PA4: Data mining

CIV8760E - Transport data management

Frédérick Chabot & François Bélisle

November 24th, 2023

This fourth practical work focuses on data mining. Different tools will be used to explore different aspects of the same dataset. In particular, you'll be using Tanagra and QGIS for different analyses.

1 Data set

The dataset used for this work is the Société de transport de Montréal's (STM) set of planned schedules, published in GTFS (General Transit Feed Specification) format. Full documentation on the GTFS data format can be found on the Google Transit APIs documentation site.

1.1 Acquire dataset

Planned transit service data in GTFS format is available, for current data, on the sites of the various operators. For archived GTFS data, the aggregator Transit Feed is a good source. To obtain the data on this site, you need to search for the city where the head office of the transport company you're looking for is located. In the case of the metropolitan region, search for "Montreal" for STM data[1].

The data is then sorted by date. For this exercise, choose the timetable for October 18, 2023.

1.2 Data format

The GTFS format is quite complex, but for the purposes of this work, only a few tables and fields are required. Firstly, the following concepts will be useful:

"**Service**": Service is defined as the set of days on which a schedule, as presented to users, is in force. For example, the same schedule may be in effect from Monday to Friday or on weekends.

"**Route**": The route, defined in the *route.txt* file, represents what we call a transit line. In GTFS format, it mainly has attributes for user display.

"**Departure**": The "trip" corresponds to a departure on a line. A departure is associated with a timetable ("calendar"), an itinerary ("route", supported by its "shape"), a series of stops ("stops") and stop times ("stops_times").

[1] "Montreal" for EXO data (still classified under the AMT name), "Laval" for Société de transport de Laval (STL) data and "Longueuil" for Réseau de transport de Longueuil (RTL) data.

Here is a brief description of the files and attributes that will be useful to perform the assignment:

- *trips.txt* : lists the different departures of a route.
  - *trip_id* : unique identifier of the departure.
  - *service_id* : links the departure to a service day.
  - *route_id* : links the departure to its line.
  - *direction_id* : allows you to identify the direction of the line for a bidirectional line (to distinguish, for example, the 51 Édouard-Monpetit Est from the 51 Édouard-Monpetit Ouest).
  - *shape_id*: links the departure to its route.
  - *wheelchair_accessible*: category variable indicating whether a line is wheelchair accessible. The categories included are :
    - *0* or empty: no information is available on accessibility features for this route.
    - *1* : the vehicle used for this trip can accommodate at least one wheelchair user.
    - *2* : no wheelchair user can be accommodated on this trip.
- *stop_times.txt* : lists the sequence of times from a departure to the series of stops covered.
  - *trip_id* : links the transit time to a departure.
  - *stop_id* : links the passage time to a stop.
  - *stop_sequence* : sequential number of the stop.
  - *arrival_time* : gives the arrival time at this stop.
- *stops.txt* : Individual locations where vehicles pick up or drop off passengers.
  - *stop_id* : Used to identify a stop.
  - *wheelchair_boarding* : category variable indicating whether a stop receives wheelchair-accessible vehicles. The categories included are :
    - *0* or empty : no information is available on accessibility features for this stop.
    - *1* : indicates that at least some vehicles allow wheelchair access.
    - *2*: no wheelchair users can be accommodated at this stop.
- *shape.txt* : geomatic route covered by a departure. The sequence is presented as a list of points.

  - *trip_id* : links the itinerary to a departure.
  - *shape_pt_sequence* : sequential number of the point.
  - *shape_pt_lon* : x-coordinates of the point in the SCR to EPSG code 4326.
  - *shape_pt_lat* : y coordinates of the point in the SCR to EPSG code 4326.

- *calendar.txt* : lists the days when the service is in effect.
  - *service_id* : unique identifier of the service.
  - *monday* : binary field indicating whether the service is active on Monday (1 for active, 0 for non-active).
  - *tuesday* : binary field indicating whether the service is active on Tuesday (1 for active, 0 for inactive).
  - ...
  - *sunday* : binary field indicating whether the service is active on Sunday (1 for active, 0 for inactive).

2 Mandates

This practical assignment is divided into three different parts to put the skills you've learned into practice. For simplicity's sake, please consider only STM bus service. **You must remove the metro and its stations!**

Please note that for each question, a methodology must be given clearly and concisely. Unless otherwise stated, the methodology should enable the reader to reproduce your results, but without necessarily using the same tools. It is therefore not necessary to mention the software used or the functions used explicitly, but rather to describe the manipulations carried out and the transformations made to your data. That said, you're free to use any tool you like to manipulate the data, and to calculate and display your various results. However, as far as sections 2.2 and 2.3 are concerned, Tanagra must be used to generate the classes and analyze the explanatory factors using a decision tree.

2.1 Service indicators

This first assignment asks you to derive various indicators for the bus service offered by the STM. These different indicators will enable you to create a new database, which will be used for the mandates that follow.

2.1.1 Data preparation

From the recovered data sets, build a database deriving, for each route/line in the territory (by direction), the following service indicators: number of departures, commercial speed, inter-stop time, amplitude and accessibility. The section 2.1.2 explains how to construct these indicators. Please present an extract from this new database in the report, and share the Excel file in which all the indicators are to be found in the same sheet. For these different indicators, please justify the following elements:

1. For **inter-arrest time**, suggest a unique value for each route (e.g. average, median, mode, etc.). Justify your choice using a distribution analysis carried out on four lines (only 1 direction per line) taken arbitrarily from the sample.
2. For **accessibility**, propose a single value (numerical or binary) for each route. Justify your transformation of the category value (empty, 0, 1 and 2) into a numerical value.
3. In your final dataset, identify if there are any variables with outliers (e.g. unrealistic speeds) and justify either a correction or the exclusion of the relevant rows from the dataset.

Important: Treat bi-directional lines as two separate lines and do not treat the metro[2]. To remove the metro, simply remove all entries with a *route_id* of 1, 2, 4 or 5. Use only the services in force at least during the week (Monday to Friday) and if a line has more than one service in force, choose only one and justify it.

2.1.2 Indicator construction

Here's how to manipulate GTFS files to cross-reference different indicators:

- The **number of departures** is obtained by counting the number of unique *trip_id* for each route and direction.
- The **commercial speed** is obtained by calculating two intermediate indicators:
  - The total distance covered can be obtained by calculating the length of the poly-line reconstructed from the list of points contained in the *shape.txt* file.
  - The travel time can be calculated using the *stop_times.txt* file by comparing the values of the first and last entries in *arrival_time*.
- The **inter-stop time** is obtained for each departure and each stop using the *stop_times.txt* file and the *arrival_time* field.

[2]The metro is published in the "frequency" format, which differs from the format used to publish other lines. For simplicity's sake, we will therefore exclude it from the analysis

- The **accessibility** is obtained from the *wheelchair_accessible* and *wheelchair_boarding* fields for each *trip_id*.
- The **amplitude** is obtained by calculating the difference between the time of the first and last departure on the line.

For this mandate, please present a statistical summary of the database (indicators) in your report and comment.

2.2 Line segmentation

Using a segmentation algorithm and the calculated indicators, group the lines into classes of service. Justify your choice of the number of classes by analyzing class homogeneity and inter-class dispersion. For this analysis, you need to put the homogeneity and dispersion values per number of classes on the same graph. Then, please present some class statistics in a table. In addition to the appropriate figures and summary tables, present the classes using a map of the region (Montreal) with the various lines. Feel free to separate the classes into 2 or more maps if this helps to visualize them better.

Please comment and if any anomalies (numerical or cartographic) appear, please give potential justifications.

2.3 Class analysis

Enrich your dataset with the class assignments you have made following the previous analysis. Using a decision tree, analyze the explanatory factors linked to membership of each class (e.g. which indicators (variables/rules) help define a class).

3 Methods

This assignment is done with the same teams as the previous one. Please contact the laboratory teacher if you have any problems. A report in PDF format, no longer than **20 pages**, should outline the terms of reference for this practical work. The due date is **December 10th, 2023 at 11:59 pm**. The file must be submitted in electronic format on Moodle. You are welcome to submit any other files that will enable you to analyze your approaches in the various tools used.

The name of the file must include the following nomenclature: EQ{team number}_TP{number of TP}_{semester of study (A, H or E)}{year of study}. For example, "EQ01_TP1_A23".

Particular attention will be paid to the writing (english mistakes will be penalized as well as poor general organization of the work), counting for a total of 5% on the final grade.

Please consult the Writing guide for civil engineers available on Moodle in the Resources section.