TP4 : Fouille de données et analyse spatiale

CIV8760 - Gestion de données en transport

Frédérick Chabot & Nicolas Saunier

10 novembre 2023 (Mise à jour le 20 novembre 2023)

Ce quatrième travail pratique porte sur la fouille de données. Différents outils seront mis à contribution afin d'explorer différents aspects d'un même ensemble de données. Notamment, vous devrez utiliser Tanagra et QGIS pour différentes analyses.

1 Jeu de données

Le jeu de données utilisé pour ce travail est l'ensemble des horaires planifiés de la Société de transport de Montréal (STM), publié au format GTFS (General Transit Feed Specification). La documentation complète concernant le format de données GTFS peut être consulté sur le site de documentation des APIs de Google Transit.

1.1 Acquérir le jeu de données

Les données du service planifié de transport en commun au format GTFS sont disponibles, pour les données courantes, sur les sites des différents opérateurs. Pour obtenir des données GTFS archivées, l'aggrégateur Transit Feed est une bonne source. Pour obtenir les données sur ce site, il faut chercher la ville où se situe le siège social de la société de transport recherchée. Dans le cas de la région métropolitaine, il faut donc chercher "Montréal" pour les données de la STM[1].

Les données sont ensuite classées par dates. Pour ce travail pratique, choisissez l'horaire en date du 18 octobre 2023.

1.2 Format de données

Le format GTFS est assez complexe, mais pour les besoins de ce travail, seuls quelques tables et champs sont nécessaires. Premièrement, les concepts suivants vous seront utiles :

"**Service**" : Le service est défini comme l'ensemble des journées où un horaire, tel que présenté aux usagers, est en vigueur. Par exemple, un même horaire peut être en vigueur du lundi au vendredi ou durant la fin de semaine.

"**Itinéraire**" : L'itinéraire, défini dans le fichier *route.txt*, représente ce que nous appelons une ligne de transport en commun. Dans le format GTFS, elle possède principalement des attributs servant à l'affichage usager.

[1] "Montréal" pour EXO (encore aujourd'hui classées sous le nom de l'AMT), "Laval" pour les données de la Société de transport de Laval (STL) et "Longueuil" pour les données du Réseau de transport de Longueuil (RTL).

"**Départ**" : Le "trip" correspond à un départ sur une ligne. Un départ est associé à un horaire ("calendar"), un itinéraire ("route", supporté par sa "shape"), une série d'arrêts ("stops") et des heures de passage à ces arrêts ("stops_times").

Voici une description sommaire des fichiers et des attributs qui seront utiles pour effectuer le travail :

- *trips.txt* : liste les différents départs d'un itinéraire.
  - *trip_id* : identifiant unique du départ.
  - *service_id* : permet de lier le départ à une journée de service.
  - *route_id* : permet de lier la départ à sa ligne.
  - *direction_id* : permet d'identifier la direction de la ligne pour une ligne bidirectionnelle (pour distinguer, par exemple, la 51 Édouard-Monpetit Est de la 51 Édouard-Monpetit Ouest).
  - *shape_id* : permet de lier le départ à son itinéraire.
  - *wheelchair_accessible* : variable de catégorie indiquant si une ligne est accessible aux fauteuils roulants. Les catégories incluses sont :
    - *0* ou vide : aucune information n'est disponible sur les aménagements en matière d'accessibilité pour ce trajet.
    - *1* : le véhicule utilisé pour ce trajet peut accueillir au moins un usager en fauteuil roulant.
    - *2* : aucun usager en fauteuil roulant ne peut être accueilli lors de ce trajet.
- *stop_times.txt* : liste la séquence d'heure de passage d'un départ à la série d'arrêts parcourus.
  - *trip_id* : permet de lier le temps de passage à un départ.
  - *stop_id* : permet de lier le temps de passage à un arrêt.
  - *stop_sequence* : numéro séquentiel de l'arrêt.
  - *arrival_time* : donne l'heure d'arrivée à cet arrêt.
- *stops.txt* : Individual locations where vehicles pick up or drop off passengers.
  - *stop_id* : permet d'identifier un arrêt.
  - *wheelchair_boarding* : variable de catégorie indiquant si un arrêt reçoit des véhicules ayant l'accès aux fauteuils roulants. Les catégories incluses sont :
    - *0* ou vide : aucune information n'est disponible sur les aménagements en matière d'accessibilité pour cet arrêt.

    - *1* : indique qu'au moins quelques véhicules permettent l'accès aux fauteuils roulants.
    - *2* : aucun usager en fauteuil roulant ne peut être accueilli par cet arrêt.

- *shape.txt* : itinéraire géomatique parcouru par un départ. La séquence est présentée sous forme de liste de points.
  - *trip_id* : permet de lier l'itinéraire à un départ.
  - *shape_pt_sequence* : numéro séquentiel du point
  - *shape_pt_lon* : coordonnées en x du point dans le SCR au code EPSG 4326.
  - *shape_pt_lat* : coordonnées en y du point dans le SCR au code EPSG 4326.
- *calendar.txt* : liste les journées où le service est en vigueur.
  - *service_id* : identifiant unique du service.
  - *monday* : champ binaire indiquant si le service est en vigueur pour la journée du lundi (1 pour actif, 0 pour non actif).
  - *tuesday* : champ binaire indiquant si le service est en vigueur pour la journée du mardi (1 pour actif, 0 pour non actif).
  - ...
  - *sunday* : champ binaire indiquant si le service est en vigueur pour la journée du dimanche (1 pour actif, 0 pour non actif).

2 Mandats

Ce travail pratique est divisé en quatre différentes parties afin de mettre en pratique les compétences acquises. À des fins de simplification, veuillez ne considérer que le service de bus de la STM. **Vous devez retirer le métro et ses stations !**

À noter que pour chaque question, une méthodologie doit être donnée de façon claire et concise. Excepté mention contraire, la méthodologie doit permettre au lecteur de reproduire vos résultats, mais sans nécessairement utiliser les mêmes outils. Il n'est donc pas nécessaire de parler du logiciel utilisé ou des fonctions utilisées explicitement, mais bien de décrire les manipulations faites et les transformations apportées à vos données. Ceci étant dit, vous êtes libres d'utiliser l'outil que vous voulez pour manipuler les données ainsi que pour calculer et afficher vos différents résultats. En revanche, en ce qui concerne les sections 2.2 et 2.3, Tanagra doit être utilisé pour générer les classes ainsi que pour analyser les facteurs explicatifs à l'aide d'un arbre de décision.

2.1 Indicateurs de service

Ce premier mandat vous demande de dériver différents indicateurs en ce qui concerne le service de bus offert par la STM. Ces différents indicateurs vous permettront de former une nouvelle base de données qui sera utilisée pour les mandats qui suivent.

2.1.1 Préparation des données

À partir des jeux de données récupérés, construire une base de données dérivant, pour chaque itinéraire/ligne du territoire (par direction), les indicateurs de service suivant : le nombre de départs, la vitesse commerciale, le temps inter-arrêts, l'amplitude et l'accessibilité. La section 2.1.2 explique comment construire ces indicateurs. Veuillez présenter un extrait de cette nouvelle base de données dans le rapport et partager le fichier Excel dans lequel doivent se retrouver dans la même feuille tous les indicateurs. Pour ces différents indicateurs, veuillez justifier les éléments suivants :

1. Pour le **temps inter-arrêt**, proposer une valeur unique pour chaque itinéraire (ex: moyenne, médiane, mode, etc.). Justifier votre choix à l'aide d'une analyse de la distribution effectuées sur quatre lignes (1 seule direction par ligne) prises arbitrairement dans l'échantillon.

2. Pour l'**accessibilité**, proposez une valeur unique (numérique ou binaire) pour chaque itinéraire. Justifiez votre transformation de la valeur de catégorie (vide, 0, 1 et 2) en valeur numérique.

3. Dans votre jeu de données final, identifiez s'il y a des variables avec valeurs aberrantes (par exemple des vitesses irréalistes) et justifiez soit une correction ou l'exclusion des lignes concernées du jeu de données.

Important : Traitez les lignes bidirectionnelles comme deux lignes distinctes et ne traitez pas le métro[2]. Pour retirer le métro, il suffit de retirer toutes les entrées dont le *route_id* vaut 1, 2, 4 ou 5. N'utilisez que les services en vigueur au moins la semaine (lundi au vendredi) et si une ligne a plus d'un service en vigueur, choisissez en un seul et justifier.

2.1.2 Construction des indicateurs

Voici comment manipuler les fichier GTFS pour croiser les différents indicateurs:

- Le **nombre de départs** est obtenu en comptant le nombre de *trip_id* uniques pour chaque itinéraire et chaque direction.

[2]Le métro est publié sous la forme "fréquence" qui diffère du format utilisé pour la publication des autres lignes. Pour fins de simplification, nous l'exclueront donc de l'analyse.

- La **vitesse commerciale** est obtenue en calculant deux indicateurs intermédiaires :
  - La distance totale parcourue peut être obtenue en calculant la longueur de la polyligne reconstruite à partir de la liste de points contenue dans le fichier *shape.txt*.
  - Le temps de trajet peut être calculé à l'aide du fichier *stop_times.txt* en comparant les valeurs de la première et de la dernière entrée de *arrival_time*.
- Le **temps inter-arrêt** est obtenue pour chaque départ et chaque arrêt grâce au fichier *stop_times.txt* et le champs *arrival_time*.
- L'**accessibilité** est obtenue grâce aux champs *wheelchair_accessible* et *wheelchair_boarding* pour chacun des *trip_id*.
- L'**amplitude** est obtenue en calculant la différence entre l'heure du premier et du dernier départ sur la ligne.

Pour ce mandat, veuillez présenter un sommaire statistique de la base de données (indicateurs) dans votre rapport et commenter.

2.2 Segmentation des lignes

À l'aide d'un algorithme de segmentation et des indicateurs calculés, effectuez un groupement des lignes en classes de service. Justifiez votre choix du nombre de classes en analysant l'homogénéité des classes et la dispersion inter-classe. Pour cette analyse, vous devez mettre dans un même graphique les valeurs d'homogénéité et de dispersion par nombre de classes. Ensuite, veuillez présenter certaines statitistiques des classes dans un tableau. En plus des figures et des tableaux résumés appropriés, présentez les classes à l'aide d'une carte de la région (Montréal) avec les différentes lignes. N'hésitez pas à séparer les classes en 2 ou plusieurs cartes si cela permet de mieux les visualiser.

Veuillez commenter le tout et si des anomalies (numériques ou cartographiques) semblent apparaitre, veuillez donner de potentielles justifications.

2.3 Analyse des classes

Enrichissez votre jeu de données avec l'assignation des classes que vous aurez effectués à la suite de l'analyse précédente. À l'aide d'un arbre de décision, analysez les facteurs explicatifs liés à l'appartenance à chaque classe (par exemple, quels indicateurs (variables/règles) permettent de définir une classe).

2.4 Analyse spatiale de l'accessibilité des arrêts

Cette question vise à étudier la distribution spatiale des arrêts accessibles aux personnes à mobilité réduite. À l'aide des outils descriptifs vus en classe (centralité, dispersion, ellipse et enveloppe) et du champ wheelchair accessible du fichier *stops.txt*, comparez les distributions spatiales des arrêts accessibles et non-accessibles. Incluez un test statistique (en utilisant le découpage de votre choix, par pavage régulier ou limites administratives).

Assurez-vous de présenter les différents éléments sous forme de carte et de décrire les démarches.

3 Modalités

Ce travail se fait avec les mêmes équipes qu'au précédent travail. Veuillez contacter le chargé de laboratoire en cas de problème. Un rapport au format PDF, ne dépassant pas **20 pages**, doit faire état des mandats de ce travail pratique. La date de rendu est le **30 novembre 2023 à 23h59**. Le fichier doit être déposé en format électronique sur Moodle. Vous pouvez déposer tout autre fichier permettant d'analyser vos démarches dans les différents outils utilisés.

Le nom du fichier doit porter la nomenclature suivante : EQ{numéro d'équipe}_TP{numéro du TP}_{semestre d'étude (A, H ou E)}{année d'étude}. Par exemple, "EQ01_TP1_A23".

Une attention particulière sera portée à la rédaction (les fautes de français seront sanctionnées tout comme une mauvaise organisation générale du travail), comptant pour un point (5%) sur la note finale.

Veuillez consulter le Guide de rédaction pour ingénieur civil disponible sur Moodle à la section Ressources..