Predictive Analytic for Local Bus System

Abstract—**There are many factors that affect bus services such as Local events (concerts, theatrical performances or opera), weather and traffics. In a rapid rate today, huge quantities of data are being generated on a technology base landscape. Embedded in these big data are previously unknown and potentially useful valuable knowledge and information that can be discovered by data scientist/data mining. In this article we focus on open transit public data because public transport such as bus as means of transportation is vital part of many people's lives. Bus delays and disruptions are unfortunately inevitable as this can negatively impact daily schedules when relying on an available time table. Time is a precious resource so predicting parameters that may affect bus system planing is very important. Here, we will examine and compare our available data and on-time bus performances. This will follow with the extraction of useful and pertinent data from the huge available data via some basic knowledge of data mining and then make some predictions rules which may help answer some useful questions in order to obtain a reliable bus system timetable as well as identify important bus stops as presented in our short illustrated case.**

I. Introduction

In any situation, a human being holds a set of data from the environment in which he is located. This could be as much weather data, geographical data as any other type of data (stock prices, sports results and data collected from social networks). This ordered and analyzed data could provide useful information for scientific research or for the improvement of activities in different sectors.

Nowadays, many European countries such as Germany through the Mobilitethek-MDM (Germany's platform for data that keeps things moving), have joined the initiative to provide open data for scientific reasons. From mid-2022, the Mobilitethek platform will gradually merge MDM portal and the open data portal mCLOUD.

One type of available data sets in these open data portals is related to transit open data. Many cities in Germany provide information about their Bus transit systems, such as bus stop, disruption, delay and cancellation. Specifically, for this article, bus system planning is of particular interest. Information such as bus delay times are provided by few cities, but even fewer cities provide information about a bus arriving early. Among one of the agencies that provide bus delay and bus earliness in Germany, where bus deviation from the scheduled time is openly available is the Stuttgart transport and transit association - VVS (Verkehrs- und Tarifverbund Stuttgart).

By studying the system of the VVS agency, we find out that each bus operated in cities is equipped with an on-board computer and a GPS (global positioning system) whose role is to record the punctuality of a bus when leaving a bus stop. It records data on the arrival and departure time of the bus and at every bus stop. This recorded data is stored on the bus's on-board computer and is regularly collected and downloaded from the computer. This is then updated to open OPNV data. The resulting large amount of data can be characterized by the fact that:

These available data can be of different level of veracity due to the GPS parameter, transmission issues may contain errors of the registered longitudes and latitudes of locations.

According to the Mobilithek, on-time performance of every bus is affected by many factor. So, to regulate this it was classified as follows:

- A bus leaving at a time greater than **3 Minutes later** to the scheduled time is considered as a **late bus**.
- A bus leaving at a time **(1 to 3) minutes after** the scheduled time is considered as **on time**
- A bus leaving at a time greater than **1 Minutes earlier** than the scheduled time is considered **early**

From the Mobilithek and Open Data OPNV we can get available data about bus disruption, earliness, existing bus timetable, delays, cancellation and much more. So, it is logic for us to examine this and we will focus on analysing data linked to our available data given in our worksheet and also focus on whether is is feasible to assist an operator of a bus system in planing new bus stops and timetables. To do this, we will use the real time bus services in Germany based on data obtained from **Mobilithek and Open Data OPNV portals.**

II. Predictive analytic on Open data OPNV

One of the biggest problems with public bus transport is reliability. Thus, designing and developing a reliable method to predict whether a specific bus would arrive early, on time or late, by identifying the bus stops and the bus most likely to reach a specific destination is in high demand. It is therefore important to improve the reliability of the bus. To do this, we first examined the real-time data from the OPNV Open Data, which was divided into several collections and each made up of characterized attributes. These attributes have been collected and grouped as follows:

1) Agency Id which corresponds to agency name
2) Trip Id, which is the unique identifier in the data-set
3) Time (Departure time and arrival time)
4) Stop sequence Route number, which is a unique set of bus route number

5) Route name; this is the routes common name which corresponds to a route number.
6) Route destination, that indicate the terminus of a route and determines the bus direction.
7) departure date and time
8) Location, which are GPS coordinates (Longitudes and latitudes)
9) Day type which takes into account one of the three values:
    - Weekday i.e. Monday to Friday
    - Saturday or
    - Sunday

    This is so because we observed that bus run on different schedule among weekdays, Saturday and Sundays
10) Delays/Disruptions between the scheduled departure time and actual departure time from a stop:
    - A **positive delay** implies a bus leaving earlier than a scheduled departure time from a bus stop; whereas;
    - A **negative delay** implies a bus leaving later than a scheduled departure time.
11) Route Id/number

We effectuated a data comparison from the available data from the open data OPNV and the data provided in our work sheet such as:

- A complete set of digital street maps for the city of OpenStreetMap.
- Places where these events take place.
- Population density of the region with a grid of cells of 100 x 100 meters providing the number of persons living in each cell of the grid.
- Locations of existing bus stops.
- Existing bus timetables
- Number of persons entering and leaving a bus at every bus stop for all historical trips of all busses.

At the first stage, we used the attributes described above and we could capture/bin a set of aggregated on-time performance of open data OPNV which is directly or indirectly related to our data made available for our analysis:

1) Agency name
2) Route number
3) Route name
4) Route destination
5) Day type
6) Date from the scheduled time
7) Time from the time period of the scheduled time. This captured time was classified into one of the following time periods within a day:
    - Morning peak hours:05:00 - 09:00
    - Off peak hours : 09:00 - 16:00
    - Afternoon peak hours: 16:00 - 18:30
    - Evening peak hours: 18:30 - 22:30
    - Night peak hours: 22:30 - 05:00
8) Total number of on-time bus stops
9) Total list bus stops

At this Second stage, From the collected data set (Set of attributes and the set of on-time bus performance of passed data), we have selected the data useful for solving our problem:

- We collected the different bus numbers.
- We collected the different route numbers
- We extracted the *day of the week* because individuals behaviours varies during the weekday (Monday to Friday), Saturdays and Sundays.
- We extracted the *time* from the scheduled bus timestamp. This collected time was classified into five:
    a) Morning peaks hours
    b) Off peak hours
    c) afternoon peak hours
    d) Evening peak hours
    e) Night peaks hours
- We collected the total number of bus stops of the different scheduled bus and classified them as:
    a) Early bus stops
    b) Late bus stops
- We collected the set of data of bus delays/Disruption and we classified as follows:
    - More than 30 minutes later than the scheduled time : **Extremely Late**
    - More than 10 minutes but no more than 30 minutes later than the scheduled time : **Very late**
    - More than 3 minutes but no more than 10 minutes later than the scheduled time : **Slightly Late**
    - More than 1 minute earlier and no more than 3 minutes later than the scheduled time : **On-time**
    - More than 1 minute earlier and no more than 3 minutes later than the scheduled time : **Slightly early**
- We ignored the agency ID as it is a unique parameter.
- We ignored the Trip Id
- We ignored the route number at it is a direct function of the route name.

    In addition to this, we also process the available data given in our work sheet and classified it as follows:
    - On a map (Digital or paper), we identified the different points (longitudes and latitudes) where our events(concerts, theatrical performances or opera) will take place and segmented it in to different bus zones
        * Bus zone 1
        * Bus zone 2
        * Bus Zone 3 etc.
    - We set the population density in percentage i.e. The surface area was set to 10 000 meter squares. This surface area was segmented into grid cells which we estimated each grid cell to contain 10 percent of the population.
    - We classified the population into different categories:
        * **Areas of Extremely Low Density**, Areas having 100 persons per sq km.

    - **Areas of Low Density**, Areas having population density of 101 to 250 persons per sq km.
    - **Areas of Moderate Density**, This class includes those areas which are having 251 to 500 persons per sq km.
    - **Areas of High Density**, These are areas having population density of 501 to 1000 per sq km.
    - **Areas of Very High Density**,Areas having more than 1000 persons per sq km are termed as areas of very high population density.

III. Analytic Output

Base on the collected and processed data from OPNV open data and the available data at our disposal from the worksheet, along with some basis in machine learning (Data mining, data optimization, supervised learning and unsupervised learning), we could come out with a schema of our work:

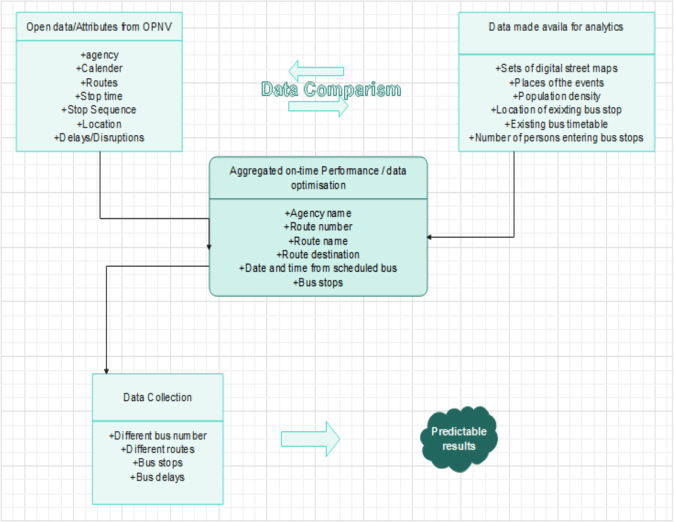

Fig. 1. Sketch of the processing

By linking the data made available to us, attribute data, Open data OPNV and those of Mobilithek, we could identify the following associative rules:

- Identify the **routes, directions and bus stops** were buses are likely to depart **early, on time or late**.
  Identifying the Routes, directions and bus stop will help us to also identify the dark points where bus stops are needed by comparing this areas with the areas and directions to reach a specific **concerts, theatrical performances or opera**.
  Identifying the frequency of extremely/very/slightly late, on time and early buses will help in optimizing the bus services after identifying the area where there is high rate of lateness.

- Identify and regrouping the days and time each bus depart or arrive a bus stop.
  This permits us to Identify the bus stop with more population congestion so as to know where more buses are required for similar trajectories. As there are many factors which may affect the time of departure/arrival of a bus, eliminating the parameter of the **required number of bus** for the same destination will help identify and solve the other parameters causing bus delays. This other parameters affecting bus delays could be identified by:
    - Identifying the relationship between the time and severity level of delays (Extremely/very/Slightly early, late or on time).
    - Identifying the relationship between the days of the weeks and severity level of delays (Extremely/very/Slightly early, late or on time).
    - Identifying the relationship between the routes and severity level of delays (Extremely/very/Slightly early, late or on time).
- Identify on the the population grid, which grid has the highest percentage of people taking a specific bus.

  Bringing the relationship between the population percentage of a grid and Identity the bus taken will permit us to identify the bus route and destination. With this data we will be able to Identify the mostly visited events (concerts, theatrical performances or opera).

- Identifying the relationship between the bus routes, bus stops and grid cells with highest population density will help optimise or identify new bus stops.

- Making a triangular research between the data obtained from each bust at a bus stop (Longitude and latitude) and the batching system when each individual gets into a bus will permit us identify the number of persons entering and leaving a bus at every bus stop.

IV. Case illustration

Let's imagine that there is a concert for young people on a Saturday at 6 p.m. in a start back out of town. People have to leave the city center to get there by bus. Our predictive analysis of local systems could allow us to:

- Identified the existing bus schedule to optimize bus frequency.
- Decide to increase other travel times due to heavy Saturday traffic in town at that time.
- Identify the roads less traveled so as not to have a lot of traffic and arrive on time; hence reducing bus delays.
- Increase other bus departure points by identifying where there are the most young people, by doing a triangulation. These would generally be places close to university centres.

V. Conclusion

The Open data OPNV and Mobilithek (mCloud MDM) provides to the world huge amount of open data about the transport services in Germany. Thanks to this available open data and data mining, we were able to adapt useful ways to unlock secrets and discover information useful for daily management.

In the case of this article it was about doing a predictive analytics to assist a bus operator in planning bus stops and bus schedules with these new bus stops. The linking of Open Data OPNV, Mobilithek and the data made available to us enabled us to respond to this problem by proposing some associative information.

References

[1] https://www.bmvi.de/SharedDocs/EN/Articles/DG/mobilithek.html

[2] https://www.mdm-portal.de/?lang=en

[3] https://www.bmvi.de/SharedDocs/EN/Articles/DG/mobilithek.html

[4] Development of an effective travel time prediction method using modified moving average approach," in KES 2009, Part I, pp. 130-138. By Chowdhury, N.K., R.P.D. Nath, H. Lee, J. Chang (2009)

[5] Bus Travel Time Predictions Using Additive Models by Matthías Kormáksson, Luciano Barbosa and Marcos R. Vieira

[6] https://www.opendata-oepnv.de/ht/de/willkommen

[7] Predictive analytics on open big data for supporting smart transportation services by Paul Patrick F, Balbina, Jackson C.R. Barkera, Carson K. Leung*a,, Marvin Trana, Riley P.