Big Data Analysis of Public Transportation Data

A picture containing drawing

Description automatically generated

Thesis - Rough Draft

Date:

2nd September 2020

Author:

Aausuman Deep

Supervisor:

Dr. Alejandro Arbelaez

Second Reader:

Dr. Gregory Provan

**Abstract**

The data generated in the public transportation sector has an enormous potential to assist us in making key decisions to improve those services over which a vast majority of population depends for their travels and daily commutes. This can be achieved by properly cleaning, treating and processing the raw available data and extract relevant information through various analytical techniques.

In the following project, I have subjected a month of public bus network data of Dublin Ireland and performed a big data analysis over it to harness some important details that can help the concerned department in better understanding the service that they work with, how it is performing and whether there is any room for improvement. Since the data generated at each instant in this network is quite detailed, the level of data analysis that can be performed in this regard is also quite comprehensive. Therefore, usage of a proper big data technology was on the cards and this is why I have used Apache Spark on Hadoop as my main tool here.

The primary concerns related to public bus transportation networks is undoubtedly the delays with which they run on their respective lines throughout the day and the major factors which directly affect these delays, like congestions. An algorithm has been initially produced to identify the most affected lines by these delays, with stepwise delve into deeper and more detailed delay analysis. A prediction model has also been made on the expected arrival times of the buses in the future dates from the data that we have as our principal repository. Furthermore, another algorithm has also been formalized which is capable of handling real time data by the help of Spark streaming, to analyze data from second to second and give us valuable insights on to how the network is performing.

All this analysis has then been compiled by me in an attempt to give out a detailed explanation with minimal background knowledge required by the reader so that anyone reading this thesis document can understand the key insights I have presented in it and utilize this analysis for their own benefit.

**Keywords**

* Public Transportation Data
* Apache Spark
* Hadoop
* PySpark
* Spark Streaming
* Big Data Analysis

**Declaration**

I confirm that, except where indicated through the proper use of citations and references, this is my original work and that I have not submitted it for any other course or degree.

Signed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Aausuman Deep

July 31, 2020

**Acknowledgement**

I would like to show the greatest appreciation to my supervisor Dr. Alejandro Arbelaez for guiding me through this project and for assisting me in the journey, considering we were miles apart in this current pandemic situation of Covid-19. His continued direction in progressing through this thesis, with the efficient scrutiny over my work with constructive feedback on a weekly basis over Microsoft Teams really helped me in completing this dissertation. This fulfilment won’t have been possible without him and for this I am deeply grateful.

I would like to thank Dr. Eric Wolsztynski and Dr. Micheal Cronin for aligning me with the industry level machine learning algorithms and data analysis techniques which form the backbone of any analysis project. I would also like to thank Dr. Gregory Provan who introduced me to Deep Learning mechanisms and how we can work with them by utilizing online computing platforms like Google Collaboration environment. This knowledge had assisted me in understanding the distributive nature of the technology I have used in this thesis. Finally, I would also like to mention Dr. Kieran Hurley, whose collective classes on Python and its various data specific capabilities laid the groundwork for me to choose PySpark as my technology in this thesis.

All the information that I gained and garnered here at UCC has cumulatively assisted me in performing the tasks I set out to do through the course of this dissertation and I appreciate everyone’s efforts for taking this venture with me.

**Dedication**

“Without effort, there are no rewards. And for the effort, we need support.”

I would like to dedicate this research thesis to my parents, without whose support and countless sacrifices, I would not have been able to achieve what I already have in life and whatever I achieve in the future.

Contents

[Introduction 7](#_Toc48090497)

[What is Big Data? 8](#_Toc48090498)

[Big Data Technologies 9](#_Toc48090499)

[Data Storage 9](#_Toc48090500)

[Data Mining 9](#_Toc48090501)

[Data Analytics 10](#_Toc48090502)

[Data Visualization 10](#_Toc48090503)

[Tools and Technologies being used in this project 10](#_Toc48090504)

[Familiarization with our dataset 10](#_Toc48090505)

[Literature Review 14](#_Toc48090506)

[Introduction to the problem statement 14](#_Toc48090507)

[Description of the publications 16](#_Toc48090508)

[An Architecture for Big Data Processing on Intelligent Transportation Systems 16](#_Toc48090509)

[Effective Bus Arrival Time Prediction based on Spark Streaming platform 18](#_Toc48090510)

[Short Term Bus Passenger Demand Prediction 20](#_Toc48090511)

[Summary of main points 23](#_Toc48090512)

[Contributions of these publications to Big Data in Public Transportation sector 24](#_Toc48090513)

[Methodology 25](#_Toc48090514)

# Introduction

What ‘buzzwords’ have you heard while being associated in the information technology industry over the last few years? Every day when we open up our workstations and our phones and arrive at the feed where we obtain our daily dosage of information before getting ready for work, we come across a few repeated words which claim to be the next big thing in the technological industry. Artificial Intelligence, Internet of Things, Blockchain and Quantum computing are to name a few. Every couple of years one of these words creates a hype and seeds itself into the minds of each and every professional trying to make in the industry with an idea of their own. Now, I am not saying that they are over-hyped, but we have to admit, people have been working at them in one or the other way for years already and it seems just a fancy way of reigniting curiosity in us.

But, sometimes one of them comes along and sweeps us off our feet. “Big Data” was already huge by the time I associated myself with it, but to be honest it was worth the wait. I thought it was again one of those hyped up frameworks which will die down to normality eventually after people get bored off it, but it did not and that’s where its sheer potential was first revealed. If a technology, after flowing through each technical valley in the world, still managed to maintain its hype amongst those same people which discarded them like a used-up tissue, it was a real deal. And am I glad that I got into it.

The usage of predictive analytics dates back to the 1940s when governments began using early computers to carry them out. So, it’s easy to say that this is also a very fundamental area of technology, but the big data introduction in recent years has unlocked a whole new level of information processing that can be performed by us.

The need for processing the vast amounts of data being generated every minute in every industry is ever increasing. It is quite important to derive methods to reduce the production of redundant data in this data pool. Technologies are there at our disposal, but how we effectively use them is what makes an analytical algorithm into an analytical funnel. This extracted knowledge can then be used to perform further tasks for predictions and decision making.

As seen in the illustration above, the amount of big data generated increases every second, and we need to model out ways to identify relevant data out of it (this is called data cleaning – we will talk about this later) and after this subsequently we need to extract knowledge from that relevant data.

## What is Big Data?

Before we go any further in this thesis, let’s clarify what exactly is meant by Big Data. By its very definition, it comprises of a set of extremely large datasets that may be analyzed computationally to reveal patterns, trends and associations. It is formed by a combination of structured, semi structured or unstructured data.

When talking about Big Data, it’s 3Vs take the forefront in explaining its significance.

1. Volume – The huge volume of data in various environments.
2. Variety – The varied nature of the data types collected.
3. Velocity – The speed at which the data is generated, collected and processed.

These characteristics of Big data were introduced by Doug Laney on 6 February 2001, who at that time was an analyst at META Group. He highlighted a few facts in his article.

1. Within a year, organizations would increase their usage of a centralized data warehouse to align internal and external practices for a more standard driven approach.
2. In a couple of years, data quality and integration problems will be eradicated by data profiling technologies (for generating metadata).
3. In a span of 4-5 years, data, document and knowledge management will amalgamate through a single schema indexed strategy.

Considering all these practices have been brought to realism over the defined timelines and not only grown but exponentially evolved over the past couple decades, it is fair to say that this marked the birth of big data in a formal manner.

More recently, ‘Veracity’ has been added as the fourth V in this structure which signifies quality and value of data at hand.

The above illustration clearly defines the major associations with Big data and what they signify.

## Big Data Technologies

There are various technologies associated with big data. They are various software utilities that are capable of analyzing, processing and extracting information from an extremely large complex dataset. From the initial step of collection, across multiple stages of cleaning, transformation and processing to the end of decision making through predictions, the tasks needed to be done are immensely complicated.

Top big data technologies are divided into 4 major categories –

1. Data Storage
2. Data Mining
3. Data Analytics
4. Data Visualization

### Data Storage

Hadoop – A framework developed by Apache Software Foundation to store and process datasets on a distributed environment i.e. a collection of systems breaking up chunks of tasks for faster processing.

MongoDB – In comparison to basic rigid schemas of a relational database system like SQL, MongoDB offers a larger flexibility making it capable of handling large datasets of varied datatypes.

### Data Mining

Elasticsearch – A search engine based on Lucene library.

Presto – It is an open source distributed SQL query engine capable of running interactive analytic queries on data sources.

### Data Analytics

Kafka – It is a distributed streaming platform with 3 major capabilities – publisher, subscriber and consumer.

Spark –Another hugely popular tech with in memory computing capabilities to deliver high speed and a generalised execution model to support various types of applications.

### Data Visualization

Tableau – A hugely popular data visualization tool used in business intelligence industry.

Plotly – Another tool used for creating graphs in an efficient manner.

## Tools and Technologies being used in this project

In this project, considering the dataset that I have at hand, which I will elaborate upon in the next section, I have decided to utilise Spark for our big data analysis procedure. The sheer amount of data available had already made sure that a proper and efficient big data tool needs to be relied upon, after that it was just the point of identifying the right one. Apache Spark has already been a proven technology for handling big datasets with ease and its distributed nature would have undoubtedly made the tasks at hand a lot easier.

The public availability of the dataset, and our usage of just a month’s data has prevented us from using a database system like MongoDB.

For the choice of programming language, Python and its integration in the Spark environment as ‘PySpark’, made going along with it the easiest decision.

## Familiarization with our dataset

The dataset that we are working on in this thesis is the government provided data of bus network of the public transportation sector of Dublin, the daily commute lifeline of the multi-million people residing in the capital city of Republic of Ireland.

It has been published by the Dublin City Council onto the data.gov.ie website, which is the centralised government repository for all generally available data for usage by subsequent parties in Ireland and internationally located. The license with which this data is available is ‘Creative Commons Attribution 4.0 International’ which allows the users of these datasets to

1. Freely share – copy and redistribute the material in any medium or format.
2. Freely adapt - remix, transform, and build upon the material for any purpose, even commercially.

We are under one important term by using this data. Attribution — We must give appropriate credit, provide a link to the license, and indicate if changes were made. We may do so in any reasonable manner, but not in any way that suggests the licensor endorses us or our use. And hereby I have done exactly that by providing an insight as to where I exactly obtained this dataset and have clearly defined usage requirements.

Now, moving on to the data structure. The dataset have been provided to us in csv format.

Each datapoint (row in the CSV file) has the following entries:

1. Timestamp
2. Line ID
3. Direction
4. Journey Pattern ID
5. Time Frame
6. Vehicle Journey ID
7. Operator
8. Congestion
9. Lon
10. Lat
11. Delay
12. Block ID
13. Vehicle ID
14. Stop ID
15. At Stop

Following is a table showing a sample of the dataset that we have at our hand.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Timestamp | LineID | Direction | JourneyPatternID | TimeFrame | VehicleJourneyID | Operator | Congestion | Lon | Lat | Delay | BlockID | VehicleID | StopID | AtStop |
| 1.357E+15 | 747 | 0 | 7470001 | 31/12/12 | 3493 | SL | 0 | -6.236852 | 53.425327 | -709 | 747006 | 40040 | 7411 | 0 |
| 1.357E+15 | 27 | 0 | null | 31/12/12 | 3883 | RD | 0 | -6.233417 | 53.342232 | 0 | 27017 | 33521 | 395 | 0 |
| 1.357E+15 | 40 | 0 | null | 31/12/12 | 2226 | HN | 0 | -6.27825 | 53.416683 | 0 | 40206 | 33142 | 6071 | 0 |
| 1.357E+15 | 7 | 0 | 71003 | 31/12/12 | 6106 | D1 | 0 | -6.231633 | 53.317768 | 0 | 7019 | 43004 | 3222 | 1 |
| 1.357E+15 | 747 | 0 | 7471001 | 31/12/12 | 3531 | SL | 0 | -6.254617 | 53.355484 | -454 | 747007 | 40039 | 1445 | 0 |
| 1.357E+15 | 56 | 0 | 056A1001 | 31/12/12 | 1830 | RD | 0 | -6.233183 | 53.342201 | 0 | 56001 | 33488 | 2379 | 0 |
| 1.357E+15 | 25 | 0 | 025A0001 | 31/12/12 | 2866 | CD | 0 | -6.296867 | 53.3475 | 0 | 25007 | 33604 | 4604 | 0 |
| 1.357E+15 | 747 | 0 | 7470001 | 31/12/12 | 3493 | SL | 0 | -6.238668 | 53.425789 | -687 | 747006 | 40040 | 7411 | 0 |
| 1.357E+15 | 27 | 0 | null | 31/12/12 | 3883 | RD | 0 | -6.2334 | 53.342232 | 0 | 27017 | 33521 | 395 | 0 |
| 1.357E+15 | 4 | 0 | null | 31/12/12 | 4243 | HN | 0 | -6.279 | 53.416683 | 0 | 4001 | 43043 | 7226 | 0 |

As can be seen in each of the above, all the columns denote some significance to the dataset at hand. Delay column will be one of the most important variable of interest to us because it is the one which will be predicted, due to its relation to the actual performance of a bus network, once the appropriate model has been trained.

For the grouping scenario for division of dataset, I will be starting at the Line ID, because logically it will be the one which ideally separates the performance of the buses within their own defined patterns.

Let’s take a deeper look at all the columns defined above.

* Timestamp – It denotes the exact timestamp at which that particular row was collected.
* LineID – It denotes the numerical ID of the Line to which that particular row belongs to.
* Direction – It informs us of the direction that particular bus is running in. It is denoted by a 0 if the bus is going from the source station towards the destination station or 1 if the bus is going in the opposite direction towards the source.
* JourneyPatternID – It is a division within the LineID. Various kinds of patterns are predefined within a LineID and JourneyPatternID is there to present that exact information.
* Timeframe – Denotes the timeframe within which that particular row of data falls into. (The start date of the production timetable - in Dublin the production timetable starts at 6am and ends at 3am).
* VehicleJourneyID – It denotes a given run within a journey pattern.
* Operator – This denotes the Bus Operator initials (not to be confused with the driver)
* Congestion – This denotes that whether at that exact timestamp, the bus is experiencing any congestion on the route it is currently running on. It is 0 if no, otherwise 1 if yes.
* Lon – Longitude value in WGS84 format. WGS stands for World Geodetic System and it is a standard used in cartography and satellite navigation including GPS. WGS84 is an earth centered, earth-fixed terrestrial reference system falling under WGS.
* Lat – Latitude value in WGS84 format
* Delay – This signifies the delay with which the bus is currently running. It is negative if the bus is running ahead of schedule.
* BlockID – This denotes a section ID of the journey pattern.
* VehicleID – As the name suggests, it denotes the ID of the vehicle within this particular row of data.
* StopID – As the name suggests, it denotes the ID of the stop that the bus is closest to (usually the next stop, or the current stop it is standing on), with respect to the current row of data.
* AtStop – It denotes the status value whether a bus is at the stop it is showing in the previous column i.e. StopID, or not. It is 0 if no, otherwise 1 if yes.

The dataset available to us, is present in this similarly structured set of 31 csv files, all containing data of each individual day in the month of January 2013. Due to the large volume of data combined in these files, this dataset at our hand can be defined as big data. Approximately 44 million records are present in these files combined.

We are going to apply a few big data technologies, which have been discussed before to this dataset in order to fulfil our objectives. All of this will comprise of our methodology. However it is always a beneficial method to delve deep into previous works that have been done in the same field to get an idea on how the industry demands information pertaining to such problems.

I went through a few such research papers, which are all related to analysis of data in the public transportation sector. The information presented in them ranges from a simple analytical project to some deep layered architectural approaches. I will be reviewing those works in the next chapter. Let’s have a look.

# Literature Review

When we talk about the public transportation sector, the data generation points are numerous and they provide us with amazing opportunities of capturing some really interesting sets of data. If we think about this topic from a real world perspective, the rising population has led the people seeking some mode of transportation to move about. Granted, in current times of COVID-19, this is something we really can’t stress upon, but when life does go back to normality, this will always be a pressing matter. In this technological age that we live in, we will undoubtedly set up some sort of a pipeline where data is collected from these modes of transportations for various reasons, be it for analytical purposes or just surveillance. Everything from the person using the transportation to the functioning attributes of it will be extracted.

## Introduction to the problem statement

Everyone out there can’t afford a vehicle of their own. A majority of inhabitants of residential demographics depend on the public transportations that have been provided by their city councils and facilitated to them as there rights. These modes range from land, aquatic and airborne and some time or the other we do need them.

Also, along with the affordability being one factor, the rising congestion problems due to increasing vehicle numbers on the roads is one such another problem. Managing this issue has a very successful tried and tested technique of focusing on proper development of public transportation which have the capacity of carrying more people with the similar on-street footprint. By ensuring there operating frequency and efficiency meets the demand of the urban population around them, we can somewhat curb this problem.

And not to mention, the rising problem of global warming and climate change is directly affected by the emissions made by the vehicles. The higher the number, the bigger the problem gets. Although, there have been numerous modifications to existing transportation technologies to reduce these emissions, a decrease in vehicle number will be of much greater help. That’s where public transportation can also chime in by carrying more amount of people with minimal regulated vehicles.

In the following data, we focus on the number of registered vehicles per 1000 people residing in different parts of the world. This data will be indicative for us to map out the rising number of vehicles in a country and how public transportation can be useful in reducing these numbers.

|  |  |
| --- | --- |
| **Country** | **Number of vehicle per 1000 population** |
| San Marino | 1263 |
| Monaco | 899 |
| New Zealand | 860 |
| United States | 838 |
| Iceland | 824 |
| Liechtenstein | 773 |
| Malta | 766 |
| Finland | 752 |
| Australia | 730 |
| Brunei | 721 |
| Switzerland | 716 |
| Canada | 685 |
| Guam | 677 |
| Luxembourg | 670 |
| Italy | 655 |

These above 15 nations rank the highest number of registered vehicles with respect to their total populations. These high numbers can definitely be reduced by introduction of higher focus on the public transportation.

As cited by researchers, in recent years there has been an immense advancement in the field on intelligent transportation systems. Due to this, the public transportation authorities across the world have developed efficient data acquisition techniques and workflows. The sweeping in of the big data era has not only helped this situation but introduced the analyst teams with amazing opportunities to mitigate forthcoming problems even before there arrival.

The availability of multiple sourced data in different forms, with cell phone data, GPS data, surveillance data, Wi-Fi data to name a few, these acquisition techniques had to be refined in an iterative manner to become efficient over time. However, the collection of these datasets was not an easy task, because traditionally, surveys has always been the go to methods for data collection at a large scale. The operation and planning of the public transportation systems depends a whole lot on these factors of demand and supply, and surveys were not able to effectively work in collection of the detailed spatial and temporal data we need.

That’s why the usage of smart transportation systems has had such an impact as it has had.

* The Geo-positioning data monitored via our consent through our phones and satellites
* Urban transportation smart cards for riders
* Occasional surveys
* Road sensors
* Radio frequency identification readers
* Social media feeds
* Cameras
* Microphones

These methods produce massive amounts of data without hindering the convenience of the system users. They agree to it once, and then the data is automatically collected according to the needs of the analytical companies who work continuously to acquire the relevant data that can assist them in their protocols. This data collected has a standard set of characteristics which are essential for big data processing. These include

* Continuous nature instead of discrete
* Wide coverage rather than a clustered set of survey data
* Contain comprehensive information
* Update dynamically

Since it was obvious to the world that prioritizing public transport was necessary, the reliability of them was the next concern. Resource allocation, network planning and frequency setting are parameters that substantially depend on the short term passenger demand. But why stop there? As we already have intelligent transportation systems in place which pour in streams of information by the second, dynamic decision making was possible for forecasting delays and subsequent scheduling of buses and tweaking there frequencies. Accuracy in these predictions will not only reduce the operation cost but also increase the service quality as a whole.

## Description of the publications

Following are a few related publications to the topic of Big Data processing and its impact on the public transportation sector.

### An Architecture for Big Data Processing on Intelligent Transportation Systems

In this publication at the Universidade Nova de Lisboa, Portugal, the researchers have aimed at proposing an ETL architecture for intelligent transportation systems for addressing an application scenario of dynamic toll charging on the highways.

The have utilised big data technologies to store and process large datasets from various sources provided by different highway operators. According to this research, the bigger challenge after collection of data is the processing of huge volumes of unstructured data for subsequent analytics. The data available is such cases is also full of inconsistencies due to missing data and un aligned data. The traditional approaches towards data analysis do not work efficiently on them, and this is the reason why they have utilized an ETL approach. There proposed architecture has three major capabilities.

1. Account for the available data quality.
2. Ability to mould itself according to already existing data standards under the Intelligent Transport System domain.
3. Furnish a robust and scalable storage system.

Elaborating a bit more on their architecture, Apache Spark lies at the soul of this. SparkSQL works simultaneously along with it for the processing needs and MongoDB has been used for data storage requirements. These technologies allow the parallel in-memory processing of data.

To successfully implement the structure defined by them for their workflows, CRISP-DM methodology has been followed by them in the project. It stands for Cross Industry Standard Process for Data Mining and is a highly regarded framework in data analysis. It consists of six basic steps –

1. Business understanding – Success criteria, objectives. From this perspective, the data was generated by interviews with several highway operators.
2. Data Understanding - Data collection, quality check and developing introductory insight into data.
3. Data Preparation – Cleaning and transformation of data to more closely suit our needs.
4. Modelling – Application of appropriate modelling technique after comparative selection.
5. Evaluation – Evaluation of obtained model in the previous step and decision making on how to interpret the results for collective benefit.
6. Deployment – Charting out use cases of the obtained knowledge and results.

Moving on to the actual model developed by the team, the primary objective set was to encourage drivers to use national highways by dynamically affecting toll prices of national roads during peak hours. This was done to restore a balance on the streets and manage financial ratios in infrastructure management in the long run, simultaneously increasing the quality of life for all inhabitants by preventing them to project themselves through the same roads causing congestion and pollution.

There were several factors in play while developing this toll pricing model.

1. Real time conditions of road networks
2. Quality of service
3. Road safety
4. Environmental data
5. Cost maintenance
6. Toll revenues
7. Congestion
8. Traffic events
9. Weather conditions

All these data values were collected and subjected through the Spark engine. MongoDB was used for storing and managing the data. The direct integration of this NoSQL database system with Spark and its scalability made it the obvious choice. After the pre-processing step of data by cleaning it, classification into two categories was done – tolling data and vehicle counting data, with both different in the essential meaning they supplied to our model.

Tolling data described the actual number of vehicles that were tolled within a 5-minute interval at a specific area of the highway in question. Vehicle Counting data described the total number of vehicles counted within a 5 minute interval at a specific area of the highway in question.

Since the data had already been subjected to transformation prior to being processed through the model, the trained model obtained was appropriate for future DATEX-II transformed data as well (DATEX-II was the data transformation standard used while data cleaning step).

The results obtained using the classic approaches took approximately double the time than that it took by using the Spark approach. The conclusion was driven that usage of big data technologies had improved efficiency and computational timelines. Since the data used now was of historical nature, plans were laid out to use it in the future on much larger datasets.

### Effective Bus Arrival Time Prediction based on Spark Streaming platform

This research touched upon the discussion we had before in our problem statement of the reliability of public transportation systems. In order for the operations to be smooth a lot of analysis has to be invested into perfecting it.

This publication at Inner Mongolia University proposed a particle filtering algorithm to formulate a bus arrival time prediction model. Going a further step in achieving higher optimization in the incepted model, the prediction error of using particle filter was reduced by accounting in latest bus speeds for collaborative data analysis. Apache Spark Streaming technique was used in the development phase, because achieving real time analysis was only possible by setting up a real time data stream and processing it on the go.

It was identified that the simultaneous processing of both real time and historical data was necessary in this project. Also, each 15 second ejected record was not required to be put through computation, and small batch processing of the real time stream was selected as a better approach.

The successful execution of particle filter algorithm is dependent on acquiring relatively recent data continuously and then correcting and updating the prediction results defined in the previous iteration with the real time computation on the recently acquired dataset. Since there is a pre-existing issue with particle filter algorithms on weight degradation in the iterative process, in order to avoid this problem of particle becoming less effective in representing the probability distribution of variable over time, they used a resampling method to curb this. Sequential Importance Sampling was selected for this step.

The two main problems in particle filter algorithms usage for a bus time prediction model, which are discussed in short in previous paragraphs, are –

1. Prediction error
2. Acquisition of records

The first problem has been mitigated by introducing average speed of the latest previous bus in the same road segment, and the second by construction of optimized observations.

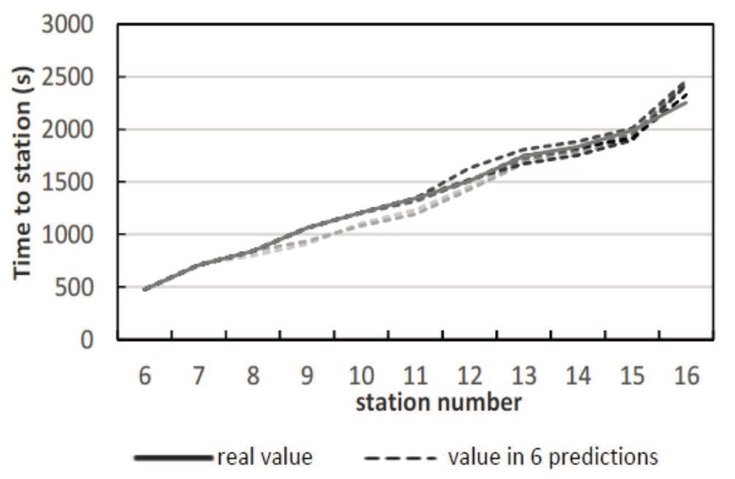
The core algorithm adopted in the model was as follows –

1. Initialise the basic parameters
2. Categorizing all pre-processed GPS records
3. Gathering all bus data which are not at terminals, i.e. in transit
4. Initialising the particle swarm and taking the sequence number of key road segments as 0 to obtain the data of latest two buses.
5. Computing number of key road sections where that particular bus trip is located at that particular timestamp.
6. Classifying the data for each bus trip according to the number of key road segments.
7. Computing the historical average speed of each bus trip on that key road segment.
8. Calculating the particle group expectation
9. Calculating the weight of each particle based on the weight calculation formula and putting it through a normalisation funnel.
10. Calculate degree of degradation and deciding whether to resample or not
11. Performing the resampling if necessary
12. Showing the final prediction result

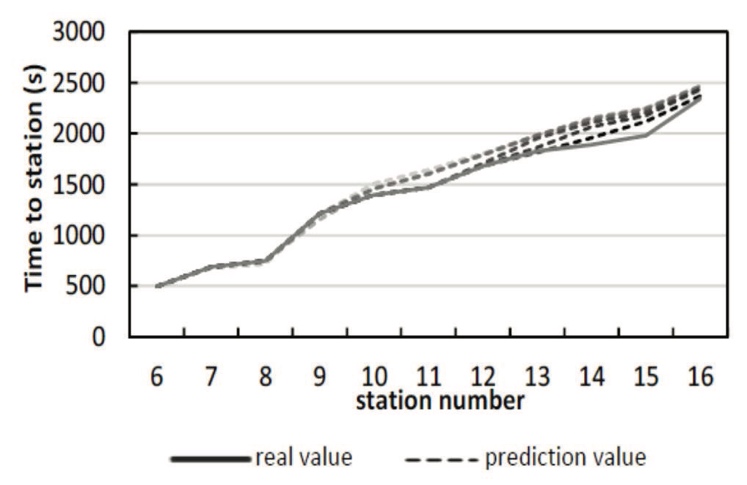
After the formulation of the final algorithm, the prediction model was designed. Since streaming in the form of batch processing had already been decided upon before, the streaming window was only set for particular amount of time. HDFS was used because of the efficiency and reliability of distributed architecture in such a case.

It was identified that as the number of particles go up, the distribution of the characterisation gets better. After setting the particle number to 2000, the prediction simulation was ran. This prediction process actually predicted the travel time in the next key road segment according to the driving condition of the previous key segment of the bus, on the constraint that the driving speed of the bus in these two sequence segments remains unchanged.

The results of this experiment were categorised for each bus into two periods – peak hours and non-peak hours. During the non-peak hours, the maximum absolute error is 78.16 seconds. During the peak hours the maximum absolute error is 270 seconds.



The above image is for the non-peak periods. The following is for peak hours.



In conclusion, the utilization of Spark platform, facilitated these researchers to effectively create a fruitful algorithm for arrival time prediction and played a key role in reducing the computational complexity considerably.

### Short Term Bus Passenger Demand Prediction

This research article is based on Time Series Model and Interactive Multiple Model Approach and was published at Shanghai Jiao Tong University. It focuses on short term passenger demand forecasting through the help of time series analysis. This will in turn assist in improving the dynamic bus scheduling techniques and their management. Also, accurately predicting the demand will also help in increasing operation efficiency and overall reliability amongst the passengers.

The objectives of this research were as tabulated below

1. Analyse characteristics of historical data with prime focus on stationarity, periodicity and volatility on differing time scales. After this, formulation of three time scales has to be done, one in a week, one in a day and last in 15 minutes which were constructed based on different characteristics.
2. Design separate prediction models which illustrate the characteristics of data to effectively predict weekly, daily and 15 minutes time series.
3. Dynamically amalgamate the prediction models estimations for further usage of outputting the hybrid predictions by applying IMM algorithm (Interactive Multiple Model) and evaluate it’s performance.

Time Series modelling on the data available to predict the bus passenger demand has been done in a hybrid manner. Both the data available in historical repositories and the real time extracted streams have been utilised in this modelling framework.

A four step process was followed in development of these time series

1. Based on the characteristics of historical data, three time series were made
2. After correlation analysis, each time series was further skimmed down to each having a separate weekly, daily and 15 min interval time series.
3. Adjust the time dependent transition probability matrix based on performance values of models based on historical data.
4. Combine the models’ predictions using IMM algorithm.

The data collected for a couple of months comprised primarily of the passenger boarding data. Since this was acquired through the smart card usage of the passengers, Intelligent Transport System standard apply to this research. Following is the demand graph for the month of August.

The x axis has the indexed number of the day and the y axis gives us an indication of the demand in trips per 15 minute intervals.

A screenshot of a cell phone

Description automatically generated

This historical data was then subjected through various analytical techniques to map out a set of descriptive statistics. To further analyse this historical data’s characteristics, three time series were constructed for the weekly, daily and 15 minutes intervals to illustrate upon the stationarity, periodicity and volatility of real time passenger demand.

Those time series are as illustrated below

A close up of a piece of paper

Description automatically generated

Above is the Weekly time series with the ARMA model (Auto-regressive moving average) implementation for constructing a stationary time series. Similarly, see below for the other time series developed in this research.

A screenshot of a cell phone

Description automatically generated

Above is the SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model fitting for the daily time series of first week of August.

Below is the comparison of the autoregressive conditional heteroskedasticity models for August 1st vs August 3rd.

A close up of a map

Description automatically generated

After these time series were developed, next came the implementation of the IMM algorithm and its validation. Its basic idea was to match the varying snapshots of data and with different models and zero in on usage of a specific model with minimal error.

The process followed while implementing this is as mentioned as follows

1. Calculate the mixed state and covariance at time t based on transition probability matrix.
2. Update the estimations for each model using Kalman filter algorithm and calculate the prediction residual and covariance with the input of real time passenger demand.
3. Update the probability for each model based on likelihood function of each model using prediction residual and covariance before.
4. Calculate the final estimation at time t by combining the updated estimations at time t.

To conclude, the real time prediction modelling facilitated by the usage of Time Series has been of prime importance in this research and has the helped in increasing the operational efficiency and reliability of the bus network. This performance evaluation has optimised real time decision making capabilities of the network administrators and furnished them a way to achieve high functioning dynamic scheduling for better usage throughout the city.

## Summary of main points

From the above literature review, we can set aside a few points that might be considered as the foundational steps in moving forward with deeper dives in works in this field. The related work mentioned above, in addition to the work I read online, has been my driving force to move ahead in this dissertation.

If we think about the points I have extracted the most knowledge out off in this literature review, they can be as listed below.

1. Data availability and variety of sources where we will accumulate it from is the first thing we need to concern ourselves with when working with a big data project. Sometimes the sheer numbers of sources make the process harder to keep a uniform nature in the incoming data and we need to formulate a plan on how to act in such a scenario.
2. Pre-processing techniques that are present in the industry have a set of standards that need to be followed in order to achieve proper data cleaning and transformations.
3. Sometimes the same source of data, may provide an irregular or alienated structure of incoming records and we need to be prepared for such a scenario with dynamic updates to our pre-processing techniques.
4. Public transportation sector is a gold mine when it comes to available streams of real time data. In order to prevent an anticlimactic situation at the point of maximum output, we need to select an appropriate model for our analysis algorithm development.
5. Big Data techniques need to subjected upon harmonized data to the platform we have decided to use and have a hybrid setting in their mind for maximum information gathering. A stream of real time data going through Kafka engine will readily be sometimes of reduced usage without the Spark engine working on historical data first.
6. When big data is in the question, the approach of using a traditional relational database may sometimes have to be pondered upon. A lot of works in this field have employed the capabilities of NoSQL database systems like MongoDB or PostgreSQL for their data storage needs.
7. Mining for data has various industry standard methodologies which are quite popular amongst miners in the data analytics industry. CRISP-DM is an example of this.

These few keynotes from the years of related works by noted scholars and professionals in this field has led us to inherit techniques which are proven to work in our cases and we move ahead with a strong foundation. Identifying out problem statement at the very beginning was essential and then walking through these publications has provided us with essential knowledge that will go a long distance in helping us achieve our benchmarked subobjectives.

## Contributions of these publications to Big Data in Public Transportation sector

The publications brushed through by me, along with the other sources cited by them, have done a whole lot of groundwork and made significant contributions to the Big Data domain as a whole. In the recent years, since the Big Data terminology has compounded within the Information Technology sector, the research works have helped pave the paths of aspirants like us.

Considering the Public Transportation sector specifically, the contributions have been marvellous. The advancement in methods to acquire data and moving those techniques from traditional survey based sheets of paper handed out to the travellers, which more often than not, they threw away because they simply didn’t have the time, to the automated collection of data using smart card in this era of Intelligent Transportation Systems is a boon.

The implementation of these big data techniques onto the vast transport datasets, and their appropriate usage in the publications above, is an example of how the public transportation sector is moving in the right direction, and with more relevant researches on the way, painstakingly trying to make all these tasks even simpler, we can only imagine the future ahead.

# Methodology

In this chapter we will talk about the process I am following for analysing the public transportation data available to me from Dublin City Council.