. https://doi.org/10.1260/147807707780913001

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition*. New York: Springer.

Hegde, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment – A review. *Safety Science, 122*(May 2019), Article 104492. https://doi.org/10.1016/j.ssci.2019.09.015

Heinimann, H. R., & Hatfield, K. (2017). *Infrastructure Resilience Assessment, Management and Governance – State and Perspectives*, PartF1.

Hernandez-Jayo, U., & Goñi, A. (2021). Zaratamap: Noise characterization in the scope of a smart city through a low cost and mobile electronic embedded system. *Sensors, 21*(5), 1–26.

Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., & Cui, H. (2020). A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Computers, Environment and Urban Systems, 80*(November 2019). https://doi.org/10.1016/j.compenvurbsys.2019.101442

Huang, B., Xie, C., Tay, R., & Wu, B. (2009). Land-use-change modeling using unbalanced support-vector machines. *Environment and Planning B: Planning and Design, 36*(3), 398–416. https://doi.org/10.1068/b33047

Ibrahim, M. R., Haworth, J., & Cheng, T. (2020). Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities, 96*(November 2019), Article 102481. https://doi.org/10.1016/j.cities.2019.102481

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science, 34*(4), 625–636. https://doi.org/10.1080/13658816.2019.1684500

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2012). Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 95–102). https://doi.org/10.1145/2346496.2346512

Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computers, Environment and Urban Systems, 69*(January), 104–113. https://doi.org/10.1016/j.compenvurbsys.2018.01.004

Boulos, K., Maged, N., Peng, G., & Vopham, T. (2019). An overview of GeoAI applications in health and healthcare. *International Journal of Health Geographics, 18*(1), 1–9. https://doi.org/10.1186/s12942-019-0171-2

Kao, A.., & Poteet, S.R.T.A.T.T. (2007). "Natural language processing and text mining.".

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research – moving away from the 'What' towards the 'Why. *International Journal of Information Management, 54*(June), Article 102205. https://doi.org/10.1016/j.ijinfomgt.2020.102205

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013). Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Part F1288* (pp. 793–801). https://doi.org/10.1145/2487575.2487616

Kauko, T.. (2009). The housing market dynamics of two Budapest neighbourhoods. *Housing Studies, 24*(5), 587–610. https://doi.org/10.1080/02673030903082328

Kaya, K., Ak, E., Yaslan, Y., & Oktug, S. F. (2021). Waste-to-energy framework: An intelligent energy recycling management. *Sustainable Computing: Informatics and Systems, 30*(November 2020). https://doi.org/10.1016/j.suscom.2021.100548

Ke, Q., Tian, X., Bricker, J., Tian, Z., Guan, G., Cai, H., Huang, X., Yang, H., & Liu, J. (2020). Urban pluvial flooding prediction by machine learning approaches – a case study of Shenzhen city, China. *Advances in Water Resources, 145*(November 2019), Article 103719. https://doi.org/10.1016/j.advwatres.2020.103719

Kim, D., Jung, S., & Jeong, Y. (2021). Theft prediction model based on spatial clustering to reflect spatial characteristics of adjacent lands. *Sustainability (Switzerland), 13*(14).

Kim, K.. (2020). Identifying the structure of cities by clustering using a new similarity measure based on smart card data. *IEEE Transactions on Intelligent Transportation Systems, 21*(5), 2002–2011. https://doi.org/10.1109/TITS.2019.2910548

Kitchin, R.. (2014). Big data, new epistemologies and paradigm shifts. *Big Data and Society, 1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Knippenberg, E., Jensen, N., & Constas, M. (2019). Quantifying household resilience with high frequency data: Temporal dynamics and methodological options. *World Development, 121*, 1–15. https://doi.org/10.1016/j.worlddev.2019.04.010

Konstantinou, C., & Stoianov, I. (2020). A comparative study of statistical and machine learning methods to infer causes of pipe breaks in water supply networks. *Urban Water Journal, 17*(6), 534–548. https://doi.org/10.1080/1573062X.2020.1800758

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy, 197*, 303–317.

Kourtit, K., Mazurencu Marinescu Pele, M., Peter, N., & Traian Pele, D. (2021). Safe cities in the new urban world: A comparative cluster dynamics analysis through machine learning. *Sustainable Cities and Society, 66*.

Kramer, A.. (2018). The unaffordable city: Housing and transit in North American Cities. *Cities, 83*, 1–10. https://doi.org/10.1016/j.cities.2018.05.013

Krishnan, S., Aydin, N. Y., & Comes, M. (2021). Planning support systems for long-term climate resilience: A critical review. *Urban Informatics and Future Cities, 465*.

Kutyłowska, M.. (2017). Prediction of failure frequency of water-pipe network in the selected city. *Periodica Polytechnica Civil Engineering, 61*(3), 548–553. https://doi.org/10.3311/PPci.9997

Lai, Y., & Kontokosta, C. E. (2019). Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems, 78*(August), Article 101383. https://doi.org/10.1016/j.compenvurbsys.2019.101383

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2015). Machine learning in geosciences and remote sensing. *Geoscience Frontiers, 7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Laskari, A., Hanna, S., & Derix, C. (2008). Urban identity through quantifiable spatial attributes: Coherence and dispersion of local identity through the automated comparative analysis of building block plans. In *Design Computing and Cognition '08 - Proceedings of the 3rd International Conference on Design Computing and Cognition* (pp. 615–634).

Lee, J., Jang, H., Yang, J., & Yu, K. (2017). Machine learning classification of buildings for map generalization. *ISPRS International Journal of Geo-Information, 6*(10). https://doi.org/10.3390/ijgi6100309

Lehmann, A., & Gross, A. (2017). Towards vehicle emission estimation from smartphone sensors. In *Proceedings - 18th IEEE International Conference on Mobile Data Management, MDM 2017* (pp. 154–163). https://doi.org/10.1109/MDM.2017.29

Leyk, S., Gaughan, A. E., Adamo, S. B., Sherbinin, A.de, Deborah, B., Sergio, F., Amy, R., Stevens, F. R., Brian, B., & Charlie, F. (2019). The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data, 11*(3), 1385–1409.

Li, F., Wang, W., Xu, J., Yi, J., & Wang, Q. (2019). Comparative study on vulnerability assessment for urban buried gas pipeline network based on SVM and ANN methods. *Process Safety and Environmental Protection, 122*, 23–32. https://doi.org/10.1016/j.psep.2018.11.014

Li, W.. (2020). GeoAI: Where machine learning and big data converge in GIScience. *Journal of Spatial Information Science, 20*(20), 71–77. https://doi.org/10.5311/JOSIS.2020.20.658

Li, X., Cheng, S., Lv, Z., Song, H., Jia, T., & Lu, N. (2020). Data analytics of urban fabric metrics for smart cities. *Future Generation Computer Systems, 107*, 871–882. https://doi.org/10.1016/j.future.2018.02.017

Li, Y., & Sun, Y. (2021). Modeling and predicting city-level CO2 emissions using open access data and machine learning. *Environmental Science and Pollution Research, 28*(15), 19260–19271.

Lin, Y. L., Yen, M. F., & Yu, L. C. (2018). Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information, 7*(8). https://doi.org/10.3390/ijgi7080298

Liu, T., Ramirez-Marquez, J. E., Chandra Jagupilla, S., & Prigiobbe, V. (2021). Combining a statistical model with machine learning to predict groundwater flooding (or infiltration) into sewer networks. *Journal of Hydrology, 603*(November 2020). https://doi.org/10.1016/j.jhydrol.2021.126916

Liu, W., Li, X., Chen, Z., Zeng, G., León, T., Liang, J., Huang, G., Gao, Z., Jiao, S., He, X., & Lai, M. (2015). Land use regression models coupled with meteorology to model spatial and temporal variability of NO2 and PM10 in Changsha, China. *Atmospheric Environment, 116*(2), 272–280. https://doi.org/10.1016/j.atmosenv.2015.06.056

Liu, X., de Sherbinin, A., & Zhan, Y. (2019). Mapping urban extent at large spatial scales using machine learning methods with VIIRS nighttime light and MODIS daytime NDVI data. *Remote Sensing, 11*(10), 1–18. https://doi.org/10.3390/rs11101247

Ma, J., & Cheng, J. C. P. (2016). Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. *Applied Energy, 183*, 182–192. https://doi.org/10.1016/j.apenergy.2016.08.079

Ma, J., & Cheng, J. C. P. (2017). Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining. *Journal of Cleaner Production, 151*, 406–418. https://doi.org/10.1016/j.jclepro.2017.03.083

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing, 152*(March), 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *ArXiv*, 1–16.

Magalhaes, R. P., Lettich, F., Macedo, J. A., Nardini, F. M., Perego, R., Renso, C., & Trani, R. (2021). Speed prediction in large and dynamic traffic sensor networks. *Information Systems, 98*.

Magazzino, C., Mele, M., & Schneider, N. (2020). The relationship between air pollution and COVID-19-related deaths: An application to three French cities. *Applied Energy, 279*(May), Article 115835. https://doi.org/10.1016/j.apenergy.2020.115835

Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A., & Crooks, A. T. (2020). Detecting and mapping slums using open data: A case study in Kenya. *International Journal of Digital Earth, 13*(6), 683–707. https://doi.org/10.1080/17538947.2018.1554010

Majumdar, S., Subhani, M. M., Benjamin, R., Ashiq, A., & Rongbo, Z. (2021). Congestion prediction for smart sustainable cities using IoT and machine learning approaches. *Sustainable Cities and Society, 64*(September 2020), Article 102500. https://doi.org/10.1016/j.scs.2020.102500

Marsland, S.. (2014). *Machine learning: An algorithmic perspective, second edition* (2nd ed.). Chapman & Hall/CRC.

Martínez, C. F., Hodgson, F., Mullen, C., & Timms, P. (2018). Creating inequality in accessibility: The relationships between public transport and social housing policy in deprived areas of Santiago de Chile. *Journal of Transport Geography, 67*, 102–109.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing, 39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Mayaud, J. R., Tran, M., & Nuttall, R. (2019). An urban data framework for assessing equity in cities: Comparing accessibility to healthcare facilities in Cascadia. *Computers, Environment and Urban Systems, 78*(August), Article 101401. https://doi.org/10.1016/j.compenvurbsys.2019.101401

Meerow, S., & Newell, J. P. (2019). Urban resilience for whom, what, when, where, and why? *Urban Geography, 40*(3), 309–329.

Milojevic-Dupont, N., & Creutzig, F. (2021). Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society, 64*(September 2020), Article 102526. https://doi.org/10.1016/j.scs.2020.102526

Mirri, S., Roccetti, M., & Delnevo, G. (2021). The New York city Covid-19 spread in the 2020 spring: A study on the potential role of particulate using time series analysis and machine learning. *Applied Sciences (Switzerland), 11*(3), 1–19.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Modai-Snir, T., & van Ham, M. (2018). Neighbourhood change and spatial polarization: The roles of increasing inequality and divergent urban development. *Cities, 82*, 108–118.

Moghaddam, H. K., & Samadzadegan, F. (2009). Urban simulation using neural networks and cellular automata for land use planning Hamid Kiavarz Moghaddam, Farhad Samadzadegan. *Real Corp 2009, 6*(April), 571–577.

Mohammed, A.F., & Baiee, W.R. (2020). "EasyChair preprint the GIS based analysis criminal events using analysing crimes using machine learning technique.".

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., & Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7). https://doi.org/10.1371/journal.pmed.1000097

de Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., Cordes, N.de, Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., & Alex "Sandy" Pentland, and Linus Bengtsson. (2018). On the privacy-conscientious use of mobile phone data. *Scientific Data, 5*(1), Article 180286. https://doi.org/10.1038/sdata.2018.286

Moretti, F., Pizzuti, S., Panzieri, S., & Annunziato, M. (2015). Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing, 167*, 3–7. https://doi.org/10.1016/j.neucom.2014.08.100

Motta, M., Neto, M.deC., & Sarmento, P. (2021). A mixed approach for urban flood prediction using machine learning and GIS. *International Journal of Disaster Risk Reduction, 56*(February), Article 102154. https://doi.org/10.1016/j.ijdrr.2021.102154

Munn, Z., Peters, M., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). *S12874-018-0611-X.Pdf* (p. 143).

De Nadai, M., Xu, Y., Letouzé, E., González, M. C., & Lepri, B. (2020). Socio-economic, built environment, and mobility conditions associated with crime: A study of multiple cities. *Scientific Reports, 10*(1), 1–12. https://doi.org/10.1038/s41598-020-70808-2

Nakalembe, C., Inbal, B.-R., Rogerio, B., Guangxiao, H., Micheal, H., Christina, J. J., John, K., Kenneth, M., Felix, R., Shraddhanand, S., Ferdinando, U., Alyssa, K., Whitcraft, Y. L., Mario, Z., Ian, J., & Antonio, S. (2021). A review of satellite-based global agricultural monitoring systems available for Africa. *Global Food Security, 29*, Article 100543. https://doi.org/10.1016/j.gfs.2021.100543

Nikparvar, B., & Thill, J. C. (2021). Machine learning of spatial data. *ISPRS International Journal of Geo-Information, 10*(9).

Nosratabadi, S., Mosavi, A., Keivani, R., Ardabili, S., & Aram, F. (2020). State of the art survey of deep learning and machine learning models for smart cities and urban sustainability. *Lecture Notes in Networks and Systems, 101*(Ml), 228–238. https://doi.org/10.1007/978-3-030-36841-8_22

Nutkiewicz, A., Yang, Z., & Jain, R. K. (2018). Data-driven urban energy simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Applied Energy, 225* (June), 1176–1189. https://doi.org/10.1016/j.apenergy.2018.05.023

Nyhan, M., Sobolevsky, S., Kang, C., Robinson, P., Corti, A., Szell, M., Streets, D., Lu, Z., Britter, R., Barrett, S. R. H., & Ratti, C. (2016). Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. *Atmospheric Environment, 140*, 352–363. https://doi.org/10.1016/j.atmosenv.2016.06.018

Oke, J. B., Aboutaleb, Y. M., Akkinepally, A., Azevedo, C. L., Han, Y., Christopher Zegras, P., Ferreira, J., & Ben-Akiva, M. E. (2019). A novel global urban typology framework for sustainable mobility futures. *Environmental Research Letters, 14*(9). https://doi.org/10.1088/1748-9326/ab22c7

Openshaw, S., & Openshaw, C. (1997). *Artificial Intelligence in Geography*.

Osborne, P. E., & Alvares-Sanches, T. (2019). Quantifying how landscape composition and configuration affect urban land surface temperatures using machine learning and neutral landscapes. *Computers, Environment and Urban Systems, 76*(August 2018), 80–90. https://doi.org/10.1016/j.compenvurbsys.2019.04.003

Osuteye, E., Johnson, C., & Brown, D. (2017). The data gap: An analysis of data availability on disaster losses in Sub-Saharan African Cities. *International Journal of Disaster Risk Reduction, 26*(September), 24–33. https://doi.org/10.1016/j.ijdrr.2017.09.026

Palafox, L., & Ortiz-Monasterio, P. (2020). Predicting gentrification in Mexico City using neural networks. In *Proceedings of the International Joint Conference on Neural Networks 0–4*. https://doi.org/10.1109/IJCNN48605.2020.9207685

Panwar, V., & Sen, S. (2020). Disaster damage records of EM-DAT and desinventar: A systematic comparison. *Economics of Disasters and Climate Change, 4*(2), 295–317.

Park, J., & Kim, J. (2018). Defining heatwave thresholds using an inductive machine learning approach. *PLoS ONE, 13*(11), 1–11. https://doi.org/10.1371/journal.pone.0206872

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare, 13*, 141–146.

Petrović, M. S., Kovačević, M., Bajat, B., & Dragićević, S. (2017). Machine learning techniques for modelling short term land-use change. *ISPRS International Journal of Geo-Information, 6*(12). https://doi.org/10.3390/ijgi6120387

Pijanowski, B. C., Brown, D. G., Shellito, B. A., & Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: A land transformation model. *Computers, Environment and Urban Systems, 26*(6), 553–575. https://doi.org/10.1016/S0198-9715(01)00015-1

Pijanowski, B. C., Tayyebi, A., Doucette, J., Pekin, B. K., Braun, D., & Plourde, J. (2014). A big data urban growth simulation at a national scale: configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. *Environmental Modelling and Software, 51*, 250–268. https://doi.org/10.1016/j.envsoft.2013.09.015

Qin, K., Xu, Y., Kang, C., & Kwan, M. P. (2020). A graph convolutional network model for evaluating potential congestion spots based on local urban built environments. *Transactions in GIS, 24*(5), 1382–1401. https://doi.org/10.1111/tgis.12641

Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy, 212*(December 2017), 372–385. https://doi.org/10.1016/j.apenergy.2017.12.051

Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies, 56*(5), 922–942. https://doi.org/10.1177/0042098018789054

Redfern, J., Sidorov, K., Rosin, P. L., Corcoran, P., Moore, S. C., & Marshall, D. (2020). Association of violence with urban points of interest. *PLoS ONE, 15*(9 September)), 1–17. https://doi.org/10.1371/journal.pone.0239840

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy, 208*(October), 889–904. https://doi.org/10.1016/j.apenergy.2017.09.060

Rodríguez-Pose, A., & Storper, M. (2019). Housing, urban growth and inequalities: The limits to deregulation and upzoning in reducing economic and spatial inequality. *Urban Studies, 57*(2), 223–248. https://doi.org/10.1177/0042098019859458

Rozos, E.. (2019). Machine learning, urban water resources management and operating policy. *Resources, 8*(4), 173. https://doi.org/10.3390/resources8040173

Sachs, J., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N., & Rockström, J. (2019). Six transformations to achieve the SDGs. *Nature Sustainability*.

Saldana-Perez, M., Torres-Ruiz, M., & Moreno-Ibarra, M. (2019). Geospatial modeling of road traffic using a semi-supervised regression algorithm. *IEEE Access, 7*, 177376–177386. https://doi.org/10.1109/ACCESS.2019.2942586

Samuel, A. L. (1959). Some studies in machine learning. *IBM Journal of Research and Development, 3*(3), 210–229.

Sangermano, F., Eastman, J. R., & Zhu, H. (2010). Similarity weighted instance-based learning for the generation of transition potentials in land use change modeling. *Transactions in GIS, 14*(5), 569–580. https://doi.org/10.1111/j.1467-9671.2010.01226.x

Sankhala, S., & Singh, B. K. (2014). Evaluation of urban sprawl and land use land cover change using remote sensing and GIS techniques: A case study of Jaipur City, India. *International Journal of Emerging Technology and Advanced Engineering, 3*(1), 1–5.

Santibanez, S. F., Kloft, M., & Lakes, T. (2015). Performance analysis of machine learning algorithms for regression of spatial variables . A case study in the real estate industry. In *Geocomputation 2015 Conference Proceedings (Bork 2015)* (pp. 292–297).

Satman, M. H., & Altunbey, M. (2014). Selecting location of retail stores using artificial neural networks and google places API. *International Journal of Statistics and Probability, 3*(1). https://doi.org/10.5539/ijsp.v3n1p67

Shahriar, S., Al-Ali, A. R., Osman, A. H., Salam, D., & Mais, N. (2021). Prediction of EV charging behavior using machine learning. *IEEE Access, 9*, 111576–111586. https://doi.org/10.1109/ACCESS.2021.3103119

Shi, W., & Zeng, W. (2014). Application of K-means clustering to environmental risk zoning of the chemical industrial area. *Frontiers of Environmental Science and Engineering, 8*(1), 117–127. https://doi.org/10.1007/s11783-013-0581-5

Shi, Z., Congbo, S., Bowen, L., Gongda, L., Jingsha, X., Tuan, V. V., Robert, J. R. E., Weijun, L., William, J. B., & Roy, M. H. (2021). Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns. *Science Advances, 7*(3). https://doi.org/10.1126/sciadv.abd6696

Smolak, K., Kasieczka, B., Fialkiewicz, W., Rohm, W., Siła-Nowicka, K., & Kopańczyk, K. (2020). Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. *Urban Water Journal, 17*(1), 32–42. https://doi.org/10.1080/1573062X.2020.1734947

Spadon, G., Carvalho, A. C. P. L. F.de, Rodrigues-Jr, J. F., & Alves, L. G. A. (2019). Reconstructing commuters network using machine learning and urban indicators. *Scientific Reports, 9*(1), 1–13. https://doi.org/10.1038/s41598-019-48295-x

Straka, M., Falco, P. D., Ferruzzi, G., Proto, D., Poel, G. V. D., Khormali, S., & Buzna, L. (2020). Predicting popularity of electric vehicle charging infrastructure in urban context. *IEEE Access, 8*, 11315–11327. https://doi.org/10.1109/ACCESS.2020.2965621

Strano, E., Viana, M., Costa, L.daF., Cardillo, A., Porta, S., & Latora, V. (2013). Urban street networks, a comparative analysis of ten European cities. *Environment and Planning B: Planning and Design, 40*(6), 1071–1086. https://doi.org/10.1068/b38216

Strohbach, M. W., & Haase, D. (2012). Above-ground carbon storage by urban trees in Leipzig, Germany: Analysis of patterns in a European city. *Landscape and Urban Planning, 104*(1), 95–104. https://doi.org/10.1016/j.landurbplan.2011.10.001

Suleiman, A., Tight, M. R., & Quinn, A. D. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research, 10*(1), 134–144. https://doi.org/10.1016/j.apr.2018.07.001

Sun, Y., Gao, C., Li, J., Wang, R., & Liu, J. (2019). Quantifying the effects of urban form on land surface temperature in subtropical high-density urban areas using machine learning. *Remote Sensing, 11*(8). https://doi.org/10.3390/rs11080924

Taleqani, A., Hough, J., & Nygard, K. E. (2019). Public opinion on dockless bike sharing: A machine learning approach. *Transportation Research Record, 2673*(4), 195–204. https://doi.org/10.1177/0361198119838982

Tang, L., Xiong, C., & Zhang, L. (2018). Spatial transferability of neural network models in travel demand modeling. *Journal of Computing in Civil Engineering, 32*(3), Article 04018010. https://doi.org/10.1061/(asce)cp.1943-5487.0000752

Tchuente, D., & Nyawa, S. (2021). *Real Estate Price Estimation in French Cities Using Geocoding and Machine Learning.*

Tehrany, M. S., Simon, J., & Farzin, S. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena, 175*(April 2018), 174–192. https://doi.org/10.1016/j.catena.2018.12.011

Thomas, I., Frankhauser, P., Frenay, B., Verleysen, M., & Samos-Matisse, S. M. (2010). Clustering patterns of urban built-up areas with curves of fractal scaling behaviour. *Environment and Planning B: Planning and Design, 37*(5), 942–954. https://doi.org/10.1068/b36039

Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2019). Analyzing large-scale human mobility data: A survey of machine learning methods and applications. *Knowledge and Information Systems, 58*(3), 501–523. https://doi.org/10.1007/s10115-018-1186-x

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1–8*. https://doi.org/10.1145/2346496.2346498

Torija, A. J., & Ruiz, D. P. (2015). A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods. *Science of the Total Environment, 505*, 680–693. https://doi.org/10.1016/j.scitotenv.2014.08.060

Truong, T. M. T., Ly, H.-B., Lee, D., Pham, B. T., & Derrible, S. (2021). Analyzing travel behavior in Hanoi using support vector machine. *Transportation Planning and Technology*, 1–17.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications, 11*(1), 1–10. https://doi.org/10.1038/s41467-019-14108-y

Walks, R. A., & Maaranen, R. (2008). *Gentrification, Social Mix, and Social Polarization: Testing the Linkages in Large Canadian Cities, 29*.

Wang, N., Guo, G., Wang, B., & Wang, C. (2020). Traffic clustering algorithm of urban data brain based on a hybrid-augmented architecture of quantum annealing and brain-inspired cognitive computing. *Tsinghua Science and Technology, 25*(6), 813–825. https://doi.org/10.26599/TST.2020.9010007

Wang, Q., Phillips, N. E., Small, M. L., & Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences of the United States of America, 115*(30), 7735–7740. https://doi.org/10.1073/pnas.1802537115

Wang, Q., & Zhang, Z. (2017). Examining social inequalities in urban public leisure spaces provision using principal component analysis. *Quality and Quantity, 51*(6), 2409–2420. https://doi.org/10.1007/s11135-016-0396-0

Wang, W., Xia, F., Nie, H., Chen, Z., Gong, Z., Kong, X., & Wei, W. (2021). Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems, 22*(6), 3567–3576.

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering, 14*(10), 1402–1411. https://doi.org/10.1080/15732479.2018.1443145

Wójcik, P., & Andruszek, K. (2021). Predicting intra-urban well-being from space with nonlinear machine learning. *Regional Science Policy and Practice*.

Xiao, D., Heaney, C. E., Mottet, L., Fang, F., Lin, W., Navon, I. M., Guo, Y., Matar, O. K., Robins, A. G., & Pain, C. C. (2019). A reduced order model for turbulent flows in the urban environment using machine learning. *Building and Environment, 148*(October 2018), 323–337. https://doi.org/10.1016/j.buildenv.2018.10.035

Xie, R., Zhang, H., & Gong, C. (2018). Hot time periods discovery for facility proportioning in urban commercial districts using POIs and mobile phone data. *Web Intelligence, 16*(2), 91–104. https://doi.org/10.3233/WEB-180375

Xu, F., Ho, H. C., Chi, G., & Wang, Z. (2019). Abandoned rural residential land: using machine learning techniques to identify rural residential land vulnerable to be abandoned in mountainous areas. *Habitat International, 84*(October 2018), 43–56. https://doi.org/10.1016/j.habitatint.2018.12.006

Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J., & Wu, H. (2016). Demand driven store site selection via multiple spatial-temporal data. In *, 1. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. https://doi.org/10.1145/2996913.2996996

Xue, C., Ju, Y., Li, S., Zhou, Q., & Liu, Q. (2020). "Research on accurate house price analysis by using GIS technology and transport accessibility: A case study of Xi'an, China." MDPI Simmetry.

Yan, B., Mai, G., Janowicz, K., & Gao, S. (2017). From ITDL to Place2Vec – reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems 2017-Novem*. https://doi.org/10.1145/3139958.3140054

Yang, C., Clarke, K., Shekhar, S., & Vincent Tao, C. (2020). Big spatiotemporal data analytics: A research and innovation frontier. *International Journal of Geographical Information Science, 34*(6), 1075–1088.

Yang, D., Heaney, T., Tonon, A., Wang, L., & Cudré-Mauroux, P. (2018). CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion. *World Wide Web, 21*(5), 1323–1347. https://doi.org/10.1007/s11280-017-0515-4

Yang, Y., Tang, J., Luo, H., & Law, R. (2015). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management, 47*, 14–24. https://doi.org/10.1016/j.ijhm.2015.02.008

Yao, H.. (2012). Simulating the total ecological footprint of Suzhou from 1990 to 2009 by Bpann. *Polish Journal of Environmental Studies, 21*(6), 1901–1910.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*(4), 825–848. https://doi.org/10.1080/13658816.2016.1244608

Yi, F., Yu, Z., Zhuang, F., Zhang, X., & Xiong, H. (2018). An integrated model for crime prediction using temporal and spatial factors. In *Proceedings - IEEE International Conference on Data Mining, ICDM 2018-Novem:1386–91*. https://doi.org/10.1109/ICDM.2018.00190

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 186–94*. https://doi.org/10.1145/2339530.2339561

Zekić-Sušac, M., Mitrović, S., & Has, A. (2021). Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management, 58*. https://doi.org/10.1016/j.ijinfomgt.2020.102074

Zhai, W., Xueyin, B., Yu, S., Yu, H., Zhong, R. P., & Chaolin, G. (2019). Beyond Word2vec: an approach for urban functional region extraction and identification by combining Place2vec and POIs. *Computers, Environment and Urban Systems, 74* (August 2018), 1–12. https://doi.org/10.1016/j.compenvurbsys.2018.11.008

Zhang, W., Robinson, C., Guhathakurta, S., Garikapati, V. M., Dilkina, B., Brown, M. A., & Pendyala, R. M. (2018). Estimating residential energy consumption in metropolitan areas: A microsimulation approach. *Energy, 155*, 162–173. https://doi.org/10.1016/j.energy.2018.04.161

Zhang, X., Yan, F., Liu, H., & Qiao, Z. (2021). Towards low carbon cities: a machine learning method for predicting urban blocks carbon emissions (UBCE) based on built environment factors (BEF) in Changxing City, China. *Sustainable Cities and Society, 69* (March), Article 102875. https://doi.org/10.1016/j.scs.2021.102875

Zhang, X., Li, W., Zhang, F., Liu, R., & Du, Z. (2018). Identifying urban functional zones using public bicycle rental records and point-of-interest data. *ISPRS International Journal of Geo-Information, 7*(12). https://doi.org/10.3390/ijgi7120459

Zhao, G., Pang, B., Xu, Z., Peng, D., & Xu, L. (2019). Assessment of urban flood susceptibility using semi-supervised machine learning model. *Science of the Total Environment, 659*(December 2019), 940–949. https://doi.org/10.1016/j.scitotenv.2018.12.217

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A review. *ArXiv*, (december)https://doi.org/10.1109/MGRS.2017.2762307

Zhu, X., Li, J., Liu, Z., Wang, S., & Yang, F. (2016). Learning transportation annotated mobility profiles from GPS data for context-aware mobile services. In *Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016* (pp. 475–482). https://doi.org/10.1109/SCC.2016.68

Zubair, O. A., Wei, J., & Weilert, T. E. (2017). Modeling the impact of urban landscape change on urban wetlands using similarityweighted instance-based machine learning and Markov model. *Sustainability (Switzerland), 9*(12). https://doi.org/10.3390/su9122223