# Joanna Riascos

Dr. Sylvain Jaume

Assignment Eight

Capstone: Big Data & Bus Analytics

**Midterm Proposal (Contribution, State of the Art, Data, Method)**

**Contribution**

According to a study by the Union of Concerned Scientists (UCSUSA), “transportation contributed more than half of the carbon monoxide and nitrogen oxides, and almost a quarter of the hydrocarbons emitted into our air.” The health risks of air pollution are shockingly dangerous. It could cause asthma, bronchitis, irritation of the eyes, and increased risk of heart attacks. Also, if a person is exposed to air pollution in the long run it could cause cancer and permanent lung damage. Not only does air pollution affect us, it also affects the environment. In reference to the United States Environmental Protection Agency air pollution can cause the following environmental effects: acid rain, eutrophication, and haze. Acid rain “is precipitation containing harmful amounts of nitric and sulfuric acids.” Eutrophication “is a condition in a water body where high concentrations of nutrients stimulate blooms of algae.” Haze “is caused when sunlight encounters tiny pollutions particles in the air.” Similarly, trucks and heavy cars use a lot of petrol and diesel triggering the release of more carbon dioxide, which causes global warming. In addition, air pollution also affects the wildlife because just as human beings they might suffer from birth effects and respiratory diseases.

If we stop and think about the serious problems that we are facing here then we would want to contribute to helping save the environment. What we believe can help reduce the air pollution produced by the vehicles’ emissions is by implementing smart parking all over the world. An investigation that we have purposed is to explore how smart parking can help solve environmental issues caused by traffic congestion. The purpose is to make cities more sustainable, healthy and happy to live in. By having a smart parking system there will be reduced traffic since there will be less cars looking for an available and open parking space. Also, there will be a reduction in air pollution because it will significantly decrease driving time therefore; it will lower the car emissions.

With the help of smart parking we will be able to contribute in helping save the environment. In addition, with the dataset that we have chosen for our data analysis project we could build models to extract real time parking data and predict future driver behaviors. As an example, a driver might have a daily routine of parking at the same time every day. If we analyze and extract the data we could have a better understanding of when and at what time the same driver parks his or her vehicle. This will help immensely in having an accurate parking management system to predict when most likely a parking spot is going to be available or vacant. Therefore, as I mentioned in the previous paragraph, I believe we should investigate how smart parking can help solve traffic and environmental issues. We need to find a way to create and implement a smart parking management system in our cities with the help of big data analytics.

**State-of-the-Art**

The smart parking management system has been implemented in various parts of the world, such as in the United Sates, United Kingdom, Europe, Japan and India. One of the cities that have introduced the concept of the smart parking assistance program is in Ellicott City, Md. They have developed a smart parking system that utilizes sensors to detect whenever a parking space is open and then it sends the data information to the drivers to let them know which parking spot is vacant. According to the Los Angeles County, the smart parking system “resulted in saving 31.3 million hours of travel time, 38 million gallons of fuel and 337,000 metric tons of carbon dioxide per year.”

There have been various related works throughout the years that discuss the importance of having and creating a smart parking management system. According to the article, “A Reservation-based Smart Parking System” by Hongwei Wang and Wenbo He. They have designed and implemented a prototype called “Reservation-based Smart Parking System (RSPS) which depicts and provides in real time parking data information to the drivers through an application and also, offers parking reservations as part of the system. The system contains a subsystem of sensory networks that deliver real time parking information in the upper layer. These sensory nodes detect and monitor the status of whether a parking space is occupied or vacant. They have also developed a parking demand model to simulate the real world traffic traces to acknowledge the parking demand.

What the authors have used for the simulator implementation is an object-oriented design that realizes the interactions between objects such as, drivers and parking. Many drivers don’t behave the same way but what they have in common is that they always look for a parking garage or lot that offers a better and more convenient price. In the simulation setup, the authors have used the map of Los Angeles Downtown as a target area. This part of Los Angeles has a lot of interstate highways so they depict the outgoing traffic in two different days of the week. “The peak time of incoming traffic is from 6am to 10am, and the rush hours of outgoing traffic is during 5pm to 8pm.” For this reason, we can see that the parking demand might be high after work since the majority of drivers go back home during those hours. Furthermore, the authors established a system hardware that is organized into three main components, the sensor network, and the mobile device. The sensors are integrated with two different wireless motes. The mote communicates the module of Bluetooth on a smartphone. As a result, “the sensor confirms the identity of users when vehicle is detected in reserved parking lot.” The mobile phone allows the Bluetooth module to communicate with the sensors to verify a user’s identity.

With the help of the smart parking system drivers can make parking reservations ahead of time. Also, the article states that the driving distance decreases at peak time, rather than increasing because drivers know ahead of time the “states” of their parking spaces. Similarly, the system also collects and stores data “about the performance metrics, including the status of parking space, reservation time, parking location, driver’s identity.” And also, the system allows drivers to check the parking, reservation information and where the parking spot is located.

Furthermore, according to the article “A Cloud-Based Smart-Parking System Based on Internet-of-Things Technology” by Thanh Nam Pham, Ming-Fong Tsai, Duc Binh Nyguyen, Chyi-Ren Dow, and Der-Jiunn Deng. The authors have developed a “novel algorithm that increases the efficiency of the current cloud based smart parking system and develops a network architecture based on the Internet-of-Things technology.” The algorithm that they have proposed will help greatly in reducing driver wait time and achieve successful parking. The system will help drivers find a free parking space at the most convenient price. The smart parking system consists of the having a vehicle park as an Internet-of-Things technology (loT) network. It will contain the data that includes the GPS location and distance between car garage areas.

All of this information then is transferred to a data center. The data center is a cloud storage server that “calculate the costs of a parking request, and these costs are frequently updated and are accessible any time by the vehicles in the network.” Moreover, the authors’ algorithm system consists of Radio-Frequency Identification (RFID) technology that monitors car parks. The Radio-Frequency Identification reader counts how much of free parking spaces there is in a parking lot or garage. “The use of RFID facilitates implementation of a large-scale system at low cost.” Whenever a driver or a user logs into the system, the driver should be able to choose any parking space he or she wants. Afterwards, the system will notify the driver with the parking spot saying that it is “pending” which doesn’t allow other drivers to reserve the parking space.

The system architecture consists of a cloud-based server, local unit, control unit, screen and RFID tag or ID card. The cloud-based server is a web entity that is in charge of the storage of the data and information that is provided by the local unit. The local unit is located in each car park and it storages the information and data of each vehicle space. The control unit is an Arduino module that connects to the RFID reader. “The card reader authenticates the user information and then displays this information on the screen.” If indeed, the information that the card reader reads is correct then it automatically open the parking garage’s door so that the vehicle can enter and park his or her car. Next, the screen depicts the “information on the capacity of the local car park, the total current percentage of free spaces, and the status RFID tag check.” Lastly, the RFID tag and the ID card are used to identify the user information and then calculate the amount of total free spaces that there might be available in a car garage or lot.

Likewise, in reference to the article “Towards Smart Traffic Management Systems: Vacant On-Street Parking Spot Detection Based on Video Analytics” by Xavier Sevillano, and Elena Marmol. They have created a sustainable solution to detect and find vacant parking spots. This work advocates fusing the information from small-scale sensor-based detectors with that obtained from exploiting the widely-deployed video surveillance camera networks.” The article discusses how video analytics is very much useful in regards to smart city solutions based on data and how they have developed a “vacant parking spot detection system based on the use of video analytics, rather than in the networks data fusion process itself.” The parking spot detection system is divided in a two-phase process. First, it consists of the training phase, in which the system detects the real time visualization of an occupied and vacant parking space. Secondly, there is a test phase in which the system analyzes and predicts the status of a parking spot.

The training phase is basically the “set up process of the system” whereas; the test phase “corresponds to the exploitation of the system.” Afterwards, they test their module the one that determines whether a parking spot is vacant or available by testing with classifiers. They employ the common algorithm, k-nearest neighbours (kNN).With the help of the kNN they will determine the status of a parking sport by looking at the sports in the training dataset that look visually very similar. The authors used the Euclidean distance “for measuring visual similarity.” Likewise, they have applied one of the most well known classifier systems called, Support Vector Machines (SVM) that differentiates vacant and occupied classes. This way a driver should be able to determine when a parking spot is vacant. As well, as part of the algorithm the authors are implementing the video data description. The authors recorded their own video in order to collect external factors. This way they were able to create their parking video analytics detector.

In conclusion, the less there exists traffic congestion the better for the environment. As I mentioned in the introductory paragraphs there are a lot of harmful effects that come from air pollution such as asthma and respiratory diseases caused by vehicle emissions. To solve these issues we need to implement smart parking in our cities. With smart parking we can create smart technology systems that will benefit animals, the environment, and us. By implementing smart parking we will be able to spend less money on gas, reduced traffic and stress, improve air quality, increased safety, optimized parking, less accidents, less time spent, decreased management costs, and enhanced driver experience.

**Data**

According to a study from the New Jersey Institute of Technology (NJIT), smart technologies “will generate revenues of more than $27.5 billion by 2013.” And 88 cities around the world will adopt these smart technologies by 2025. Parking is the foundation of a smart city. Smart cities are in the need to look out for reducing parking and street congestion issues. Parking plays a major role in finding a solution to these issues. As reported by the Information Technology Services Reports, “Thirty percent of all traffic congestion in urban areas is caused by drivers looking for a parking space.” Traffic congestion has always been a frightening and global issue that is increasing since there are more drivers in the streets. Also, searching and finding a parking space is sometimes nearly impossible. According to a report, “Smart Parking could result in 2,20,000 gallons of fuels saving till 2030 and approx. 3,00,000 gallons of fuels saved by 2050, if implemented successfully.” With the implementation of smart technologies in parking, the driver should be able to obtain information and see in real-time available parking spaces. As well, it will allow the drivers to reserve their parking spaces ahead of time. This implementation will reduce car emissions since drivers won’t have the need to drive in circles until they find a parking spot. Similarly, it may reduce stress for the drivers because the majority of times drivers argue over parking spaces.

Many others of the benefits that the cities may have by employing “smart parking” is help traffic in the city to flow more freely, increased mobility, greener cities, optimize parking space usage, enable smart decisions by using big data analytics, monitor and manage available parking spaces, increased revenue, and savings of car fuel. Likewise, there will be savings in paper costs since smart phone applications can take care of the billing aspects of paying for parking reservations. The development for creating “smart parking” consists of building a workflow. The smart parking workflow analysis consists of six different processes. The first process is the one single gateway covers one area of 1 square Km. Next; the gateway sends the data via the Internet to the database. Then, if a vehicle is parked the occupancy is reported instantly to the users. Afterwards, the central control system gets real-time data about parking occupancy during the day. Subsequently, when a user is connected to the payment system the parking authorities can identify if a driver paid or not. Lastly, when a car is parked the sensor is detected.

* The parking dataset comes from the City of Aarhus in Denmark. It contains data from February 2015 until October 2015.
* The data file contains the following fields: vehicle count, update time, id, total spaces, and garage code.
* The data was loaded and read in Zepellin.
* We imported the “apache spark sql function” and the date time format packages.
* Then we adjusted the path and added the location of the data to the script.
* The vehicle count field is the number of vehicles that park in the parking lots.
* The total spaces is the number of available parking spots.
* Garage code is just the code to enter the garage.
* Postal code is just the city’s postal code
* Street is where the house is located.
* House number is just the house’s number.
* Latitude and the longitude are the geographical coordinates.
* We merged the parking and the pollution dataset in order to add the "ozone” column.
* The data fields that we didn’t include when loading our data was the id, update time, and stream time since we believed that those fields didn’t give us any meaningful information.

In conclusion, smart cities and smart parking are the future of the world. They are the way to go. Smart technologies help the city’s assets and improve its citizens’ way of life. Likewise, big data analytics plays a major role in the process of implementing these technologies since big data helps in the identifying new and efficient ways of doing smarter businesses. The more data that we collect, the better because with a lot of data we can analyze and make more meaningful conclusions and decisions to help solve a problem or issue. Smart cities and smart parking are a result of big data analytics since they go hand in hand. The smartness of a city’s information is extracted from the data sources of the government and healthcare, etc. That’s the reason why we should all get together to implement these smart technologies all over the world.

**Method**

The algorithm consists of running a linear regression and a time series analysis. What is linear regression? Linear regression is a methodology for modeling many relationships between a scalar dependent variable y and one or more variables denoted by x. If it’s only one explanatory variable then we called it simple linear regression. If it’s more than one explanatory variable then it’s called multiple linear regression. Linear regression is a supervised machine-learning algorithm. It is the foundation of many complex and long algorithms. In linear regression the relationships are modeled by using and implementing predictor functions where the unknown model parameters reside in the data. These types of models are linear models.

Likewise, linear regression has many useful and practical uses. Linear regression can be used to fit a predictive model to an observed dataset of x and y values. Afterwards, when developing this type of model, an additional value x is then given without its accompanying value of y, which is the fitted value. If we get a variable, y and a number (x1,..xp) it might be related to y. Linear regression analysis can be implemented to count the relationship between y and x1 which may result in not having a relationship at all and to analyze which subsets of the x1 contain useful information about y. Similarly, linear regression implements a statistical model that even though some relationships between independent and dependent variables are almost linear they will still show optimal results and observations. Linear regression models, most of the time are fitted using the least squares method. The least squares technique is the easiest and most common used form of linear regression and delivers an optimal solution to the issue of having to find the best fitting straight line through a set of points. The data is then transformed so that the resulting line is straight. Linear regression models may be fitted in many other ways however; they may also be fitted by minimizing the “lack of fit” or by minimizing a penalized version of the least square function as in a ridge function. The main assumptions of the linear regression equation are the following:

* Linearity – the mean of the response variable is a linear combination of the parameters and the predictor variables. This is very much less restrictive than it may at first seem.
* Homoscedasticity (constant variance) – means that the different response variables have the same variance in their errors regardless of the predictor variables.
* Independence of errors – this assumes that the errors of the response variables are uncorrelated with each other. No correlation between consecutive errors in such a case of time series data.
* Multicollinearity - happens when the independent variables are not independent from each other. Also, the error of the mean has to be independent from the independent variables.
* Normality – when we draw a histogram of the residuals and then examine the normality of the residuals. If we see that the residuals are not skewed, then that means that the assumption is satisfied.
* Equality of variance – if we create a scatter plot and the residuals do not depict in a triangular manner then that means that the equal variance assumption is met.

Ordinary Least Square (OLS)– is a method for estimating the unknown parameters in a linear regression model. It is theoretically simple and computationally straightforward. It is used to establish a line of best fit by minimizing the sum of squares created by the mathematical function. Generalized least squares is an extension of the OLS method that allows efficient estimation of when either heteroscedasticity or correlations or both are present among the terms of the model as long as the form of heteroscedasticity and or correlations or both are present among the error terms of the model as long as the of heteroscedastically and correlation is known independently of the data. GLS can be implemented to perform linear regression when there exists a particular degree of correlation between the residuals in a regression model.

The reason for implementing linear regression is to identify the impact of the parking management system by using different and various tool such as: Zeppelin and Tableau. By running a linear regression we will read in the data and analyze the dataset. We will take as a dependent variable the “Ozone” field and the other fields vehicle count, total spaces, and garage code as the independent variables. What we have done here is that we have merged the two dataset files, the parking and the pollution dataset. We believed that it was a good idea of adding the “Ozone” field from the pollution dataset to our parking dataset since our idea was to investigate how smart parking can help the environment. The reason as to why we chose the “Ozone” column is because car emissions are one of the major contributors of air pollution.

First of all, we will load the parking dataset “aahrus\_parking.csv” file on Zeppelin. Next, we will choose our dependent and independent variables to run the linear regression analysis. Subsequently, we added the following import packages for the linear regression and created the Spark environment. We then added the %r in the script to run our R code in each sections of the code. After running the code we can see that the significant coefficient is vehicle count. The vehicle count is for how many cars there is, the total spaces is for the amount of spaces in the parking lot or garage, the garage code is the name of the garage space and ozone is the amount of ozone. In conclusion, our algorithm consists of the linear regression and the time series analysis. As mentioned previously, all of the analyses will run on Zeppelin and there we will import the parking and pollution datasets. Afterwards, we will merge these two files and then we will take the ozone field count as the dependent variable and the vehicle count, total spaces, and garage code fields as independent variables. Then we will call the lm function and run our analysis. Finally, we will plot these values to look at the residuals vs fitted, and qq-plots. By doing this we will get a better understanding of our data values and be able to perform more accurate analysis.

Time series analysis is composed of a series of methods for analyzing time series data that extracts meaningful statistics and other characteristics and parts of data. Time series is an ordered order of values of a variable at equally spaced time intervals. With the help of time series we can acquire an understanding of the structure that produced a certain observed data. Time series forecasting is the use of a model to predict future values based upon previously observed values. Time series analysis is used for many applications such as: economic forecasting, sales forecasting, budgetary analysis, stock market analysis, yield projections, process and quality control, inventory studies, workload projections utility studies, and census analysis. Objective of time series analysis consists of data compression (provides compact description of the data, descriptive (identify patterns in correlated data), explanatory (understanding and modeling the data), forecasting (prediction of future trends from old trends), intervention analysis (how a single event changes within time), prediction (use of the model to predict future values of time series). Also, the advantages of time series models are quite useful models when you have serially correlated data. The best way to deal with temporal effects is by running the time series analysis. Spectral analysis (frequency domain)- aims to isolate periodic or cyclical components in a time series. Intervention analysis – it is used to establish if an event can lead to a variation in the time series. Explanative analysis – studies the cross correlation or relationship between two series and the dependence of one on another. We then ran a time series analysis using the vehicle count and the amount of ozone. First, we took the count of the vehicle count, the sum of the vehicle count, and the rank of the vehicle count. We chose the rank as the data calculation because we wanted to visualize the amount of ozone.