Big Data Analytics in Smart Mobility: Modeling and Analysis of the Aarhus Smart City Dataset

Johannes Zenkert, Mareike Dornhöfer, Christian Weber, Charly Ngoukam and Madjid Fathi

University of Siegen
Institute of Knowledge Based Systems and Knowledge Management
Department of Electrical Engineering and Computer Science
Germany

Abstract—**A smart city is a modern and visionary approach for a city to provide intelligent and smart urban services by using information and communication technologies (ICT). The Internet of Things (IoT), emerging through intelligent networking and sensing technologies, is seen as the data-driven enabler for smart cities with current and future infrastructures. The Open Data Aarhus datasets have been created from sensor data in the city Aarhus in Denmark. This paper presents how big data technologies in the context of smart cities are used to implement a framework with a prototype R Shiny application to analyze road traffic and pollution data to make a step towards smart mobility. The main objective of the approach is the calculation and visualization of the least polluted route from a chosen start to an end point by applying an algorithm utilizing the MapReduce framework running on a Hadoop cluster.**

Keywords—**Smart City, Smart Mobility, Big Data Analytics, MapReduce**

I. Introduction

The growing rate of global urbanization is a central challenge of the 21st century. Urbanization has many negative consequences for cities' environmental pollution, transportation and traffic, but also water and energy management. The continuous consideration and monitoring of all these urban services is the objective of an intelligent city, also called knowledge-based city or smart city. In this regard, a smart city can be defined as a city that uses digital technologies to improve the quality and performance of the aforementioned urban services.

Due to the urbanization changes, cities are increasingly confronted with large scale problems such as pollution or the lack of mobility. To deal efficiently with these problems, smart cities require an intelligent mobility strategy to inform, guide and support the inhabitants. In this regard, pedestrians, cyclists and motorists have different preferences. While some prefer the least crowded path to reach their destination, even if the distance is longer, some prefer the direct or fastest route. Others consider the least polluted way to avoid health problems due to existing city pollution or low air quality. Moreover, available parking places around the destination are an important factor to be considered in routing and navigation decisions. Smart mobility is a concept which offers decisions based on various data gathered about the current traffic and pollution situation inside the smart city, recommending routes based on the users preferences or to solve traffic jams.

In this paper, we propose a smart mobility strategy based on the analysis of large datasets which are collected from different sensors, all deployed at various places in a smart city. The use case applies the Open Data Aarhus datasets, which have been created between August and September 2014 in the smart city of Aarhus in Denmark. The data contains various features such as pollution, road traffic, weather, parking, as well as cultural and library event information. The presented smart mobility approach uses the pollution and road traffic information from the data. The analysis and evaluation of the data is done with a Hadoop cluster using MapReduce to concurrently calculate routing possibilities and to find the least polluted way from start to end points based on current conditions indicated by the sensor data.

The paper is structured as follows: Section 2 provides a background on the topics smart city, smart mobility and smart environment and introduces big data analytics. Section 3 describes the framework and method which has been used to conceptualize the prototype. The implementation of the framework and the application is presented in Section 4. Section 5 provides additional results of the analysis and evaluation. Section 6 summarizes the paper and gives an overview of potential future work.

II. Background

Smart city can be defined as urban areas that exploit operational data, such as traffic congestion, power consumption statistics and public safety events, to optimize the operation of urban services [1]. Smart city focuses on the use of network infrastructure to consider social, cultural and urban scenarios and improve the efficiency of the city [2]. It uses ICT and semantic technologies [3] to improve the quality and performance of its urban services. Urban services in this definition relate to traffic management, environmental management, energy management, or environmental protection. The aim is to manage urban services intelligently in order to reduce city spending and thus improve the living conditions of citizens. Going one step forward, a smart city is a complex construct composed of smart buildings, smart energy, smart environment, smart energy or smart health [4].

978-1-5386-6531-2/18/$31.00 ©2018 IEEE

A. Smart Environment

The high rate of urbanization and traffic considerably increases the degree of pollution in cities. To solve these problems, cities use intelligent environmental management solutions to monitor and evaluate different environmental parameters. In addition to intelligent management of waste, water, electricity or lighting, the conditions of air and soil are crucial for a healthy climate and the health of inhabitants. Disaster management is another key aspect of smart environment, which deals with the observation of potential high water levels or the early detection of earthquakes. Sensors and cameras are deployed at strategic important locations to gather, monitor and analyze reference data for immediate actions (e.g. limitation of traffic in highly polluted area).

B. Smart Mobility

Smart mobility, or “intelligent mobility” is a core pillar of a smart city strategy. Basically, smart mobility is considered as an intelligent traffic management to reduce traffic jams in cities. Large cities have an increasingly high urbanization rate and for this reason it is difficult to organize activities in the inner city due to congestion. Urban traffic experts estimate that 30% of the vehicles in the inner areas of big cities are looking for a parking space, and an average of 7.8 minutes is needed to find one [5]. Smart mobility projects can improve the flow of traffic in cities and thus make a city more attractive and encourage businesses to expand their activities. Smart mobility analysis is able to predict traffic in cities using big data technologies.

C. Big Data Analytics

Big data is a popular term used to describe the exponential growth, availability, and use of data - structured, semi-structured and unstructured. A general definition for big data is given by Gartner[1] in their 3V-Model in which it is characterized by three V’s, namely the size (volume) of the information, the type of data generated (variety) and its rate of production (veracity).

The big data technology Hadoop utilizing MapReduce - a programming model for distributed parallelizable problems - is used in the presented work. Hadoop and MapReduce have been applied in previous research related to knowledge discovery from social media analysis using big data-provided sentiment analysis [6].

1) Hadoop Distributed File System (HDFS): HDFS is a Java-based distributed file system that allows reliable and persistent storage as well as fast access to large data volumes. It is used to store and retrieve data within a Hadoop architecture [7]. In order to be able to store the data distributed in the cluster, they are split into blocks (128MB) and distributed to the nodes of a Hadoop cluster. For each of these blocks, three copies are created on different nodes in the cluster to guarantee fault and failure safety. HDFS is a master-slave system and the architecture of HDFS is described by services as the NameNode, DataNodes, and the SecondaryNameNode which are responsible for the storage and management of data in Hadoop.

2) MapReduce: MapReduce handles the assignment and execution of queries for data stored in HDFS and is a core part of the Hadoop framework for calculating large amounts of data in a small amount of time [8]. The MapReduce framework is designed for large parallelizable problems and is able to process data on a very large number of nodes. The MapReduce architecture is based on a master-slave system. The components required for job parallelization are the JobTracker and the TaskTracker.

The MapReduce principle consists of two main phases. The first phase is the map phase and the second is the reduce phase. Figure 1 illustrates the MapReduce principle.

Map Phase: The data to be processed is divided into splits. A map task is started for each split. For the map task to run normally, the JobTracker queries the NameNode for the required information and the location in the cluster. Once the response is received, the JobTracker sends the map function to the TaskTrackers. The TaskTracker retrieves data from its DataNode and executes the map functions. Once the map functions complete their processing, the results are saved. The data is processed in the form of key-value pairs.

Reduce Phase: The reduce part combines the different results of the map phase and consolidates them into a single final result. The reduce tasks get the output of the map tasks for further processing. All data that has the same key is processed in the same reduce task.

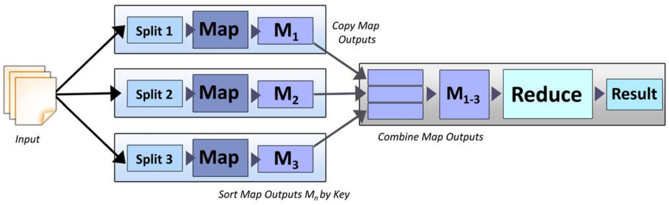


Fig. 1. MapReduce job with three map tasks and one reduce task, adapted from [9]

III. Method

One of the major challenges in smart mobility is intelligent routing of traffic in combination with environmental aspects. In order to address this problem, an algorithm has been developed to find the least polluted route from one city location to another (based on available sensor data). The algorithm is performed in five steps. In the following, the individual steps of the algorithm are detailed.

A. Algorithm

1) Creating a sensor map: The first step of the algorithm is to create a graph with all the road sections of the smart city. Therefore, the longitude and latitude information of sensors is processed. An example of a graph is shown in Fig. 2 a).

[1] https://www.gartner.com/it-glossary/big-data/

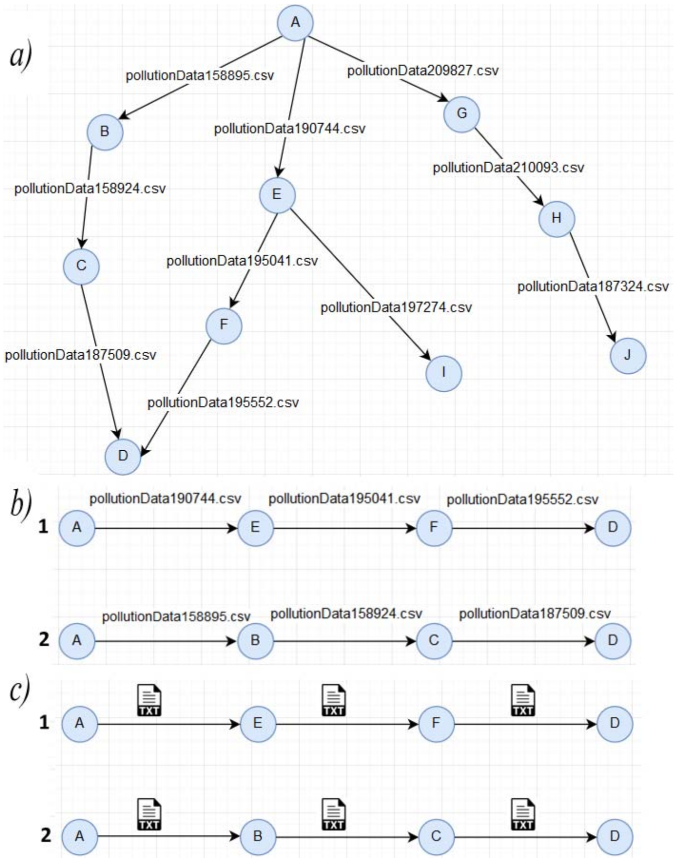


Fig. 2. a) A graph created from road section information, b) The shortest routes from point A to point D, c) Data extracted from pollution sensor data.

2) Determination of a set of shortest paths: The second step is to determine a set of shortest paths from the selected start location to an end location. In this step, potential combinations of routes are determined and a selection of the shortest paths is done in order to avoid long road section combinations. As an example, the road section graph from Fig. 2 a) is used to implement this step. Assuming that the shortest paths from point A to point D are to be determined, two possible paths are found, which are illustrated in Fig. 2 b).

3) Extraction of pollution data: The third step is to extract the current pollution information of each road section. For the set of shortest paths, a request is sent to extract relevant road section's pollution data corresponding to a given time-stamp. The extracted data is stored in .txt format.

4) Calculating the least polluted route: The fourth step of the algorithm is to calculate the least polluted route using the Hadoop cluster. To perform this calculation, the largest pollution values (highest contamination) are extracted from each .txt file. The values of road section pollution are compared with each other to determine the lowest values (least contamination). The smallest valid value combination indicates the least polluted route.

5) Visualization of "healthiest" route: The information of the least polluted route is provided as final result. A list of the longitude and latitude of sensors and road sections is used to show the least polluted route on a map visualization.

Figure 3 summarizes the overall process of the algorithm and components used.

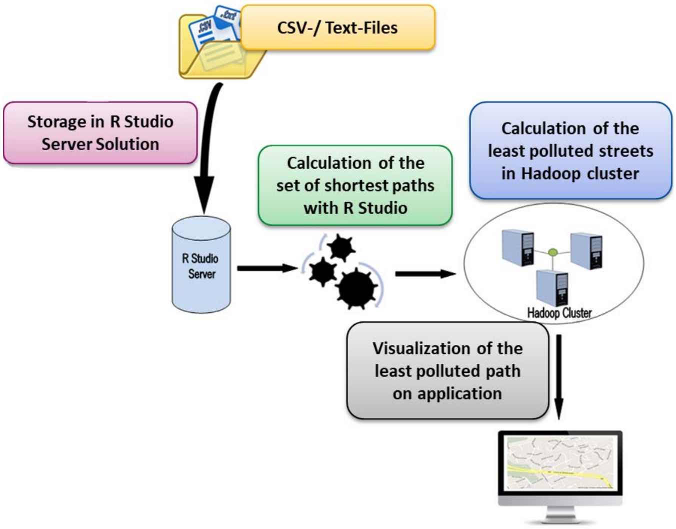

CSV-/ Text-Files
csv
txt
Storage in R Studio Server Solution
Calculation of the set of shortest paths with R Studio
Calculation of the least polluted streets in Hadoop cluster
R Studio Server
Hadoop Cluster
Visualization of the least polluted path on application

Fig. 3. Overview of the algorithm to calculate the least polluted path in Hadoop cluster

IV. Implementation

For the implementation of the proposed smart mobility algorithm, an interactive R web application has been created in R Studio. The Shiny package has been used to rapid prototype the web application in order to create an interface for users to determine the least polluted road. The datasets and the web application are described in the following.

A. Datasets

The selected datasets refer to the city of Aarhus in Denmark. Further information about Open Data Aarhus is available at the main project website[2]. The CityPulse EU FP7 Project[3] provides a detailed overview and introduction to the Open Data Aarhus (ODAA) datasets.

1) Road Traffic: The traffic metadata provides global information about all sensor locations in the city. The sensors are mounted on road sections. Each row in the data of the *trafficMetaData.csv* file provides information about two sensors. For each pair of sensors, a larger set of information is available. These include, for example, the name of the road, the measuring time, the names of points and streets, parts of the city, longitude and latitude, measured speed, type of road or the distance between the two points in meters. The analyzed file *trafficMetaData.csv* has a total of 26 variables and 449 rows of data. The collection of *trafficData*.csv* files provides information about the state of traffic on each road section. For example, the records contain the number or average speed of cars that drive on the road section (every five minutes). Each file has a total of nine variables.

[2] https://www.odaa.dk/

[3] http://www.ict-citypulse.eu/page/

2) Pollution: The collection of *pollutionData*.csv* files contains information about the pollution of the air in each section of the road. Every five minutes, the environmental pollution level is documented in each section of the road. In order to measure the pollution in each road section, a sensor is simulated on each road section. Each file has a total of eight variables, including values for ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, longitude, latitude of sensors and a time-stamp.

The data set also contains weather and parking information. These were not further considered in our analysis method, but could play a decisive role in smart mobility and are therefore listed hereafter.

3) Parking: The parking metadata files contain the addresses of eight parking spaces in the city of Aarhus. For each parking lot there is the information about its garage code which is the name and also the geographical position (city, street, postal code, house number, latitude and longitude). Each metadata file has a total of seven variables. The parking records contain information about the conditions of the parking lots (e.g. the total parking space, the number of cars at a certain time, the name of the car park).

4) Weather: The weather data provides information on the weather conditions in the city of Aarhus. This data is grouped into five files which contain information about dew point, humidity, air pressure, temperature, wind direction, wind speed and date as well as time of recording.

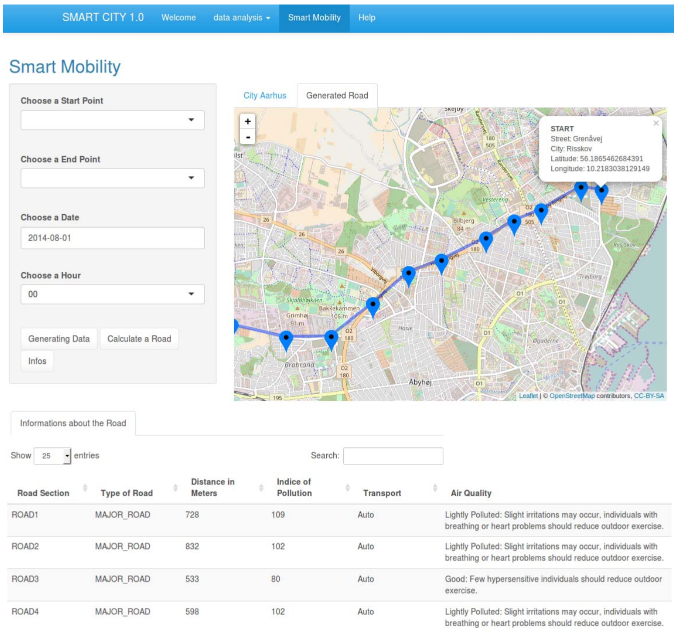
B. Framework, Cluster, IDE and Configuration

A cluster typically consists of a collection of interconnected standalone computers that work together as a single integrated computing resource [10]. Many organizations use computer clusters to maximize processing time, increase data retention, and implement faster data storage and retrieval techniques [11]. As part of this work, the test cluster consisted of three machines, created using Hadoop 2.7.2. R (3.3.0 Beta) and R Studio (1.0.136) have been installed on the master of the cluster. Since R is a free and scalable software, additional packages could be downloaded and integrated easily. The package Shiny allowed the development of a GUI for the visualization of the analytics results in form of an application. Eclipse has been installed on the cluster and was used as part of this work to develop the "Smart Mobility" module. The MapReduce algorithm to calculate the least polluted route has been programmed in Java.

C. Application

The functionality of the R application is divided into two main modules for data analysis and smart mobility. The module for data analysis has four components (analysis of traffic, pollution, weather and parking). The smart mobility module has the functionality to calculate the least polluted road given a start point, end point, date and time information. To create the Shiny webpage-based user interface in R, two separate files have been created. The *ui.R* file contains the user interface script, and therefore all implementation details related to the GUI and shiny package. The file is also used to manage inputs through widgets. The *server.R* file contains the server script which is used to implement actions performed by the ui.R file. The file allows the creation of tables or graphics with R language. The file creates the outputs that are displayed in the web application.

1) Data Analysis: In the data analysis compontent, a visualization highlights a selected road section. The road section is displayed on a map with sensor locations and provides additional information of selected sensor locations. Below the map, the user has the option to analyze the data with different instruments to explore the data set using a descriptive analysis and visualizations. By selecting the tabs, corresponding R functions are executed and integrated as responsive widgets into the interface.



Choose a Date: 2014-08-01

Choose a Hour: 00

START
Street: Grenåvej
City: Risskov
Latitude: 56.1865462684391
Longitude: 10.2183038129149

Road Section	Type of Road	Distance in Meters	Indice of Pollution	Transport	Air Quality
ROAD1	MAJOR_ROAD	728	109	Auto	Lightly Polluted: Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
ROAD2	MAJOR_ROAD	832	102	Auto	Lightly Polluted: Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
ROAD3	MAJOR_ROAD	533	80	Auto	Good: Few hypersensitive individuals should reduce outdoor exercise.
ROAD4	MAJOR_ROAD	598	102	Auto	Lightly Polluted: Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.

Fig. 4. Prototype application: Smart mobility through big data provided traffic and pollution analysis - Visualization of the least polluted route between two selected points.

2) Smart mobility: Smart mobility is the second component of the prototype. Figure 4 shows the smart mobility module of the developed R Shiny application. The implementation of the smart mobility module has been created with two separate files *mobilityUI.R* and *mobilityServer.R*. The mobilityUI.R file implements the widgets. The mobilityServer.R file implements the actions which are performed in the interface using the mobilityUI.R file. The mobilityServer.R file is the server-side logic of the mobilityUI.R script. The analysis starts with the selection of a start and an end point. Here, a point is characterized by the longitude and latitude of the sensor location. This information is retrieved by using the *trafficMetaData.csv* file.

The least polluted route is calculated on the Hadoop cluster using algorithms implemented using Java. The Java code implements both, the map() and the reduce() function. The map()

function creates a file from the road section's pollution and returns the name of this file and the values of ozone, particulate matter, carbon monoxide, sulfur dioxide and nitrogen dioxide contained in this file. The reduce() function returns the name of each file and the maximum value of the pollutants. It should be noted that the outputs of the function reduce() are actually the name of the path of each file in HDFS (key) and the maximum value of the pollution values contained in each file. Using a Hadoop command, the output of the reduce() function is copied to a new .txt file.

Finally, the values of the reduce() function are compared to identify the lowest values among the pollutants. If the smallest value has been identified, the key of this smallest value is returned. All keys together, allow the identification of the best combination of road sections. After the route has been identified, the web application shows the least polluted road on a map visualization. The user can find additional information about each road section with the type of the road, transport, distance in meters and the value of the maximum pollution. The air quality is also further described with a short text about the health implications. The air quality, quality levels and health implications are based on the Air Pollution Index (API) from China's State Environment Protection Agency (SEPA) and the Air Quality Index (AQI) from China's Ministry of Environmental Protection (MEP) [12]. These indicators match to the Open Data Aarhus dataset and therefore have been selected as reference indicators to evaluate the pollution. The API/AQI levels indicate the concentration of six atmospheric pollutants, namely ozone, suspended particulates smaller than 10 and 2.5 μm, carbon monoxide, sulfur dioxide and nitrogen dioxide. An individual score is assigned to the level of each pollutant and the final API/AQI level is the highest of those six scores. The quality levels and health implications are described in Table I.

TABLE I
AIR POLLUTION / QUALITY INDEX, BASED ON THE AQI FROM CHINA'S MINISTRY OF ENVIRONMENTAL PROTECTION (MEP)

API / AQI	Air Pollution Level	Health Implications
0 - 50	Excellent	No health implications
51 -100	Good	Few hypersensitive individuals should reduce outdoor exercise.
101-150	Slightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
151-200	Lightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
201-250	Moderately Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
251-300	Heavily Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
300+	Severely Polluted	Healthy people will experience reduced endurance in activities. There may be strong irritations and symptoms and may trigger other illnesses. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid outdoor activities.

V. Results

The data analysis module of the R application has been used to retrieve, analyze, but also to visualize and interpret results from the datasets. The smart mobility module using MapReduce has been implemented to find the best route in order to avoid negative health implications. In the following, three use cases are given, which point out further potentials of the analysis of traffic, pollution and parking data.

A. Traffic analysis and prediction

Figure 5 a) visualizes the traffic data for a specific day (02.08.2014) from the road section Nordre Ringgade 3 to Vestre Ringgade 61 in a box plot. The box plot shows the distribution of data based on the minimum, first quartile, median, third quartile, and maximum. The Y-axis corresponds to the speed of the cars on this route. The X-axis corresponds to the different hours of the day. The second box plot b) shows the speed of the cars during August 2014 on the same route. The Y-axis represents the speed of the cars, and the X-axis represents the different days of the month. Since the sensor data is available for all road sections, the traffic situation on the proposed route can be analyzed in more detail based on historical values and used for additional modification and prediction of changes in the "healthiest" route.

B. Monitoring of pollution levels

The second scenario is the analysis of environmental pollution for the same road section from Nordre Ringgade 3 to Vestre Ringgade 61. Figure 5 c) shows an example of a bar plot created using the web application. The figure represents the pollution status for the day of 01.08.2014 at 0 am. The X-axis represents the various pollutants that were evaluated. To evaluate the pollution in a road section for a defined time, the maximum levels of pollutants are determined during the full hour. The pollutants shown in Figure 5 c) are ozone, fine particles, carbon monoxide, sulfur dioxide and nitrogen dioxide. According to the API, the pollutant with the highest value determines the level of pollution. In said case, the pollutant having the highest value is carbon monoxide. Its value is approximately 85. In this example, the value is less than 100 and the application assumes that the degree of pollution is acceptable (Table I).

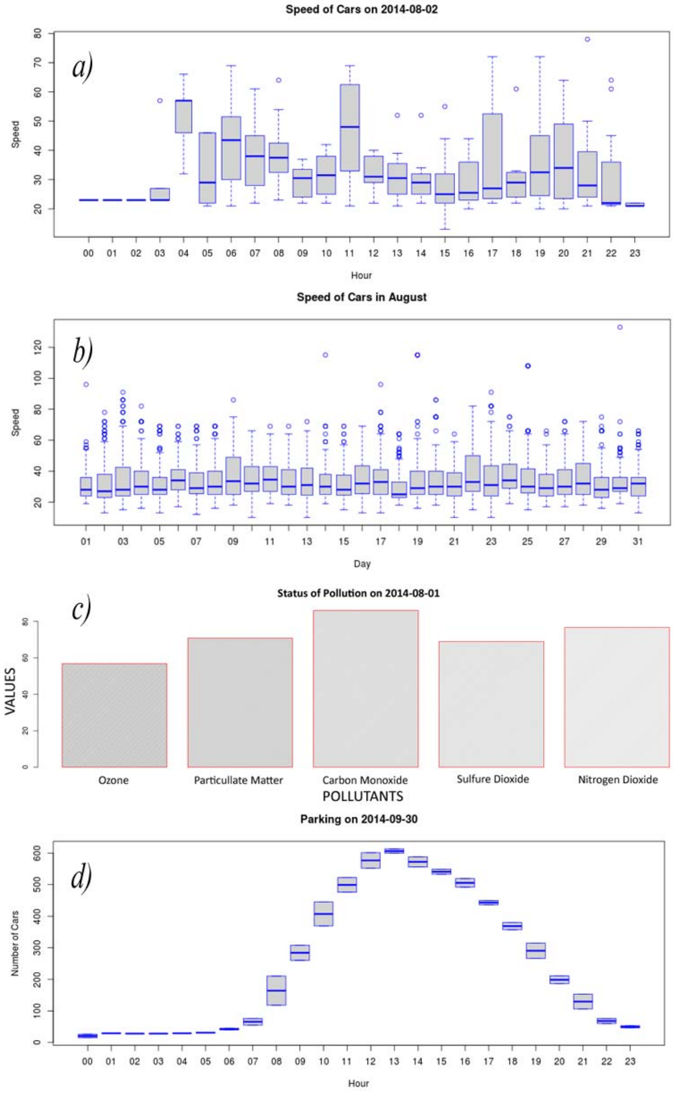


Fig. 5. Boxplot visualizations: a) Speed of cars on 02.08.2014, b) Speed of cars in August 2014, c) Pollution conditions at road section from Nordre Ringgade 3 to Vestre Ringgade 61 on 01.08.2014 (0 am), d) Parking data: BRUUNS parking garage on 30.09.2014

C. Analysis of parking data

The presented smart mobility approach can be further improved by the analysis of parking data. As an example, another scenario shows the state of the Bruuns parking lot in Aarhus on 30. September 2014. A total of 8 parking lots can be analyzed considering the whole parking dataset. Figure 5 d) shows the state of the Bruuns car park for the selected day. The X-axis represents the hours of the day, the Y-axis represents the number of cars. It is seen, that from 0 am to 6 am the parking lot is almost empty (about 20 cars). From 8 am to 1 pm, the number of cars in the parking lot increases. From 2 pm to 11 pm, the number of cars in the car park decreases.

VI. Conclusion and Future Work

The main goal of this research was to implement a use case of smart mobility considering environmental aspects. The aim was to find the least polluted road between one location and another location on the basis of smart city sensor data. A further aim of this work was to analyze the available datasets by statistical evaluation and open potentials for prediction through machine learning in future work.

In this paper, open datasets were used from the smart city of Aarhus in Denmark between August and September 2014. The datasets refer to the state of pollution, traffic, weather, and parking. It has been shown that parallel processing of sensor data can be implemented with a Hadoop cluster to improve the processing speed with MapReduce algorithm. An interactive web application using the Shiny package in R was created for visualization and analysis purpose.

In future work, the authors envisage extending the scenario to include weather data, a traffic simulation and the consideration of environmental conditions.

Acknowledgment

The authors would like to thank the Open Data Aarhus (ODAA) initiative and CityPulse EU FP7 Project for publishing open datasets.

References

[1] K. B. Ahmed, M. Bouhorma, M. B. Ahmed, and A. Radenski, “Visual sentiment prediction with transfer learning and big data analytics for smart cities,” in *Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on*. IEEE, 2016, pp. 800–805.

[2] V. Lisena, M. Paschero, V. Gentile, P. Amicucci, A. Rizzi, and F. F. Mascioli, “A new method to restore the water quality level through the use of electric boats,” in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–4.

[3] S. Bischof, A. Karapantelakis, C.-S. Nechifor, A. P. Sheth, A. Mileo, and P. Barnaghi, “Semantic modelling of smart city data,” 2014.

[4] A. Gaur, B. Scotney, G. Parr, and S. McClean, “Smart city architecture and its applications based on iot,” *Procedia Computer Science*, vol. 52, pp. 1089–1094, 2015.

[5] A. I. Niculescu, B. Wadhwa, and E. Quek, “Technologies for the future: Evaluating a voice enabled smart city parking application,” in *User Science and Engineering (i-USEr), 2016 4th International Conference on*. IEEE, 2016, pp. 46–50.

[6] M. Bohlouli, J. Dalter, M. Dornhöfer, J. Zenkert, and M. Fathi, “Knowledge discovery from social media using big data-provided sentiment analysis (somabit),” *Journal of Information Science*, vol. 41, no. 6, pp. 779–798, 2015.

[7] R. K. Chawda and G. Thakur, “Big data and advanced analytics tools,” in *Colossal Data Analysis and Networking (CDAN), Symposium on*. IEEE, 2016, pp. 1–8.

[8] P. Lathiya and R. Rani, “Improved cure clustering for big data using hadoop and mapreduce,” in *Inventive Computation Technologies (ICICT), International Conference on*, vol. 3. IEEE, 2016, pp. 1–5.

[9] M. Steyer and H.-P. Grahsl, “Big-data-analyse mit apache hadoop in der windows azure cloud,” 2015. [Online]. Available: https://entwickler.de/online/big-data-analyse-apache-hadoop-windows-azure-cloud-167430.html

[10] L. Li and L. Gruenwald, “Smopd-c: An autonomous vertical partitioning technique for distributed databases on cluster computers,” in *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*. IEEE, 2014, pp. 171–178.

[11] Q. Sun, P. Lin, and C. Wang, “Implementing dynamical pattern recognition algorithm on computer cluster,” in *Control Conference (CCC), 2016 35th Chinese*. IEEE, 2016, pp. 5249–5254.

[12] A. Hsu, “Chinas new air quality index: How does it measure up,” 2012. [Online]. Available: https://datadriven.yale.edu/air-quality-2/chinas-new-air-quality-index-how-does-it-measure-up/