PA3 : Statistical analyses and modeling

CIV8760E - Transport data management

Frédérick Chabot & François Bélisle

October 27th, 2023

This third practical assignment aims to have you apply different methods of statistical analysis of transportation data as well as work with a linear regression model.

1 Dataset

The data used are the travel times collected by the City of Montreal using Bluetooth sensors available on the open data portal. The data is collected for pre-defined origin-destination pairs, described in the road segments file. Travel time data are available for multiple years. The attributes of both data files are described on their respective pages.

2 Mandates

This practical work is divided into two main sections in order to practice different skills.

Note that for each question, a methodology must be given in a clear and concise manner. The use of SQL queries and coding is strongly recommended and if used, it would be appropriate to present them in your report. Otherwise, you are free to use any tool you wish to manipulate the data and make your graphs/tables.

2.1 Analysis of a road segment

For the following two sections, select **ONE** road segment to be studied that goes in both directions of travel (from A to B **AND** from B to A, e.g. "Sherbrooke_N06: Curatteau to SaintDonat" and "Sherbrooke_S07: SaintDonat to Curatteau") that is longer than 1000 m. Clearly indicate the selected segment. It is important to select a segment that contains data for each year (2016 to 2019).

2.1.1 Analysis of travel times and speeds

Describe the travel times in each direction with descriptive statistics and figures representing the entire distribution by year. Each year does not necessarily contain the same amount of information for each segment. Therefore, it is important in your analyses/interpretations to take this aspect into account. In addition, based on the previous analyses, propose a method for filtering/cleaning out the most outlier travel times. The filtering method should be applied to all data used in the following.

Finally, once the data have been filtered/cleaned, perform a goodness-of-fit test of the travel speeds in each direction on your section using the normal distribution (or any other distribution you feel is more appropriate). Use the method seen in class and describe the

various steps (you can compare with other automated methods on the various tools at your disposal such as R or Python). Justify your choice of theoretical law and explain clearly the initial hypothesis of your test and the result obtained. Give any intermediate results and comment on the reliability of such a test.

2.1.2 Analysis of observations and travel times as a function of time

Choosing a direction of travel, describe the temporal distribution of observation numbers and travel times, by month, day of the week (including weekends) and hour of day. Justify the statistics used to represent travel times. Make at least six graphs presenting the following information per year (2016 to 2019) and comment:

- Number of observations and travel times by month;
- Number of observations and travel times by day of the week;
- Number of observations and travel times per hour of the day by day type (weekday/weekend).

For this last point, I recommend that you make one graph for weekdays and a second for weekend days. This will facilitate visualization and analysis.

Finally, you need to study the correlation of travel times between two different years of your choice (include a figure). To do this, group the travel time values by hour of the day, for two different years. This will give you one time value per year for each hour of the day. Then, in an Excel chart (scatter plot), simply put the values for one year in "x" and those for the other year in "y". You can then add a linear regression line and add the coefficient. Note that there are similar ways of making a graph by adding a regression line, as well as calculating the coefficient of determination (or R^2) using Python.

2.2 Study of factors associated with travel conditions

For this last part, you need to consider only the year 2016 and five road segments in different Montreal neighborhoods. Briefly introduce these neighborhoods. Next, create a variable describing travel conditions. For example, the ratio of average speed on the segment and its speed limit (speed limits available on OpenStreetMap). You could also think of another way of deriving travel conditions on the sections. The latter will be your dependant variable, the variable to be described by the model.

In addition, before proceeding with model development, you need to enrich the data with weather conditions (Historical data - Government of Canada) and relevant road segment characteristics such as number of lanes, orientation (cardinal points), number of junctions,

distance to downtown, etc. Explain the steps involved in adding these new characteristics to the road segments selected for analysis. You'll need to have at least one temperature variable, one travel variable and one road segment variable in your model. This means you'll need to add several variables of each type and test them in the model.

In this way, you'll be able to study the temporal, weather and segment characteristics associated with travel conditions across all the selected segments using a linear regression model. It is important that you describe/illustrate your model development and verification process. Finally, comment on your final model (in particular, the significance and sign of the coefficients) and whether its estimation conditions are verified (in particular, concerning residuals).

3 Specifications

A report in PDF or Word format, not to exceed **20 pages**, must report the mandates of this fieldwork. The due date is **November 9, 2023 at 11:59 PM**. The file must be submitted in electronic format on moodle. If needed, you can, for example, submit an Excel file containing your different linear models and the related analysis figures.

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 two points (5%) on the final grade.

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