The Big Data Deluge for Transforming the Knowledge of Smart Sustainable Cities: A Data Mining Framework for Urban Analytics*

Simon Elias Bibri
The Norwegian University of Science and Technology (NTNU), Trondheim, Norway
simoe@ntnu.no

John Krogstie
The Norwegian University of Science and Technology (NTNU), Trondheim, Norway
john.krogstie@ntnu.no

ABSTRACT

There has recently been much enthusiasm about the possibilities created by the big data deluge to better understand, monitor, analyze, and plan modern cities to improve their contribution to the goals of sustainable development. Indeed, much of our knowledge of urban sustainability has been gleaned from studies that are characterized by data scarcity. Therefore, this paper endeavors to develop a systematic framework for urban sustainability analytics based on a cross–industry standard process for data mining. The intention is to enable well–informed decision–making and enhanced insights in relation to diverse urban domains. We argue that there is tremendous potential to transform and advance the knowledge of smart sustainable cities through the creation of a big data deluge that seeks to provide much more sophisticated, wider–scale, finer–grained, real–time understanding, and control of various aspects of urbanity in the undoubtedly upcoming Exabyte Age.

CCS CONCEPTS

• **Information Systems → Information Systems Applications**; Data Mining, Decision support systems Data Analytics

KEYWORDS

Smart sustainable cities, big data analytics, data mining

ACM Reference format:

S.E. Bibri and John Krogstie 2018 The Big Data Deluge for Transforming the Knowledge of Smart Sustainable Cities: A Data Mining Framework for Urban Analytics In Proceedings of the 3rd International Conference on Smart City Applications Tetouan, Marocco, October 2018 (SCA 2018) , 9 pages.
DOI: 10.1145/123 4

1 INTRODUCTION

The amount of data in the world is soaring, amounting to hundreds of extra exabytes every year. An urban data deluge results from the increasing availability of the data being generated in continuous streams and on daily basis in the urban environment from heterogeneous sources [1]. It is estimated that more data are being produced every two days at present than in all of history prior to 2003 [2]. Such explosive growth in data is due to a number of different enabling and driving technologies, infrastructures, techniques, and processes as new developments in ICT, and their rapid embedding into everyday practices and spaces, enabling the accessing and sharing of data, in addition to the means by which much big data are generated [2]. These data are considered to be the most scalable and synergic asset for smart sustainable cities. The role of big data processing technologies lies in collecting, storing, processing, analyzing, and interpreting large masses of data on every urban system and domain to discover useful knowledge and employ it to enhance decision–making and insights. The value of this knowledge lies in improving physical forms, infrastructures, resources, networks, facilities, and services by developing urban intelligence functions for automating and supporting decisions pertaining to control, optimization, management, and (short–term and long–term) planning for the purpose of improving the contribution of smart sustainable cities to the goals of sustainable development. However, merely keeping up with data flood, and storing the bits that might be useful, is challenging enough, not to mention analyzing datasets to spot patterns and extract useful knowledge. Even so, the data deluge is already starting to transform the knowledge of smart sustainable cities. It has great potential for good—as long as city governments make the right choices about when to restrict the flow of data, and when to encourage it.

Data mining is one of the techniques needed for creating value from big data. Indeed, it is gaining a strong footing in the urban domains, and presents a tremendous challenge due to the interdisciplinary and transdisciplinary nature of urban data [1]. However, over the last 20 years or so, research within the area of data mining has been mainly in other domains. Accordingly, there is a paucity of research on this topic in the domain of urban sustainability. In particular, while big data analytics has recently become of focus in the context of smart cities ([3], [4], [5], [6], [2]) as well as smart sustainable cities of the future ([1], [7], [8]), research on the next wave of urban analytics based on data mining remains scant, while in city–related research, it is argued that "small data" studies—surveys, focus groups, case studies, participatory observations, audits, interviews, content analyses, and ethnographies—are associated with high cost, infrequent periodicity, quick obsolescence, bias, and inaccuracy.

This paper is about thinking data–analytically about urban sustainability problems to transform and advance the knowledge of smart sustainable cities, which is aided by a conceptual framework with well–defined stages to help structure urban data–analytic thinking. Specifically, it develops, illustrate, and discuss a systematic framework for urban sustainability analytics based on a cross–industry standard process for data mining. The intention is to enable well–informed or knowledge–driven decision–making and enhanced insights in relation to diverse urban domains with regard to operations, functions, strategies, designs, and policies for the purpose of enhancing the contribution of smart sustainable cities to sustainability.

The remainder of this paper is structured as follows. Section 2 introduces and describes the relevant conceptual and theoretical constructs. Section 3 explains how urban sustainability problems can be transformed into data mining tasks. Section 4 presents the proposed data mining framework for urban analytics, with a particular focus on data–analytic solutions to urban sustainability problems. This paper ends, in Section 5, with concluding remarks, reflections, and future research avenues.

2 CONCEPTUAL AND THEORETICAL BACKGROUND

2.1 Smart Sustainable Cities

As echoed by Höjer and Wangel [10], the interlinked development of sustainability, urbanization, and ICT has recently converged under what is labelled "smart sustainable cities" which is a new techno–urban phenomenon that materialized around the mid–2010s ([1], [7], [9], [10]). As an integrated framework, they amalgamate the strengths of sustainable cities in terms of the design concepts and planning principles of sustainability and those of smart cities in terms of the innovative ICT-solutions being developed in cities [1]. The development of smart sustainable cities is gaining increasing attention and traction across the globe among research institutes, universities, governments, policymakers, and ICT companies as a promising response to the imminent challenges of sustainability and urbanization.

The term "smart sustainable city," despite not always explicitly discussed, is used to describe a city that is supported by the pervasive presence and massive use of advanced ICT, which, in connection with various urban systems and domains and how these intricately interrelate and are coordinated, enables the city to control available resources safely, sustainably, and efficiently to improve economic and societal outcomes [9]. Höjer and Wangel [10] define a smart sustainable city as "a city that meets the needs of its present inhabitants without compromising the ability for other people or future generations to meet their needs, and thus, does not exceed local or planetary environmental limitations, and where this is supported by ICT". This entails primarily unlocking and exploiting the potential of ICT of pervasive computing as an enabling, integrative, and constitutive technology with embodied transformational, substantive, and disruptive effects for achieving the environmental, social and economic goals of sustainability.

2.2 Big Data Analytics

The term "big data" essentially denotes datasets that are too large and complex for conventional data processing systems. This implies that the existing computing models and practices remain unfit for handling big data. Big data are often characterized by a number of Vs the main of which are volume, variety, and velocity. Additional Vs proposed include veracity, validity, value, volatility, and variability [2]. While there is no canonical or definitive definition of big data in the context of smart sustainable cities, the term can be used to describe a colossal amount of urban data, typically to the extent that their manipulation, analysis, management, and communication present significant computational, analytical, logistical, and coordinative challenges. Important to note is that such data are invariably tagged with spatial and temporal labels, largely streamed from various forms of sensors, and mostly generated automatically and routinely.

The term "big data analytics' refers commonly to any vast amount of data that has the potential to be collected, stored, retrieved, integrated, selected, preprocessed, transformed, analyzed, and interpreted for discovering new or extracting useful knowledge. This can subsequently be evaluated and visualized in an understandable form prior to its deployment for decision–making purposes. In the context of smart sustainable cities, big data analytics refers to a collection of dedicated software applications and database systems run by machines with very high processing power, which can turn a large amount of urban data into useful knowledge for well–informed decision–making and enhanced insights in relation to various urban domains " [1].

The common types of big data analytics include predictive, descriptive, diagnostic, and prescriptive analytics. These are applied to extract different types of knowledge or insights from large datasets. Urban analytics involves the application of various techniques based on data science fundamental concepts, including data mining, machine learning, statistical analysis, regression analysis, database querying, data warehousing, or a combination of these. The use of these techniques depends on the urban domain as well as the nature of the urban problem to be tackled. The main difference between data mining and other analytics techniques is that the former focuses on the automated search for or extraction of useful knowledge from data. This paper puts emphasis on predictive and descriptive data mining.

2.3 Data Mining and its Process

Data mining is the computational process of probing large datasets in order to find frequent, but hidden patterns; to make useful and valid correlations from these discoveries; and to summarize the results in novel ways and then visualize them in understandable formats prior to their deployment for decision–making purposes. Among the data mining models used to perform data processing and analysis functions include distributed data mining, multi–layer mining, data mining from multi–technology integration, and grid-based mining [11]. There are a variety of data mining algorithms that can be used to solve problems pertaining to urban sustainability, including classification, clustering, regression, profiling, similarity matching, causal modeling, predictive link, and co–occurrence grouping [1].

According to several codifications of the process of data mining [1], this process consists of well–defined stages, namely problem understanding, data understanding, data preparation, model building, result evaluation, and result deployment. In the context of smart sustainable cities, the process of data mining targets optimization and intelligent decision support pertaining to the control, efficiency, management, and planning of urban systems, as well as to the enhancement of the associated ecosystem and human services related to energy, water, healthcare, education, safety, and so on [1]. Additionally, the process targets the improvement of practices, strategies, and policies by changing them based on new trends. In all, the analytical outcomes of data mining can serve to improve urban operational functioning, optimize resource utilization, reduce environmental risks, enhance the quality of life and well–being of citizens, and streamline planning and governance processes.

There is a large body of work that addresses cross–industry standard process for data mining ([12], [13]). This process emphasizes the idea of iteration. This implies that solving a particular problem may require going through the process more than once. The data mining process model exists in many variations, e.g., a simplified process like (1) Pre-processing, (2) Data Mining, and (3) Results Validation. The literature shows that the CRISP-DM methodology is the leading methodology used by data miners.

3 FROM URBAN SUSTAINABILITY PROBLEMS TO DATA MINING TASKS

Each data–driven decision–making problem as part of urban analytics is unique, consisting of its own combination of objectives, requirements, and constraints. However, there are sets of common data mining tasks that underlie the urban sustainability problems pertaining to various urban domains and how they may be interrelated or coordinated. In collaboration with urban stakeholders, data scientists decompose an urban sustainability problem into subtasks, and the solution can subsequently be composed to solve such problem. This can relate to energy, transport, and mobility, traffic, built environment, healthcare, education, safety, or other urban (application) domains. The know–how of data scientists resides in their ability to decompose a data analytics problem of a particular aspect of urban sustainability into subtasks for which tools and techniques are available and can be used separately or combined. Some of these subtasks remain common data mining tasks, and others are unique to the particular context of an urban sustainability problem.

In recent years, there has been a major shift in knowledge used to solve the common data mining tasks. Overall, what matters in data analytics problems in the context of urban sustainability is to possess the ability of recognizing urgent and common problems and their solutions, and doing this in ways that avoid wasting resources and time by reinventing the wheel when considering new urban projects associated with data analytics for advancing urban sustainability with regard to various urban domains. This implies that the data mining process in this context is not only about the automated extraction of useful knowledge from data for enhanced decision–making and insights, but also about creativity, common sense, acumen, expert knowledge, and so on.

Data mining focuses on the automated search for extracting useful knowledge through finding patterns, regularities, and correlations in data. And it is important for the urban analysts to be able to recognize what sort of the analytic techniques among the available ones is appropriate for addressing a particular urban sustainability problem within a given urban domain. Although there are a large number and variety of specific data mining algorithms (from such fields as machine learning, statistics, artificial intelligence, database systems, and pattern recognition) to perform different data analysis tasks, there are only a small amount of fundamentally different kinds of tasks these algorithms perform. The tasks apply to, in the context of smart sustainable cities, different kinds of human or inanimate entities (e.g., citizen, mobility system, utility system, traffic system, transport system, energy system, travel behavior, typology, etc.) about which we have data. In many urban sustainability analytics projects, the desire is to find correlations between one or more particular variables describing an entity and other variables.

The data mining tasks relate to different analytics methods, including descriptive (what happened?), diagnostic (why did it happen?), predictive (what will happen?), and prescriptive (what should be done?) used to solve different decision–making problems related to urban sustainability in terms of urban operations, functions, services, designs, strategies, practices, and policies [1]. The first three types of analytics are concerned with decision–making and its support, which entails human intervention, the level of which would vary depending on the nature and complexity of the application in connection with various urban domains. The last one is associated with decision automation and some kind of decision support. The targets of decision–making and action–taking are associated with the operating and organizing processes of urban life in line with the goals of sustainable development.

4. A DATA MINING FRAMEWORK FOR URBAN ANALYTICS

This section presents some of the fundamental principles of data science underlying the common types of data mining tasks based on Provost and Fawcett [14].

Next, a systematic framework for urban analytics studies is presented (see Figure 1), which places a structure on the problems pertaining to urban sustainability, allowing reasonable consistency, repeatability, and objectiveness. The derivation of this data mining framework is based on cross–industry standard process for data mining based on several sources as referenced in Bibri [1]. The deep technical details of the components of this data mining framework and thus how they are linked to urban domains in relation to sustainability dimensions is beyond the scope of this paper.

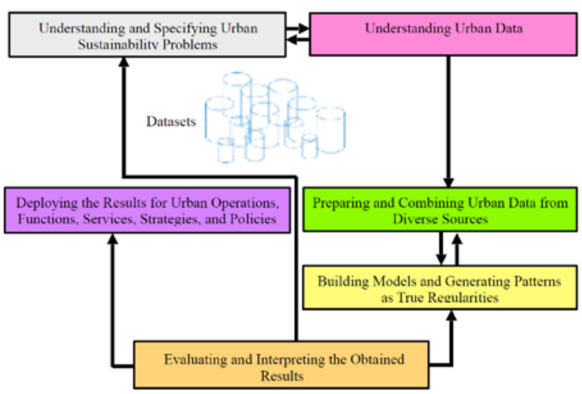

Figure 1: A data mining framework for urban analytics

Applicable to the domains of smart sustainable cities, this framework emphasizes the idea of iteration as the case for all existing codifications of the process of data mining. The entire process is an exploration of the urban data and how it can be integrated and harnessed. The rationale for the iteration is for the group of data scientists in collaboration with urban researchers, scholars, and planners to increase their understanding and thus gain more knowledge as they explore the problem and devise the right solution to it. The lessons learned during the process can trigger new, often more focused and fine–grained sustainability questions, and subsequent data mining procedures will benefit from the experiences of previous ones. In addition, the sequence of the components is not strict, and moving back and forth between different components is always required. The arrows indicate the most important and frequent dependencies between the components. A data mining process continues after a solution has been deployed. The components are discussed next in more detail in relation to urban sustainability.

4.1 Understanding and Specifying Urban Sustainability Problems

Initially, it is crucial to understand the urban sustainability problem that is to be tackled. This is not an evident or clear–cut thing per see. Sustainability endeavors within diverse urban domains and how they relate to physical, environmental, social, and economic dimensions of smart sustainable cities hardly ever come pre–packaged as unambiguous or straightforward data mining problems. Usually, recasting the problem associated with one or more dimensions of urban sustainability and devising an acceptable solution is an iterative process of discovery taking the form of cycles within a cycle. The initial formulation of any urban sustainability problem is unlikely to be complete, thereby the need and relevance for multiple iterations to achieve an adequate, or ideally optimal, problem formulation. The understanding stage of the urban sustainability problem represents part of competence and skillfulness where the urban analysts or specialized data scientists' creative and innovative ideas become determining with regard to how to cast the urban sustainability problem as a set of data science problems. In the domain of urban sustainability, both specialized and interdisciplinary knowledge help urban analysts come up with novel problem and solution formulations.

Typically, the early stages of any problem in this regard entail designing a solution that takes advantage of data mining techniques. This can signify engineering the urban problem in ways that consist of one or more sub-problems that involve building clusters or constructing models for classification, clustering, probability estimation, regression, or causal modelling. To understand the urban sustainability problem as a first stage, the urban analysts who are in charge of structuring the problem should think carefully about the use scenario. Provost and Fawcett [14] devote two entire chapters (Chapter 7 and 11) to this concept, which is one of the most important concepts of data science. The understanding stage involves understanding the project objectives and requirements from an urban sustainability perspective, converting this knowledge into a data mining problem definition, and creating a preliminary plan to achieve the objectives. That is to say, based on the concept of use scenario, what exactly we want to carry out, how exactly we would carry it out, what aspects of use scenario constitute possible models of data mining, and which kind of models are most relevant. As to the latter case, the way forward is to begin with a simplified view of the use scenario (e.g., traffic light control, energy demand management, car sharing implementation, etc.). As the process will evolve, we will loop back and adjust the use scenario to better reflect the actual urban need. This entails framing the urban sustainability problem in ways that can allow us to systematically and effectively decompose it into data mining tasks. In relation to urban sustainability use scenarios, several environmental, social, and economic indicators could be extracted from historical urban data, assembled into predictive models and then deployed in the operating and organizing urban processes and activities as part of data–centric applications spanning a variety of urban domains in relation to the operational functioning, management, and planning of smart sustainable cities. In this regard, the predictive model constituting an aspect of a particular use scenario abstracts away most of the complexity of the urban world, focusing on a specific set of environmental, social, or economic indicators that correlate in some way with the value of the target variable to be predicted. For example, an environmental urban sustainability problem would be to find out if a group of dwellers will be environmentally sustainable. A subtask of data mining that will likely be part of the solution to such problem is to estimate from historical data the probability of dwellers being environmentally sustainable in a given district characterized by certain typologies (e.g., density, diversity, mixed land–use, etc.) on the basis of a set of descriptive attributes entailing different environmental and spatial indicators.

4.2 Understanding Urban Data

In this stage, the focus is on matching the problem with the data, a process that involves understanding the strengths and weaknesses of the available data. In this respect, the solution to the urban sustainability problem as a goal is to be built from the available raw material. Historical urban data often are collected and stored for purposes unrelated to the current urban problem. Different databases are available across diverse urban domains and owned by different urban entities, covering different information on citizens, transactions, movements, observations, and so on, and may have varying degrees of reliability, different formats and costs. In the urban domain, some data are open and thus freely accessible to the public while other data are confidential. Also, some data are available virtually for free while other data require effort to obtain or even need to be acquired. Still not all the data needed for building solutions to a given urban sustainability problem exist. Hence, some data are likely to necessitate entire ancillary projects to arrange their collection and storage.

In the context of smart sustainable cities, it is necessary to put in place a cross–service domain system to ensure that access to the data from different urban domains is available at all times [1]. Prior to this, it is important to ensure that the diverse kinds of datasets pertaining to the areas associated with urban sustainability are open for use by the city constituents with respect to data–driven applications in relation to operations, functions, services, designs, strategies, and policies.

In light of the above, data warehousing is of critical importance in terms of urban data understanding. It serves to collect and coalesce data from across urban systems involving multiple urban entities, each with a set of its own databases. Residing usually at a single site, a data warehouse is a massive repository of information collected from multiple sources, but stored under a unified schema. In this sense, data warehousing can be regarded as a facilitating technology of data mining. Smart sustainable cities rely on data warehouses or storage facilities so that they can apply data mining more broadly and deeply. For example, if a data warehouse integrates records from travel behavior, energy consumption as well as traffic system and mobility patterns, it can be used to find effective typologies and design concepts in urban planning.

A critical part of the data understanding stage in the context of smart sustainable cities is estimating the costs and benefits of data sources and data repositories, and deciding whether further effort is merited. In relation to data warehousing and processing, urban data do require additional effort to be collated after all datasets are acquired and accessible to be subsequently mined. As data understanding progresses in relation to attempting to solve various urban sustainability problems, solution paths may change direction and new insights come into play in response, and the efforts of urban analysts may even fork, as sometimes one problem may have different solutions, or two problems of the same concern may be categorized significantly different in terms of which techniques are more suitable. It is important, when attempting to understand data, to dig beneath the surface to uncover the structure of the urban sustainability problem in terms of what analytical tasks are of more relevance to solving that problem, as well as in terms of the data that are available. Afterwards, we come to match these data to relevant data mining tasks for which there are substantial scientific and technological methods and systems to apply. It is not unusual for an urban sustainability problem, whether be it physical, environmental, economic, or social, to contain several data mining tasks, often of different types, whose solutions should be integrated for effective outcomes (see Chapter 11 from Provost and Fawcett [14] for more detail).

In sum, urban data understanding encompasses the following steps:

- Creating familiarity with the data
- Data collection and warehousing
- Identification of data quality problems
- Discovery of first insights
- Detection of interesting subsets
- Formation of hypotheses for hidden information

4.3 Preparing and Combining Urban Data from Diverse Sources

The analytic techniques to bring to bear as to solving urban sustainability problems impose certain requirements on the data they employ. They require the data be in a certain form, which often is different from how the data are provided originally. Therefore, some conversion of the data into suitable forms is necessary. Typical examples of data preparation in the urban domain entail coordinating data from different sectors by bringing the different elements of urban entities or complex activities into a relationship that will ensure efficiency or harmony, converting data to tabular format, combining tables with similar or shared attributes, inferring or completing missing values, removing repetitive values, and converting data to achieve various types for consistency. Concerning the latter, some data mining techniques are, as discussed above, designed for categorical data (e.g., for classification) while others deal only with numerical values (e.g., for regression). These must often be normalized so that they become computationally comparable and yield the desired results. The interested reader can be directed to Chapter 3 from Provost and Fawcett [14] for a detailed discussion of the most typical format and conversion of mining data.

A related aspect to conversion, which pertains to the early stages of the automatic processing of the data, is to provide guidelines on how all the different cross–thematic data categories can be integrated. More generally, urban analysts team may spend considerable time early in the process defining and crafting the variables used later in the process as part of structuring the urban sustainability problem, an aspect that has a lot to do with expert knowledge as well as creativity and common sense. Overall, data preparation entails constructing the final dataset to be fed into the data mining algorithms.

4.4 Building Models and Generating Patterns as Stable Regularities

Generally speaking, a model is an (over)simplified view of the real world created to serve a purpose. In the context of smart sustainable cities, this oversimplification is based on assumptions about what is and is not important for the purpose of advancing their contribution to the goals of sustainable development. The output of modeling in the process of data mining is some sort of patterns capturing stable regularities in the urban data. These data pertain to different urban domains. It is during modeling when data mining techniques are applied to the data through building models from both historical and real–time data depending on the urban sustainability problem that is to be solved. In this respect, it is crucial to have some understanding of the sorts of available techniques and algorithms. In this stage, the urban analysts ensure that appropriate data mining techniques and algorithms are selected and applied in relevance to the urban sustainability problem, and parameters are learned. Also, some methods may have specific requirements on the form of input data, and therefore going back to the data preparation stage may be needed.

There are mainly two types of modeling in data mining based on supervised and unsupervised learning methods: predictive data mining and descriptive data mining. Predictive data mining generates models described by particular urban data and uses some variables, with important, informative gains, in the dataset to predict future or unknown values of target variables. Speaking of informative gains and in relation to supervised segmentation, finding important, informative variables in the dataset pertaining to the entities described by the data is one of the fundamental ideas of data mining means the variables that are most predictive of the target, or alternatively, have the best correlation with the target. While for variables to be informative usually varies among applications, underlying informativeness is the idea that information generally represents a quantity that reduces uncertainty about something in the sense that the better the information provided, the more the uncertainty is reduced by that information. Having a target variable crystalizes the idea of finding informative attributes with regard to identifying whether there is one or more other variables that reduces the uncertainty about the value of the target or in it. Determining these correlated variables is intended to provide key insights into the urban sustainability problem on focus.

Prediction signifies forecasting a future event. A predictive model is a formula (mathematical or logical statement, but usually a combination of the two) for estimating the future or unknown value of the target. In the urban domain, data mining may deal with historical or real–time data, and hence models are built and tested using events from both the past and present. We can think of many urban questions involving predictive modeling, such as how can we segment the citizens with respect to mobility as something that we would like to predict or estimate in relation to a particular spatial scale (e.g., district, city, and region). The target of this prediction can be something we would like to relate to environmental sustainability or social sustainability, such as which individual movements are likely to increase emissions and which ones are likely to be influenced by existing typologies and design concepts on a particular spatial scale, or which individual movements are likely to be associated with enhanced spatial accessibility to services and facilities as an aspect of the quality of life. To extract patterns from data (useful knowledge pertaining to mobility) in a supervised manner entails segmenting the citizens dwelling in a particular district into subgroups with instances of similar values for the target variable as part of sub-groups that have different values for the target variable. In this case, the segmentation is performed using values of variables that are known when the target variable is not in order to predict its value accordingly. In addition, the segmentation may concurrently provide a human–understandable set of segmentation patterns. An example of a segment expressed in English would be: "people who live in density and mixed land–use oriented areas and prefer walking and cycling on average have an emission rate of 5%."

As regards to descriptive data mining, it produces new, non-trivial characteristic or grouping information based on the available urban dataset while focusing on finding patterns for human interpretation. In descriptive modeling, the primary purpose of the model is to gain meaningful insights into the underlying phenomenon of urban sustainability. A descriptive model of citizen travel behavior or mobility mode would tell us what citizens who use sustainable transport or cycling and walking typically look like. A descriptive model must be judged in part on its intelligibility and easiness of understanding for an effective deployment in relation to urban services, designs, strategies, or policies, in contrast to a predictive model which may be assessed solely on the basis of its predictive performance. As some of the same techniques and algorithms can be used for both descriptive and predictive modeling, the difference between supervised and unsupervised data mining models is not as strict.

Below are some examples of predictive and descriptive data mining in relation to smart sustainable cities taken from Bibri [1]:

Predictive questions:

- Classify citizen travel behavior
- Classify household energy consumption
- Predict how GHG emissions will develop in the next month or next year
- Predict areas of dense traffic in the near future
- Predict travel behavior related to particular typologies at a particular spatial scale
- Predict spatiotemporally the development and propagation of congestion

Descriptive questions:

- Find useful travel behavior categories
- Find interesting collective or individual mobility patterns
- Find characteristic information about traffic jams and road congestion
- Describe normal accessibility to facilities in urban areas characterized by mixed–land use
- Find groups of typologies that share similar features of environmental performance
- Find groups of citizens that share similar travel behavior patterns
- Find association rules between mobility or commuting behavior and environmental performance
- Discover the subgroups of travel characterized by common behavior, time length, and purpose

4.5 Evaluating and Interpreting the Obtained Results

After building the desired models (or patterns capturing regularities in the data), it is important to assess the data mining results and to gain confidence that they are valid and reliable before moving on in the process. This is what the evaluation stage is about. It is desirable to have confidence that the generated models represent stable regularities and not just sample anomalies, odds, or idiosyncrasies. The underlying assumption is that an urban analyst can always find patterns from any dataset, but these patterns may not survive careful scrutiny. This relates to one of the fundamental concepts of data science: overfitting. Moreover, it is inadvisable to deploy the results of data mining immediately as to their use for decision–making purposes. Prior to their deployment, results have to be evaluated in ways that ensure the generated models satisfy the original urban goals in terms of supporting decision–making. This involves finding a data–analytic solution to the urban sustainability problem being explored or investigated. A data mining solution often is only a piece of the larger solution to the urban sustainability problem in question. And it needs to be evaluated as such. There are different external factors that should be taken into account when evaluating models, which might make them impractical, although they pass strict evaluation tests in a controlled laboratory setting.

The complexity and scale of urban sustainability projects imply that they involve various urban stakeholders that have interests in the urban sustainability decision–making that will be supported by the resultant models. This signifies that the results of data mining is to be evaluated using both qualitative and quantitative metrics. However, the stakeholders need to be satisfied with the outcome of the evaluation with regard to the quality of the models' decisions in order to "sign off" on the deployment of the resultant models. Irrespective of the domain application of urban sustainability in this context, the basic idea guiding this quality is to ensure the model is unlikely to make mistakes in directing urban operations, functions, services, designs, strategies, practices, and polices. To facilitate this kind of qualitative evaluation, the urban analyst must think about the comprehensibility of the model to urban stakeholders, and accordingly attempts to find ways to render the behavior of the model comprehensive, if the latter happens to be not so due to some complex mathematical formula, for example. The rationale for performing the evaluation early on to provide a comprehensive assessment framework before its use (production) is that it may be difficult to obtain detailed information on the performance of a deployed model due to the limited access to the production environment where such model is applied. Adding to this is that the model may be deployed as part of decision–making systems related to diverse urban domains. In addition, deployed systems typically contain many parts that they tend to move around in the underlying system, and assessing the contribution of a single part of the system remains difficult. To obtain the most realistic evaluations before taking the risk of deploying the resultant models across urban systems, it is recommended to build testbed environments that can reflect production data as closely as possible. In sum, the evaluation stage requires current model have a high quality from a data mining perspective, and before final deployment, it is important to test whether the model achieves all urban sustainability project objectives.

4.6 Deploying the Results for Urban Operations, Functions, Services, Strategies, and Policies

In the deployment stage, the results of data mining are put into real use in terms of making, supporting, or automating different kinds of decisions associated with urban operational functioning, planning, and service delivery. The clearest cases of deployment in this regard entail implementing predictive or descriptive models in the processes operating and organizing urban life or information systems associated with public services as part of decision support systems across various urban entities. In this context, a model for predicting or describing travel behavior could, for example, be used in public transport system engineering or management. In relation to this, travel "data are potentially extremely useful for figuring out disruptions on the (transport) system. We do need, however, to generate some clever cognitive analyses of how people make their way through the various transport systems, just as we need to assign travelers to different lines to ensure that we can measure the correct number of travelers on each line… The state–of–the–art in what we know about navigation in complex environments is still fairly primitive. Many assumptions have to be made and we have no data on what different users of the system have actually learned about their routes. New users of the system will behave differently from seasoned users and this introduces further error. We can see disruption in the data by determining the times at which travelers enter and exit the system, but to really predict disruption on individual lines and in stations, we need to match this demand data to the supply of vehicles and trains that comprise the system" [4]. Many other kinds of models can be built into environmental systems, energy systems, water distribution systems, communication systems, building systems, traffic systems, and a wide variety of information systems to increase the contribution of smart sustainable cities to the goals of sustainable development in terms of environmental regeneration, economic efficiency, and social equity and well–being.

Deployment can also be much less technical when a set of rules discovered by data mining techniques could help to quickly diagnose and fix a common error in some systems (e.g., typologies, design concepts, bicycle or car sharing approaches, etc.). In this case, the deployment can be in the form of disseminating new practices containing the rules or principles in question. Deployment can moreover be much more subtle, such as a change to operations, functions, services, and strategies resulting from insights gained from mining the urban data.

Results deployment often returns to the urban sustainability problem understanding phase, irrespective of whether it is successful. This is illustrated in Figure 1: the link from results evaluation back to urban sustainability problem understanding.

To sum up, it is worth pointing out that in urban analytics, it is critically important for urban analysts to be able to formulate urban sustainability problems well in relevance to each urban domain, to prototype (analytical) solutions quickly, to make realistic assumptions in the face of ill–defined and–structured problems, to design scientific procedures for making meaningful discoveries, and to analyze results. In light of this, new partnerships and alliances among different urban entities are necessary for the use of big data analytics in the context of smart sustainable cities, especially city authorities are likely to lack data scientists and hence must borrow them from academic institutions and industrial organizations.

5 CONCLUSIONS

We stand at a threshold in beginning to make sense of big data analytics and data–driven decision–making and related processes, systems, and methods that will be of massive use in, and interwoven into the very fabric of, smart sustainable cities of the future as complex and dynamic systems. With big data analytics, and particularly the use and combination of different tasks as part of the data mining processes to handle and solve various decision–making problems pertaining to urban sustainability, we will be able to better monitor, understand, and analyze smart sustainable cities so as to be more intelligently operated, managed, planned, developed, and governed in terms of improving and maintaining their contribution to sustainability.

This paper intended to develop, illustrate, and discuss a systematic framework for urban sustainability analytics based on cross–industry standard process for data mining in response to the emerging wave of city analytics in the context of smart sustainable cities. The proposed data mining framework for urban analytics consists of six components, namely:

1. Understanding and specifying urban sustainability problems
2. Understanding urban data
3. Preparing and combining urban data from diverse sources
4. Building models and generating patterns as stable regularities
5. Evaluating and interpreting the obtained results
6. Deploying the results for urban operations, services, strategies, and policies

The prominence of this framework as a set of conceptual tools lies in the value of the related well–defined stages in enhancing the likelihood of successful and relevant results. This work can serve to bring together city analysts, data scientists, urban planners and scholars, and ICT experts on common ground in their endeavor to transform and advance the knowledge of smart sustainable cities.

The unique features of the proposed framework lie in its novelty in terms of extending its applicability to sustainability problems in the context and domain of smart sustainable cities. The proposed framework is based on a general perspective of data mining rather than on a perspective specific to each urban domain or even sub–domain.

For future work the focus will be on examining the key distinctive features of various urban domains from technical, computational, and analytical perspectives so as to design a set of frameworks tailored to each urban domain and, optimistically, to each sub–domain, while drawing on the general framework proposed in this work. Of equal importance is to develop effective evaluation methods for these frameworks.

REFERENCES

[1] S. E. Bibri. Smart sustainable cities of the future: the untapped potential of big data analytics and context aware computing for advancing sustainability, Springer, 2018.

[2] R. Kitchin, "The real-time city? Big data and smart urbanism," Geography Journal, 2014, vol.79, pp.1–14.

[3] E. Al Nuaimi, H. Al Neyadi, M. Nader and J. Al-Jaroodi, "Applications of big data to smart cities," Journal of Internet Services and Applications, 2015, vol. 6, no. 25, pp. 1–15.

[4] M. Batty, "Big data, smart cities and city planning," Dialogues Human Geography, 2013, vol. 3, no. 3, pp. 274–279.

[5] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, Y. Portugali, "Smart cities of the future," European Physical Journal, 2012, vol. 214, pp.481–518.

[6] Z. Khan, A. Anjum., K. Soomro and M. A. Tahir, "Towards cloud based big data analytics for smart future cities," Journal of Cloud Computing and Advanced System Applications, 2015, vol. 4, no. 2.

[7] S. E. Bibri, "The IoT for smart sustainable cities of the future: an analytical framework for sensor-based big data applications for environmental sustainability," Sustainable Cities and Society, 2018b, vol. 38, pp. 230–253.

[8] S. E. Bibri and J. Krogstie, "ICT of the new wave of computing for sustainable urban forms: their big data and context-aware augmented typologies and design concepts," Sustainable Cities and Society, 2017b, vol. 32, pp. 449–474.

[9] S. E. Bibri and J. Krogstie, "Smart sustainable cities of the future: an extensive interdisciplinary literature review," Sustainable Cities and Society, 2017a, vol. 31

[10] M. Höjer and S. Wangel S, "Smart sustainable cities: definition and challenges," in L. Hilty and B. Aebischer (eds), ICT innovations for sustainability, Springer, 2015, pp.333–349.

[11] S. Bin, L., Yuan and W. Xiaoyi, "Research on data mining models for the internet of things," in Proc. of the Int. Conference on Image Analysis and Signal Processing, 2010, p. 127–132.

[12] R. L. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," Knowledge Engineering Review, 2006, vol. 21, no. 1, pp. 1–24, Cambridge University Press, New York, NY, USA

[13] C. Shearer, "The CRISP–DM model: the new blueprint for data mining," Journal of Data Warehouse, 2000, vol. 5, no. 4, pp. 13–22.

[14] F. Provost and T. Fawcett, Data science for business, O'Reilly Media Inc, 2013.