UNIL | Université de Lausanne
Faculté des géosciences
et de l'environnement

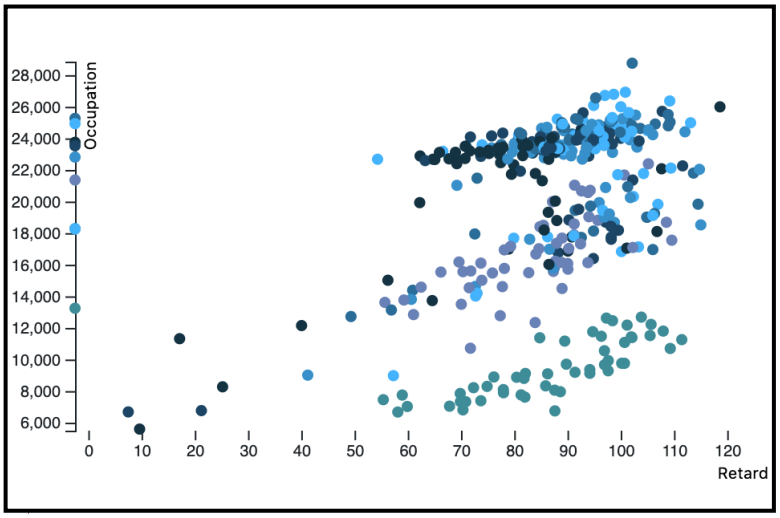
Master of Science in Geography

Visualisation d'un jeu de données

Analyse des retards sur deux lignes des Transports Lausannois

Romain Loup

Sous la direction du Dr. Christian Kaiser
Et l'expertise de Raphaël Bubloz



Relation entre la charge et le retard

Résumé

L'analyse d'un réseau de transport et l'amélioration de ses performances est une problématique récurrente dans le domaine de la mobilité. L'informatisation de la société et la création de larges jeux de données offrent de nouvelles possibilités dans l'analyse des transports. S'inscrivant dans ce courant, ce mémoire utilise des données issues des Transports Lausannois (tl) et tend grâce à une représentation graphique de visualiser les retards des bus et d'en comprendre les causes.

La recherche s'articule en trois parties : la mise en forme des données, leur exploration et leur visualisation. Une application interactive de visualisation est créée afin que l'utilisateur puisse, en choisissant divers filtres proposés, visionner sur un seul écran les caractéristiques des voyages effectués par les tl. Les résultats affichés permettent une analyse simple et rapide de la situation.

Mots clé : transports publics, données, exploration de données, visualisation, analyse, retards

Abstract

Analyzing a transport network and improving its performance is a recurring statement in the field of mobility. Informatization of society and the creation of large data sets offer new possibilities in transport analysis. As part of this trend, this master thesis uses data from Transports Lausannois (tl) and tends, thanks to a graphic representation, to visualize bus delays and to understand the causes.

The research is organized in three parts: the shaping of the data, their exploration and their visualization. An interactive viewing application is created so that the user can, by choosing the various proposed filters, view on a single screen the characteristics of the trips made by the tl. The results displayed allow a quick and easy analysis of the situation.

Keywords: public transportation, data, exploratory data analysis, visualization, analysis, delays

Remerciements

Je tiens à témoigner ma reconnaissance à toutes les personnes qui grâce à leur soutien m'ont permis d'élaborer ce mémoire.

Je souhaite tout d'abord remercier Christian Kaiser, mon directeur de mémoire, pour ses cours riches et intéressants et son intérêt pour la géovisualisation qu'il a su me transmettre. Son soutien a été important pour conception de ce travail. Un grand merci également aux Transports Lausannois, plus particulièrement à Michel Joye, Christophe Jemelin et Irina Karpushova pour leur confiance et la fourniture de leurs précieuses données.

Merci aussi à Kerria Grize pour ses relectures et notre partage d'idées tout au long de ce mémoire et à Raphaël Bubloz pour son aide généreuse lors de la conception du code et pour son rôle d'expert. Tous les deux ont été un soutien moral apprécié et les nombreuses pauses partagées ensemble au Géopolis des moments de plaisir.

Enfin merci à ma famille pour son soutien et plus particulièrement à ma mère pour ses relectures et suggestions. Un grand merci aussi à Vania pour ses relectures, sa présence et son soutien.

Je remercie finalement la communauté de développeurs pour la qualité des nombreuses informations mises en ligne.

Table des matières

Résumé ..... **ii**

Remerciements ..... **iii**

Introduction **1**

1. Problématique **4**

1.1. Plan 5

2. Cadre conceptuel et théorique **6**

2.1. Géographie quantitative 6

2.2. Géographie des transports 10

2.3. Exploration de données 12

2.4. Visualisation de données 19

2.5. Time series 24

3. Méthodologie **26**

3.1. Délimitation du territoire 26

3.2. Lignes 2 et 18 27

3.3. Jeux de données 31

3.4. Logiciels et langages de programmation 35

4. Cadre opératoire **38**

4.1. Mise en forme des données 38

4.2. Exploration des données 39

4.3. Visualisation 39

5. Résultats **43**

5.1. Mise en forme des données 43

5.2. Exploration des données 45

5.3. Visualisation 57

6. Discussion **69**

Conclusion **72**

Bibliographie **74**

Annexes **81**

Table des figures

Figure 1 : Valeurs d'outliers multivariées (Aggarwal, 2015) 7
Figure 2 : Attributs des formes et des arrangements (Pointet, 2007) 8
Figure 3 : Sept variables visuelles de Bertin (Bertin, 1983) 9
Figure 4 : Caractéristiques des données géospatiales (Kraak & Ormeling, 2010) 10
Figure 5 : Quatre paradigmes de la science (Hey, 2009) 14
Figure 6 : Détection des anomalies dans un jeu de données à deux dimensions (Chandola et al., 2009) 17
Figure 7 : Composantes clés pour la détection d'anomalies (Chandola et al., 2009) 18
Figure 8 : "Rose diagram" de Florence Nightingale schématise la réduction du nombre de morts dans les hopitaux militaires à Scutari grâce aux changements qu'elle a effectué (Spence, 2014) 20
Figure 9 : Carte du métro de Londres avant 1931 (Spence, 2014) 21
Figure 10 : Nouvelle carte du métro de Londres de Harry Beck en 1931 (Spence, 2014) 21
Figure 11 : Plan du réseau des Transports Lausannois (TL, 2020) 22
Figure 12 : Quatre jeux de données ayant les mêmes caractéristiques générales (Anscombe, 1973) 23
Figure 13 : Représentation graphique des quatre jeux de données d'Anscombe (1973) 23
Figure 14 : Dashbord de la propagation du COVID-19 créé par ArcGIS, consulté le 28 mai 2020 24
Figure 15 : Tracé de la ligne 2 28
Figure 16 : Tracé de la ligne 18 30
Figure 17 : Fonctions du JavaScript, CSS et HTML 36
Figure 18 : Premières idées de visualisation 40
Figure 19 : Filtres de l'application, première version 41
Figure 20 : Schéma de compréhension du processus de développement 42
Figure 21 : Création des tables « tjm » et « tps » en SQL 44
Figure 22 : Une des étapes nécessaires à l'élaboration de la colonne du retard depuis la base « tps » 45
Figure 23 Répartition du retard pour 90% des données centrales, sur les lignes 2 et 18 47
Figure 24 Répartition du retard pour 90% des données centrales, sur la ligne 2 48
Figure 25 Répartition du retard pour 90% des données centrales, sur la ligne 18 48

Figure 26 : Boxplot de la charge par jour 51
Figure 27 : Boxplot du retard par jour 51
Figure 28 : Fonction de densité de la charge par jour 52
Figure 29 : Test de normalité de la répartition de la charge par jour 52
Figure 30 : Fonction de densité du retard par jour 52
Figure 31 : Test de normalité de la répartition du retard par jour 53
Figure 32 : Charge des lignes 2 et 18 54
Figure 33 : Retard médian des lignes 2 et 18 54
Figure 34 : Retard par rapport à la charge, régression sur le total 55
Figure 35 : Retard par rapport à la charge, régressions différenciées entre les lignes 2 et 18 56
Figure 36 : Retard annuel moyen selon la météo 57
Figure 37 : Page d'accueil de la version par jour, avec le menu des filtres ouvert 58
Figure 38 : Page d'accueil de la version par arrêt, version 1 58
Figure 39 : Lancement de la visualisation 59
Figure 40 : Jours et couleurs correspondantes 59
Figure 41 : Correspondance entre les différents graphiques et tooltip 60
Figure 42 : Rectangle de sélection 61
Figure 43 : Exemple de fichier téléchargé via l'application 61
Figure 44 : Données de charge et retard sur plus de deux mois 62
Figure 45 : Retard représenté par des symboles proportionnels sur chaque arrêt 62
Figure 46 : Page d'accueil de tlDataViewer 63
Figure 47 : Données complètes de l'année 2018 avec la sélection d'un outlier 64
Figure 48 : Charge et retard, retards maxima de 5 minutes 65
Figure 49 : Charge et retard, jours ouvrables 66
Figure 50 : Charge et retard, jours ouvrables dont le retard par arrêt est de moins de 5 min 66
Figure 51 : Charge et retard, jours ouvrables hors des vacances scolaires dont le retard aux arrêts est de moins de 5 minutes 67
Figure 52 : Charge et retard pour l'arrêt du Flon, Ligne 18 68

Introduction

L'urbanisation rapide de ces dernières décennies est un enjeu majeur pour notre environnement, notre espace de vie et les villes d'aujourd'hui. La façon dont nous voyageons à travers les villes se modifie très vite et l'offre des transports publics est toujours plus variée pour répondre aux besoins et à la demande du public. Le maintien d'une clientèle et l'attraction de nouveaux usagers peuvent être améliorés par l'augmentation de leur satisfaction (Transportation Research Board of the National Academies, 1999). Les infrastructures de transport sont amenées à évoluer et doivent se développer tout en s'adaptant au tissu urbain existant dans le but de satisfaire au mieux les utilisateurs en tant qu'individus, ceci en prenant en compte les capacités des autorités à aménager la ville et ses transports publics (Abenoza, Cats, & Susilo, 2017). La stratégie d'investissement des autorités dans les transports a un impact immense sur la façon de se déplacer. Les trajets ne seront pas effectués de la même manière si une ville décide de miser sur le transport individuel ou sur les transports publics par exemple. Par ailleurs, une fois qu'un réseau est établi, les décisions intermédiaires de gestion sont tout autant importantes pour un fonctionnement optimal des infrastructures (Magnanti & Wong, 1984).

Parallèlement au développement des transports, l'informatisation de la société a permis le stockage d'une masse énorme de données, accumulée chaque jour dans les serveurs informatiques des diverses entités publiques et privées (Goodchild, 1992). Ces données peuvent provenir de sources très variées, telles que des traces GSM, des antennes WI-FI ou encore des capteurs implantés dans les transports publics. Cette production quotidienne de données, qui forme en quelque sorte les *big data*, représente aujourd'hui un outil incontournable pour gérer et comprendre l'organisation du territoire et des moyens de transports (Zhong, Huang, Müller Arisona, Schmitt, & Batty, 2014). En effet, cette récolte d'information contribue notamment au développement d'une meilleure communication entre les usagers et le transporteur, à l'amélioration des réseaux de transport, à l'identification de potentiels encore non découverts ainsi qu'à la réduction du coût des transports.

Bien que ces données soient une source d'information inestimable dans la société actuelle, elles doivent tout d'abord être traitées et souvent adaptées afin de mettre en évidence tout leur potentiel. Les données peuvent être traitées selon des méthodes statistiques et économétriques (Washington, Karlaftis, & Mannering, 2011). De plus, les données sont souvent collectées en masse sans objectifs d'analyse clairement définis préalablement. De ce fait, lors du traitement

des données disponibles, le travail d'adaptation peut se révéler conséquent pour en tirer des informations pertinentes (Bivand, 2010).

Par ailleurs, une fois traitées, les données doivent également être mises en forme de façon à être comprises par le public cible. Un important travail de visualisation doit être effectué afin d'utiliser de manière optimale le potentiel de chaque jeu de données (Vuillemot, 2010). La visualisation est le résultat visuel explicite de l'important traitement des données. Ces étapes sont déterminantes car la façon dont est mise en forme l'information a une influence sur son interprétation et donc sur les prises de décisions y relatives.

Les agences de transport ainsi que les gouvernements font face à des problèmes stochastiques et doivent donc déterminer comment évaluer la performance d'un réseau et comment l'améliorer tout en suivant les demandes des usagers. En Suisse, les transports publics ont pour objectif principal de transporter dans les meilleures conditions un maximum de passagers, de limiter les utilisateurs délaissés par le réseau et de maximiser le bénéfice de la moyenne des voyageurs en conservant leur avantage écologique (Office fédéral des transports, 2019).

Or, l'optimisation du réseau et de la satisfaction de l'usager passent également par une compréhension et une évaluation des retards, afin d'améliorer la ponctualité des transports. Selon une enquête menée à Genève auprès d'usagers des transports publics, la ponctualité représente le deuxième critère de qualité le plus important après la sécurité (Jemelin, 2008). En effet, un réseau ponctuel influence directement la confiance de l'usager et lui permet d'emprunter sereinement les transports publics pour ses déplacements quotidiens. La qualité et la facilité des déplacements sont primordiales pour diriger et favoriser le choix de l'usager vers les transports publics et permettre de décongestionner les villes du trafic individuel motorisé. Ainsi, le développement de la qualité du service de transports publics contribue également à la promotion de la mobilité durable.

Par ailleurs, l'optimisation du réseau et la réduction des retards dépendent non seulement des structures de transport préexistantes, mais également de diverses contraintes, qu'elles soient géographiques, politiques ou économiques. Les services sont en général repartis selon un modèle gravitaire, c'est-à-dire où les densités d'habitations, d'emplois et de services sont les plus élevées (Bougheas, Demetriades, & Morgenroth, 1999). Selon Gastner et Newman (2006), la densité optimale des services doit être proportionnelle à la densité de population à la puissance deux tiers. Cette densité peut se traduire en termes de population, mais aussi d'emplois ou d'activités humaines. Les lignes de transports et les arrêts suivent probablement le même type de

répartition, mais le facteur pourrait être un peu différent. Par ailleurs, chaque ville possède sa propre écologie, par conséquent les transports publics ne peuvent pas être uniformes (Park, Burgess, & McKenzie, 1984). Par exemple, la ville de Lausanne compte de nombreuses collines, un centre historique hérité du Moyen-âge et est bordée dans sa partie sud par le Léman. Ces spécificités rendent l'exploitation des transports publics compliquée et peuvent avoir une influence importante sur l'aménagement des lignes de transports et potentiellement sur la ponctualité des véhicules.

Les éléments discutés précédemment mettent en avant les enjeux vastes du transport en commun. C'est pourquoi il est important de formuler clairement la problématique ainsi que de délimiter la zone d'étude. Les choix d'aménagements sont effectués par les politiques publiques. Pour le Canton de Vaud (Etat de Vaud, 2020), les aménagements urbains et urbains en site propre[1] sont financés respectivement à maximum 50% par le canton et 50% par la commune et 70% par le canton et 30% par la commune. Il est donc primordial de comprendre l'intention de l'institution dans le domaine des transports, afin d'apporter des réponses pertinentes aux problèmes rencontrés.

Ce travail porte sur l'acquisition et la mise en forme de données brutes fournies par les Transports Lausannois (tl). L'objectif de ce projet est d'extraire des informations sur les retards d'exploitation de cette compagnie et de mettre en évidence des structures répétitives en suivant une démarche exploratoire. Grâce à cette mise en forme, une application de visualisation des données est développée dans le but de rendre les données lisibles et exploitables sans être statisticien ni géographe. Les données spécifiques traitées sont celles des tl, auxquelles ont été adjointes des données météorologiques qui sont également un facteur important à intégrer dans l'analyse des retards des transports publics lausannois. Bhattacharya et al. (2013) affirment d'ailleurs que l'occupation des transports peut être prédite par plusieurs facteurs comme le jour de la semaine et l'heure mais aussi la météo.

[1] Site propre : voie réservée à l'exploitation d'un bus (Xu & Zheng, 2012)

1. Problématique

Ce travail de recherche se base sur les données des Transports Lausannois des lignes de bus numéros 2 et 18 durant l'année 2018. Il vise à comparer les données de circulation effectives récoltées par les tl lors des trajets des bus aux horaires théoriques de ces derniers. D'autres données issues de MétéoSuisse viennent compléter ce jeu de données. Cette démarche a pour but de créer un support montrant la répartition des retards en fonction de chaque arrêt et de la charge des bus. Il s'agit aussi de déterminer si la forme actuelle de la récolte de données se prête à une analyse ou si les données pourraient être articulées différemment afin d'en faciliter l'utilisation. Par ailleurs, il est également question dans ce travail de définir ce qu'est un retard et à partir de quand il a une influence sur la performance du déplacement. Ces résultats pourraient aider les tl à effectuer d'éventuelles améliorations de leur réseau et de leur saisie de données.

De ce fait, l'objectif de ce travail est d'offrir un outil pertinent de visualisation des retards sur le réseau des tl qui pourrait potentiellement appuyer certaines décisions de planification ou d'amélioration des lignes de transport.

La question de recherche principale est la suivante :

- Sous quelle forme un outil de visualisation doit-il être créé pour permettre une visualisation pertinente des occupations et retards des bus ?

Plusieurs questions secondaires peuvent être formulées afin de guider ce travail. Elles portent sur la mise en forme et l'exploration des données qui permettent de mettre en place des outils nécessaires à la visualisation des retards des lignes numéros 2 et 18 des tl. Elles sont les suivantes :

- De quelle manière les données d'exploitation récoltées par les tl peuvent-elles être mises en valeur pour des analyses statistiques ?
- Dans quelle mesure est-il possible d'identifier et de visualiser la répartition des retards en fonction de divers facteurs comme l'occupation ou la météo ?
- Comment différencier les retards systématiques et les retards occasionnels ?

Afin de répondre à ces questions de recherche, il est nécessaire de traiter les données en fonction du choix de la visualisation de ces dernières. En effet, l'objectif est que les données de base puissent être transformées en visualisation parlante et être comprises par des personnes non

initiées à l'analyse de données. Il s'agit aussi de s'adapter au jeu de données disponible et d'en tirer le maximum. La recherche se focalise sur les retards systématiques plutôt que sur les retards occasionnels et aléatoires, liés par exemple à une manifestation. En effet, les retards systématiques et leur régulation représentent un enjeu majeur pour les sociétés de transports publics.

1.1. Plan

La première partie de ce travail de recherche se base sur la mise en forme des données, processus qui débute par l'acquisition des données jusqu'à l'utilisation à proprement parler de ces dernières. En effet, un grand travail de mise en forme est nécessaire afin de tirer des enseignements sur la répartition des retards.

La deuxième partie de ce travail traite de l'exploration des données, de l'analyse des retards ainsi que de la façon de les visualiser. La visualisation des données sur les retards a pour objectif de les rendre plus lisibles, accessibles et compréhensibles. Cette partie est envisageable grâce au jeu de données mis en place précédemment.

Finalement, une application interactive est construite afin de proposer aux utilisateurs une interface qui leur permet de choisir différents filtres dans le but de visualiser un phénomène précis. L'occupation des bus, leur retard ainsi que diverses données météorologiques sont confrontés et forment des visualisations liées mettant en avant le retard des bus.

2. Cadre conceptuel et théorique

Afin de comprendre comment procéder pour l'analyse d'un réseau de transports publics, il est important de définir un cadre théorique clair qui présente les concepts utilisés tout au long de cette recherche.

Pour analyser un réseau de transports publics, de nombreux domaines touchant la géographie ainsi que d'autres champs scientifiques doivent être parcourus. Au-delà de la notion de transports à proprement parler, une large palette de disciplines doivent être prises en compte, incluant bien sûr l'ingénierie des transports, mais aussi l'urbanisme, l'économie, la logistique, la psychologie, la sécurité, le droit et encore la théorie des consommateurs (Washington et al., 2011). Tous ces domaines ne sont évidemment pas traités dans ce mémoire, mais cela montre que l'enjeu dépasse la simple notion de transport et de géographie.

Les données et leur traitement proviennent du domaine quantitatif, comme l'analyse de données géographiques, ou la cartographie, mais aussi de la géographie des transports et de la mobilité. L'informatique et les statistiques sont aussi des champs importants pour le traitement de données ainsi que la visualisation qui permet de traduire les résultats pour une personne non initiée.

2.1. Géographie quantitative

L'analyse spatiale apparaît dès les années 1950 comme un lien entre les statistiques et la géographie (Goodchild & Haining, 2004). Délaissée jusque-là par les géographes, cette discipline fait son retour vers la fin du XX^{ème} siècle (Johnston, 1997). Les méthodes de la géographie quantitative moderne se distinguent de celles de leurs des pionniers des années 1950. En effet, les données spatiales ont des propriétés qui rendent l'utilisation de méthodes empruntées à des disciplines aspatiales discutables. La géographie quantitative peut se résumer comme étant l'analyse de données numériques spatiales, du développement de ces méthodes et de la construction d'un modèle mathématique spatial (Fotheringham, Brunsdon, & Charlton, 2000). Il s'agit d'une valorisation de l'information grâce à la modélisation du phénomène (Pointet, 2007). Ce qui sépare cette discipline de l'économie, de la sociologie quantitative ou encore de la physique, ou de l'ingénierie est l'accent prédominant mis sur la notion d'espace et de territoire. Les données spatiales combinent information avec localisation.

2.1.1. Application de la géographie quantitative

L'application de la géographie, qu'elle soit qualitative ou quantitative, est de générer du savoir. Selon Fotheringham et al. (2000), le premier objectif de la géographie quantitative est de passer d'un large jeu de données à une information synthétisée et lisible.

Deuxièmement, le rôle de la géographie quantitative est l'analyse exploratoire de données (*exploratory data analysis*). Celle-ci représente une première analyse des données qui permet de mettre en évidence des *outliers*, de suggérer des hypothèses ainsi que de construire des visualisations. Un *outlier*, ou donnée aberrante, est un point qui diffère fortement du reste des données. « *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism »* (Aggarwal, 2015, p. 1). Différents *outliers* sont illustrés à la figure 1.

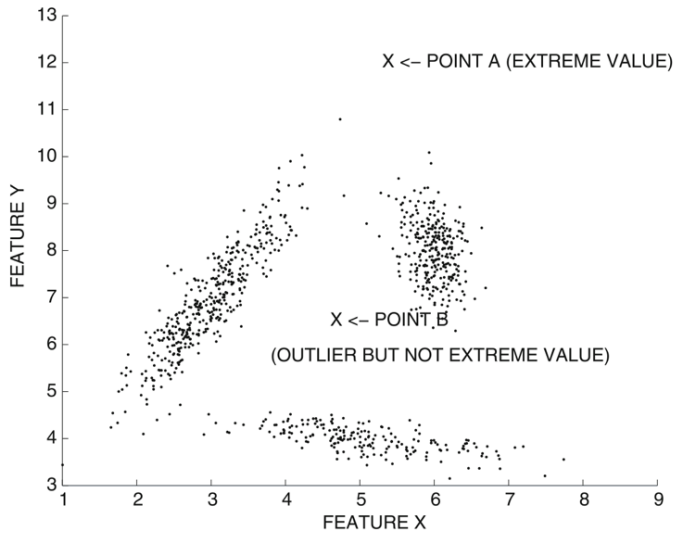


Figure 1 : Valeurs d'outliers multivariées (Aggarwal, 2015)

Troisièmement, la géographie quantitative permet de tester, via des analyses statistiques, si les résultats sont dus au hasard ou s'il y a effectivement un *pattern*, ou une observation récurrente. Un *pattern* peut définir si deux observations appartiennent au même groupe de données ou non (Duda, Hart, & Stork, 2001). Des analyses statistiques permettent d'avoir une meilleure idée à ce propos et des modèles mathématiques peuvent être utilisés sous différents scenarii afin de comparer les résultats issus des données à la réalité.

En résumé, l'analyse quantitative de données spatiales permet de mieux comprendre les processus spatiaux, non seulement en géographie mais également dans d'autres disciplines.

2.1.2. Dimension spatiale

Tout ce qui a trait de près ou de loin à un territoire présente une dimension spatiale (Aldenderfer & Maschner, 1996). La localisation d'éléments dans l'espace et leurs attributs font partie intégrante de la dimension spatiale, ce qui pose le contexte géographique (Brunet, 2006). Selon Pointet (2007), l'espace géographique peut être décomposé en deux champs : la mesure et l'analyse. Le premier se décrit comme la position dans l'espace alors que le second champ montre *« une détermination d'attributs géographiques issus de la mesure »* (Pointet, 2007), comme il est le cas dans ce projet. C'est à ce niveau que se distinguent les attributs des formes et des arrangements comme illustré à la figure 2.

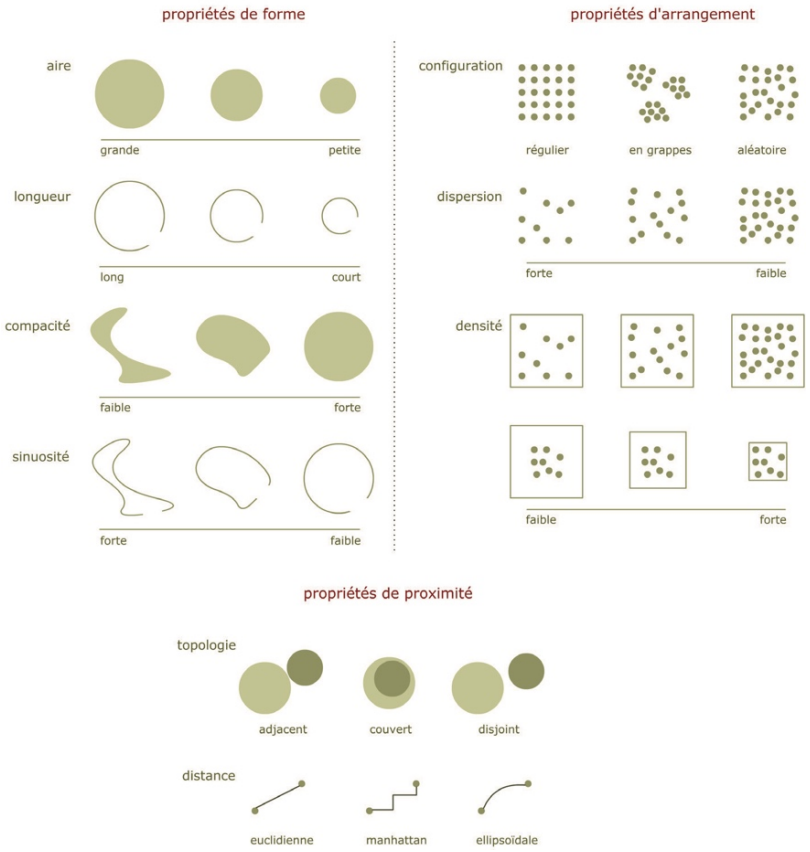


Figure 2 : Attributs des formes et des arrangements (Pointet, 2007)

2.1.3. Données spatiales

Les données spatiales sont caractérisées par des attributs de localisation, de surface, de distance et d'interaction (Anselin, 1989). Ces caractéristiques répondent à la Première Loi de Géographie de Tobler (1979) « *Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux*

objets éloignés. » : citation où les notions de « *proche* » et « *loin* » sont retranscrites par les données elles-mêmes.

Les informations géographiques sont différentes des informations aspatiales car elles ont une caractéristique qui leur est propre : la localisation. Ces données peuvent donc être projetées dans l'espace grâce à la localisation des objets et définies par les trois primitives que sont le point, la ligne et la surface (Figure 3) (Bertin, 1983).

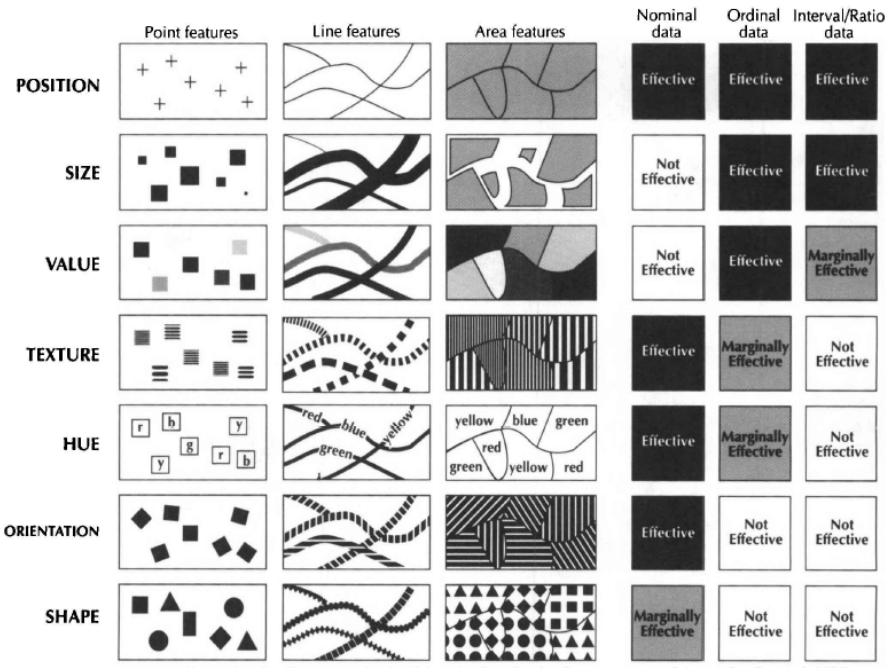


Figure 3 : Sept variables visuelles de Bertin (Bertin, 1983)

Quand les données sont stockées dans une base de données, l'attribut localisation est accompagné d'une valeur qui représente un indicateur et souvent d'un attribut temporel. Ces trois aspects sont étroitement liés aux questions élémentaires « Où ? », « Quoi ? » et « Quand ? » et définissent la nature de l'objet. Tous ces objets peuvent avoir différentes caractéristiques et même différentes temporalités au sein d'une temporalité comme illustré à la figure 4 (Kraak & Ormeling, 2010). La temporalité « jour » possède une sous temporalité « heure » par exemple.

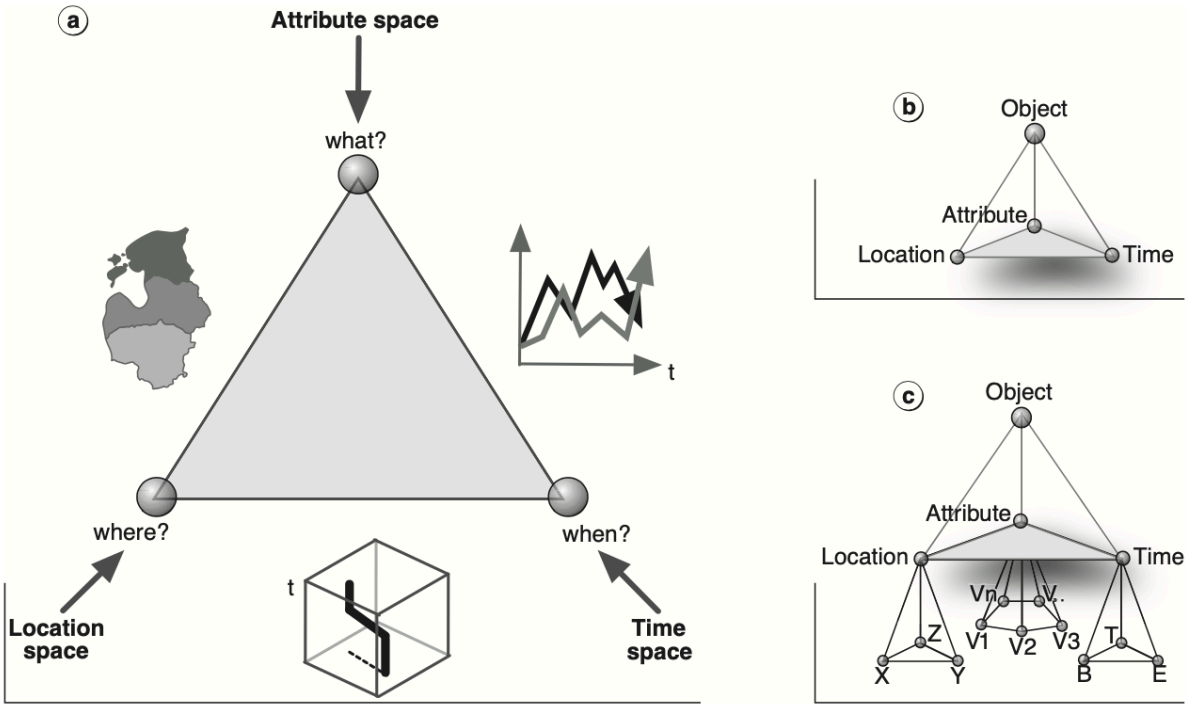


Figure 1.3 The characteristics of geospatial data: (a) its components location, attribute and time, and their related elementary questions where, what and when; (b) the object view; (c) detailed characteristics of the data components

Figure 4 : Caractéristiques des données géospatiales (Kraak & Ormeling, 2010)

2.2. Géographie des transports

« Le secteur des transports est fondamental dans la vie de nos sociétés où chacun se déplace continuellement, où la plupart des produits consommés viennent d'ailleurs et où circulent continuellement l'argent, les images et les informations » (Mérenne, 2013, p. 9). En effet, tous les types de transports sont omniprésents aujourd'hui. La société actuelle est entièrement basée sur les échanges et est complètement paralysée si ces flux sont interrompus. C'est pour cette raison que le champ de la géographie des transports est incontournable dans ce travail. L'activité humaine et économique restent au stade embryonnaire dans les régions excentrées et isolées et sont au contraire en explosion dans les régions urbaines, ce qui renforce encore l'importance de ce champ.

Bien que multidisciplinaires, les transports touchent particulièrement le domaine de la géographie car (Mérenne, 2013) :

- les infrastructures (arrêts de bus, gares ou encore aéroports) sont à la base d'un maillage spatial qui lie les différentes échelles, allant du quartier au monde entier ;

- les transports influencent la localisation des activités, accélèrent ou freinent leurs développements ;
- les transports sont à eux seuls un secteur d'activité avec leur propre logique spatiale et de localisation et ont un impact sur d'autres secteurs d'activité.

Dans le cadre des Transports Lausannois (tl), les infrastructures sont importantes, elles nécessitent des terminaux représentés par les arrêts de bus et de métros, mais aussi d'un balisage important au sol. Le réseau est parfois en site propre mais aussi en site partagé, c'est pour cela qu'il est possible que des retards puissent s'accumuler, ou que les charges de transports peuvent fortement différer. Une étude géographique doit donc prendre en compte ces paramètres afin de comprendre au mieux la situation du réseau et d'en tirer de bonnes observations.

2.2.1. Données de localisation et de circulation

Lors des dernières années, les données récoltées issues du trafic et du transport ont explosé, ce qui renforce l'idée que nous sommes entrés dans l'ère des *big data*. Cette situation d'omniprésence de données pousse à repenser la façon dont le monde scientifique analyse et construit des modèles sur les déplacements des gens et des biens (Lv, Duan, Kang, Li, & Wang, 2014). Ces flux d'informations ont le potentiel de permettre aux planificateurs et utilisateurs des transports de prendre de meilleures décisions quant aux déplacements. Cet enjeu majeur prend une importance toujours plus grande par rapport aux préoccupations environnementales afin de réduire les émissions de carbone par exemple. La prédiction du flux de trafic dépend fortement des données collectées de différentes sources, comme les caméras, les compteurs de voyageurs, les données de téléphones portables (GSM et GPS), etc. (Lv et al., 2014). Actuellement la gestion des transports devient de plus en plus guidée par les données (*data driven*) et leurs interprétations.

2.2.2. Optimisation des réseaux

Les offres de transport, surtout en ville, se sont développées séquentiellement au fur et à mesure de la croissance d'aires urbaines, mais cette offre ne correspond plus toujours aujourd'hui aux besoins des utilisateurs. Afin d'évaluer un réseau de transport, les distances entre les points x et y doivent être connues, ainsi que les temps d'attente sur ces nœuds ou le temps de chargement des passagers (Mandl, 1980). Le développement des transports cause une demande accrue de mobilité. Il est donc nécessaire d'organiser les systèmes de transport de façon bien distribuée.

L'ajustement des réseaux de transport doit se faire selon une organisation structurée pour maximiser la fonctionnalité des transports pour les usagers tout en réduisant les coûts. La conception et l'organisation du réseau est un des principaux problèmes de planification pour mettre en place des transports publics performants. Il est important d'utiliser des méthodes efficaces pour l'évaluation du réseau, quand il s'agit d'investir dans de larges infrastructures ou de modifier des infrastructures existantes (Bielli, Caramia, & Carotenuto, 2002). Ainsi, il est essentiel de trouver des outils efficients dans l'évaluation du trafic et sa visualisation, afin de décider au mieux d'une stratégie efficace de développement et d'optimisation du réseau.

2.2.3. Variabilité temporelle et analyse des retards

Dans les transports publics, les passagers doivent souvent changer de ligne afin de relier leur origine à leur destination. Les transferts modaux entre les bus, les métros ou les trains prennent du temps et peuvent provoquer du retard dans tout un réseau. Cela a pour conséquence que certains véhicules doivent en attendre d'autres provoquant une réaction en chaîne et des retards plus ou moins conséquents (Heilporn, De Giovanni, & Labbé, 2008). Si un transport attend une correspondance, les usagers déjà présents subiront un retard. Au contraire, si un véhicule n'attend pas la correspondance, des passagers peuvent endurer un retard encore plus conséquent dans l'attente du véhicule suivant provoquant une rupture de charge. Les protocoles et modèles de décisions temporels doivent aider à effectuer de meilleurs choix pour minimiser les retards, ce qui ne peut être réalisé qu'avec un moyen efficace qui parvient à mettre en évidence les retards en fonction de divers paramètres.

2.3. Exploration de données

Les données peuvent être traitées de différentes façons, les techniques utilisées pour extraire l'information évoluent constamment et se développent en même temps que les données sont accumulées. L'imprimerie et l'écriture ont mis mille ans à se développer pour en arriver à ce qu'elles sont actuellement. Par analogie, les données computationnelles, leur compréhension et leur utilisation ont pris également plusieurs dizaines d'années. Aujourd'hui, certaines données numériques collectées au XX^{ème} siècle sont devenues inaccessibles, illisibles à cause de leur format de conservation. Il existe cependant des données qui ont traversé les années, mais qui

proviennent majoritairement de grandes entités comme Microsoft ou Google. Les données sont aujourd'hui collectées 24 heures sur 24, 7 jours sur 7 (Hey, 2009).

La conservation de ces données fait l'objet de nombreuses études. En 2005, le National Science Board of the National Science Foundation publie « Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century », conférence qui est à la source de l'idée de l'importance de la conservation des données (National Science Board, 2005).

Dans la science des données, il existe trois étapes primordiales à identifier qui sont la récolte des données, leur mise en forme et finalement leur analyse (Hey, 2009) :

- **Récolte** : Des données sont produites continuellement chaque jour par des entités multiples, à tout type d'échelles et de formes diverses.
- **Mise en forme des données** : Cette partie peut être conséquente, il s'agit ici de trouver sous quelle structure enregistrer les données et les arranger correctement dans les bonnes bases de données. Il s'agit aussi de créer de bonnes métadonnées afin de répertorier au mieux les différentes variables. Sans métadonnées ou schémas explicites, les données peuvent rester inutilisables ou incompréhensibles. Les données qui ne sont pas mises en forme ont le risque d'être perdues ou de devenir illisibles. Il est donc important de sélectionner quelles données devraient être sauvées et quelles métadonnées doivent être présentes pour assurer une longévité.
- **Analyse** : L'analyse est la partie principale, qui définit le style de travail, soit l'usage à proprement parler de la base de données, de la modélisation et de la visualisation des données. Une base de données doit être construite afin de pouvoir répondre aux questions que pose l'analyste.

2.3.1. Paradigme de gestion de données

Aujourd'hui, les échelles d'observation ont changé, l'information n'est plus directement observée par l'œil humain, mais captée et analysée par un ordinateur. Par exemple, pour observer l'Univers, le chercheur n'est plus assis derrière son télescope, mais il analyse une série de fichiers de données produite automatiquement par celui-ci. Les techniques scientifiques sont maintenant différentes, ce qui amène à de nouveaux paradigmes, comme celui de l'exploration de données (Kitchin, 2014).

Hey et al. (2009) identifient quatre paradigmes de la recherche scientifique qui sont illustrés à la figure 5. Ce tableau montre que le quatrième paradigme, qui correspond à notre époque actuelle, se développe dans le cadre de l'émergence d'une collecte intensive de données analysées à travers des programmes informatiques, des bases de données ainsi que des statistiques.

Table 1. Four paradigms of science.

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Compiled from Hey et al. (2009).

Figure 5 : Quatre paradigmes de la science (Hey, 2009)

Changement de paradigme

Selon Kuhn et Hacking (2012, p. 164), "*a paradigm constitues an accepted way of interrogating the world and synthesizing knowledge common to a substantial proportion of researchers in a discipline at any one moment in time*". L'apparition des *big data* est un tournant radical dans la façon d'effectuer des recherches, ce qui ouvre de nouvelles possibilités d'analyse et de nouvelles définitions de la vie sociale (Boyd & Crawford, 2012).

Les *big data* ne sont pas simplement des critères de volumes, mais répondent aux caractéristiques suivantes (Kitchin, 2013) :

- *volume* : grand en volume de données, pouvant aller jusqu'à des petabytes de données ;
- *velocity* : vitesse de création de données en temps réel ou presque ;
- *variety* : données structurées et non structurées ;
- *exhaustive* : but de collecter des données pour une population ou un système entier (n = tout) ;
- *resolution and uniquely indexical* : données fines et indexées ;
- *relational* : contient des champs communs permettant une jointure de différents jeux de données ;
- *flexible and scaleability* : possibilité d'ajouter de nouveaux champs facilement et possibilité de grandir rapidement en taille.

Ces diverses caractéristiques ont un coût qui doit être contrôlé et limitent donc leurs possibilités, leurs temporalités et leurs tailles (Miller, 2010).

La difficulté avec les analyses de *big data* est de jouer avec l'abondance de ces données qui n'ont pas toujours été créées pour répondre à des questions spécifiques. L'analyse de *big data* met en place une approche entièrement nouvelle, plutôt que de tester une théorie en analysant les données, les analyses peuvent être « nées depuis les données ».

Empirisme

"*The data deluge makes the scientific method obsolete*" (Anderson, 2008, p. I). D'après Anderson, il est possible d'analyser des données sans poser d'hypothèses. Il suffit de lancer les données dans un modèle et d'analyser les *patterns* qui ressortent des données.

Avec une méthode de travail empirique, les observations suivantes peuvent être déclarées :

- les *big data* peuvent enregistrer tout un domaine à grande résolution ;
- pas besoin d'hypothèses ;
- les données peuvent parler d'elles-mêmes sans parti pris ;
- interprétation possible par quiquonque comprend une statistique ou une visualisation de données.

Data-driven

Cette méthode est dérivée des méthodes scientifiques traditionnelles mais diffère du fait qu'elle cherche à générer des hypothèses à la source des données. Elle est donc plus souple que les méthodes traditionnelles, tout en gardant de l'induction dans la recherche (Kelling et al., 2009).

2.3.2. Exploratory Data Analysis (EDA)

Traduit de John Tukey (1992, p. 408), l'*Exploratory Data Analysis* (EDA), est une « *procédure d'analyse de données, une technique d'interprétation de résultats de ces procédures, une méthode de planification et de collecte de données pour en faciliter l'analyse, pour les rendre plus précises, et est l'ensemble des mécanismes et des résultats des statistiques (mathématiques) qui s'appliquent à l'analyse des données.* ». L'EDA permet d'explorer un jeu de données et d'appliquer des techniques efficaces pour rendre la description plus facile et compréhensible avec un jeu de données (Tukey, 1977). L'EDA n'analyse pas de fond en comble un jeu de données mais sert de base à son analyse. Le but est de résumer les caractéristiques les plus fréquentes apparaissant dans un jeu de données et aide ainsi à limiter un objectif.

Il serait illusoire de s'attendre à mettre en lumière des éléments inattendus, mais cela doit donner un résumé et une idée première quant aux données. Plus le jeu de données est important, plus il est difficile de passer à côté d'une EDA. Il faut par exemple sortir des indicateurs simples comme les extrêmes et les valeurs médianes qui ne sont évidemment que rarement suffisants pour une analyse pertinente. Cependant, cela permet de comprendre où aller chercher les premières informations (Tukey, 1977).

2.3.3. Exploratory Spatial Data Analysis (ESDA)

Le développement récent de l'informatique a permis d'être en interaction pratiquement instantanée avec de larges bases de données et d'effectuer de nombreuses opérations avec des systèmes d'information géographiques (SIG). La géographie spatiale des années 1960 n'avait pas la technologie suffisante pour traiter efficacement l'information, ni pour la visualiser, ce qui est maintenant rendu possible avec des SIG modernes (Anselin, 1996). C'est grâce à l'informatisation que l'EDA a pu évoluer en ESDA, *Exploratory Spatial Data Analysis* qui ajoute la dimension spatiale à l'exploration de données.

Par nature, les modèles statistiques sont statiques et limitent donc l'interaction possible avec les données, les modèles et l'analyse de ces données. Les approches dynamiques ou interactives permettent d'obtenir un environnement graphique qui permet une manipulation en direct des données, ce qui confère aux EDA et EDSA une importance considérable dans les SIG. L'affichage des données est primordial et il faut choisir des indicateurs simples et parlants, tout en prenant garde aux *outliers* et aux observations atypiques (Anselin, 1996).

Dans la méthode EDA, Il faut être attentif à l'autocorrélation, car la méthode est basée sur l'indépendance des données. Il faut donc développer une autre méthode et c'est ici que l'ESDA est utile, car elle prend en compte l'espace.

Une des motivations d'une méthode spatiale est d'impliquer le facteur humain plus directement pour l'exploration des données. Il s'agit de l'utilisateur et non le statisticien qui détermine complètement quelles données visualiser. Cela devient particulièrement efficace quand les jeux de données sont très larges ou comportent de multiples dimensions. Les données peuvent alors devenir dynamiques et se visualiser simultanément sur une carte ou sur un graphique par exemple.

Les ESDA peuvent être définies comme une collection de techniques pour décrire et visualiser des distributions spatiales, identifier des localisations atypiques, des *outliers*, découvrir des patterns d'associations spatiales (*clusters*) et suggérer différents régimes spatiaux ou différentes formes d'instabilité spatiale (Anselin & Bao, 1997). Les données sont sensibles à l'autocorrélation car elles sont influencées par les valeurs voisines et cette autocorrélation peut être expliquée par les ESDA. Un autre facteur important qu'il est nécessaire de prendre en compte est que les données sont souvent collectées pour des buts différents que leur utilisation. Il faut donc faire au mieux avec les données disponibles (Bivand, 2010) et il est mieux d'obtenir une réponse approximative à une bonne question que de répondre exactement à la mauvaise question (Chambers, 2008).

2.3.4. Anomaly detection

Une anomalie est une donnée qui ne se conforme pas à une distribution qui suit un certain comportement (Chandola, Banerjee, & Kumar, 2009).

La détection d'anomalie (*anomaly detection*) dans l'exploration de données permet de mettre en avant et d'identifier des éléments statistiques inhabituels dans un jeu de données. Comme le montre la figure 6 (Chandola et al., 2009), deux *outliers* (o1 et o2) et un groupe *d'outliers* (O3) sont détectés. Cela peut être valable pour des retards de bus par rapport à un jour, mais aussi pour un trafic inhabituel sur un site internet qui pourrait mettre en avant un piratage de la plateforme.

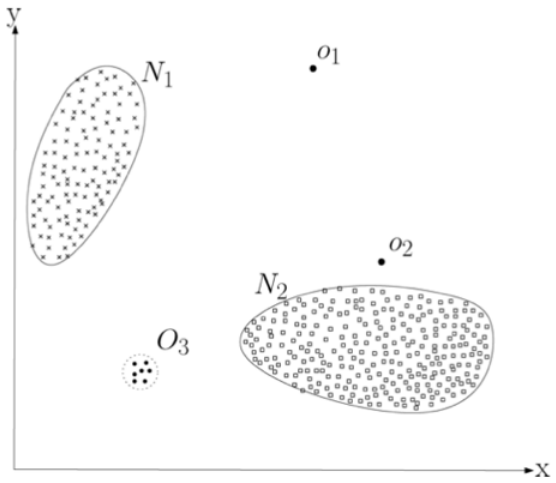


Figure 6 : Détection des anomalies dans un jeu de données à deux dimensions (Chandola et al., 2009)

Cette pratique n'est pourtant pas récente, Edgeworth (1887, p. 364) la définissait déjà au XIX^{ème} siècle : "*Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined.*" C'est-à-dire des observations qui diffèrent de la loi de fréquence des autres observations.

Ces anomalies peuvent arriver pour diverses raisons. Dans le cadre des transports, elles peuvent arriver pour cause de neige, de manifestation, de panne ou encore d'accident. Il est aussi possible que des données n'aient pas été répertoriées correctement ou qu'il y ait eu un biais dans la transformation de ces dernières. La figure 7 montre les composantes clés pour la détection d'anomalies.

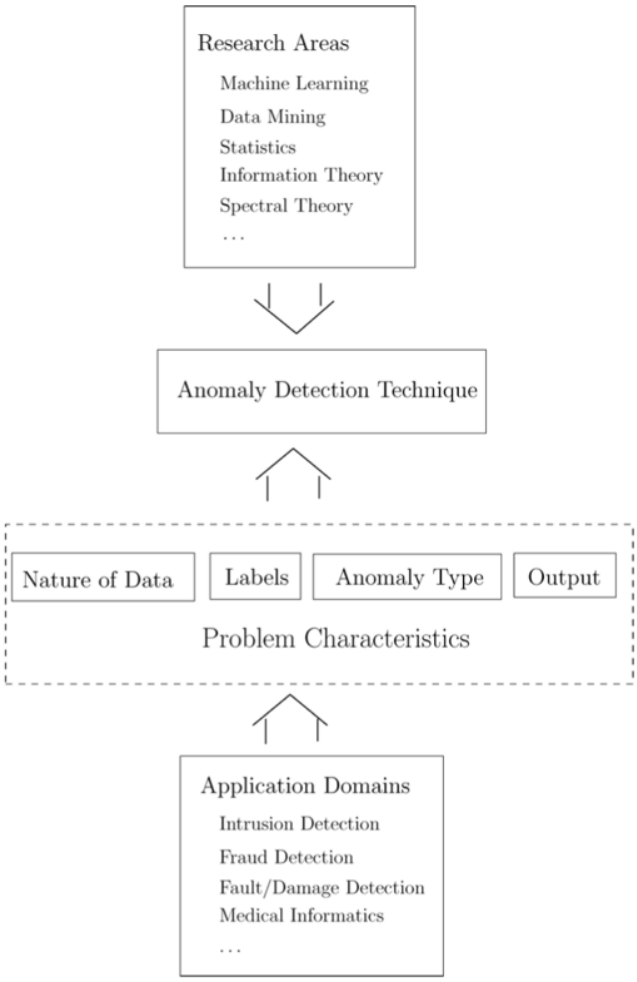


Figure 7 : Composantes clés pour la détection d'anomalies (Chandola et al., 2009)

Le type de donnée a aussi une influence sur l'anomalie, il peut s'agir de données binaires, catégorielles ou continues.

2.3.4.1. Type d'anomalie

Les anomalies peuvent être classées suivant trois différentes catégories (Chandola et al., 2009) :

- les anomalies de point : une donnée d'un jeu de données n'est pas en accord avec la tendance des points alentours ;
- les anomalies de contexte : une donnée peut être conforme à un jeu de données, mais pas dans son contexte ;
- les anomalies collectives : une collection de données peut ne pas être conforme au jeu de données complet.

Il s'agit aussi dans ce travail de prêter attention aux données qui ne sembleraient pas être conformes au jeu de données ou qui fausseraient l'interprétation d'un résultat.

2.4. Visualisation de données

Les données brutes peuvent être exploitées par un public averti et manipulées par des statisticiens, des mathématiciens ou encore des *data scientists*. Pour transmettre de l'information à un plus large public, ces données doivent être traduites et interprétables et c'est là que la visualisation de données est importante.

La base de la visualisation est de présenter des données sous forme visuelle, permettant à l'être humain de se faire une idée de ce que représentent les données, de tirer des conclusions et de directement interagir avec les données (Keim, 2002).

L'exploration visuelle de données est un excellent outil d'exploitation de larges bases de données, en particulier quand les connaissances sur le jeu de données sont restreintes et que le but de l'exploration est vague. Elle permet aussi de mieux comprendre ce que veulent dire les données et aide à formuler de nouvelles hypothèses (Keim, 2002).

Au sens le plus étendu du terme, la visualisation est l'activité de former un modèle mental de quelque chose (Spence, 2014). La visualisation n'est pas un phénomène récent : déjà dans les années 1850, Florence Nightingale schématisa les améliorations des performances des hôpitaux à Scutari durant la guerre de Crimée. Chaque segment de cette visualisation (Figure 8) est proportionnel au nombre de décès dans les hôpitaux.

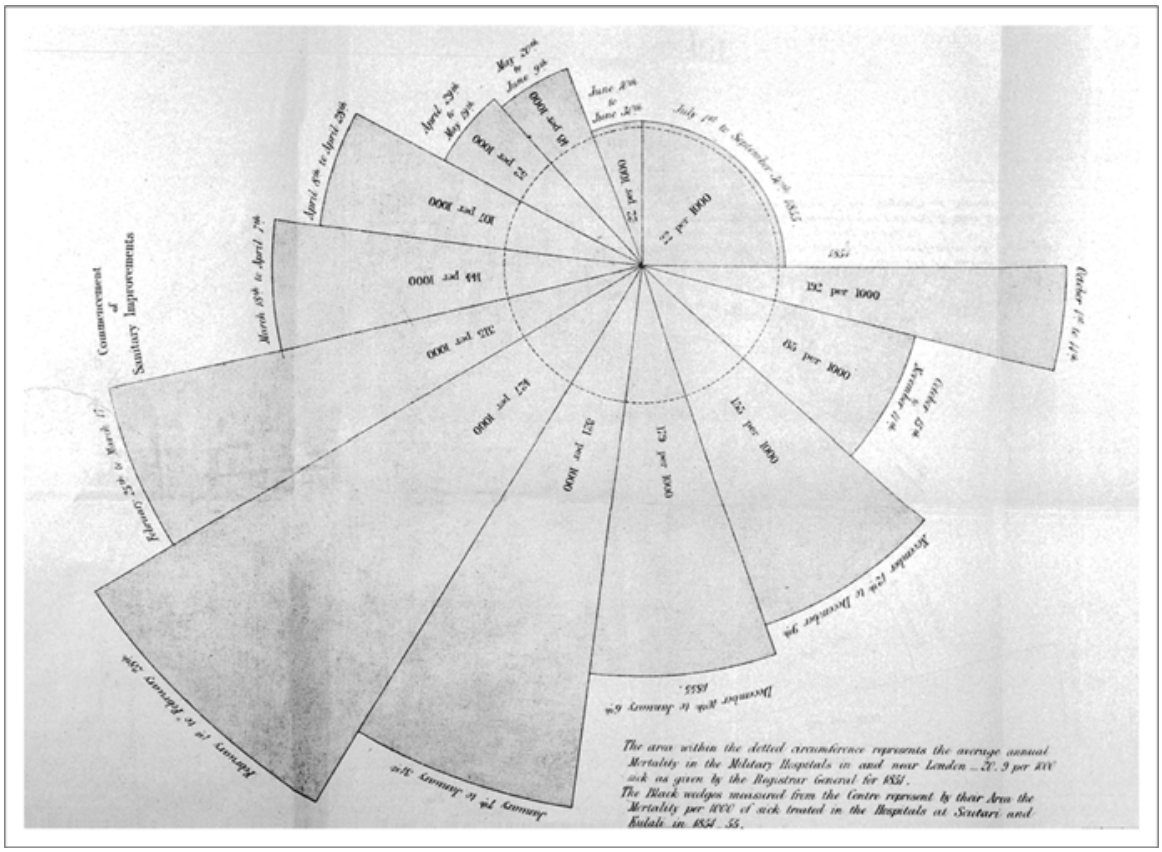


Figure 8 : "Rose diagram" de Florence Nightingale schématise la réduction du nombre de morts dans les hopitaux militaires à Scutari grâce aux changements qu'elle a effectué (Spence, 2014)

Dans le thème des transports, déjà en 1931, Harry Beck créa une nouvelle carte du métro londonien afin de donner plus de clarté à la lecture du plan (Figure 10). Ce type de carte ne respecte plus les distances et la topographie du lieu mais a pour avantage d'être bien plus lisible que les anciennes cartes bien plus réalistes (Figure 9) (Spence, 2014). Ce genre de plan est toujours utilisé dans la plupart des réseaux de transports comme dans celui des tl (Figure 11).

Figure 9 : Carte du métro de Londres avant 1931 (Spence, 2014)

Figure 10 : Nouvelle carte du métro de Londres de Harry Beck en 1931 (Spence, 2014)

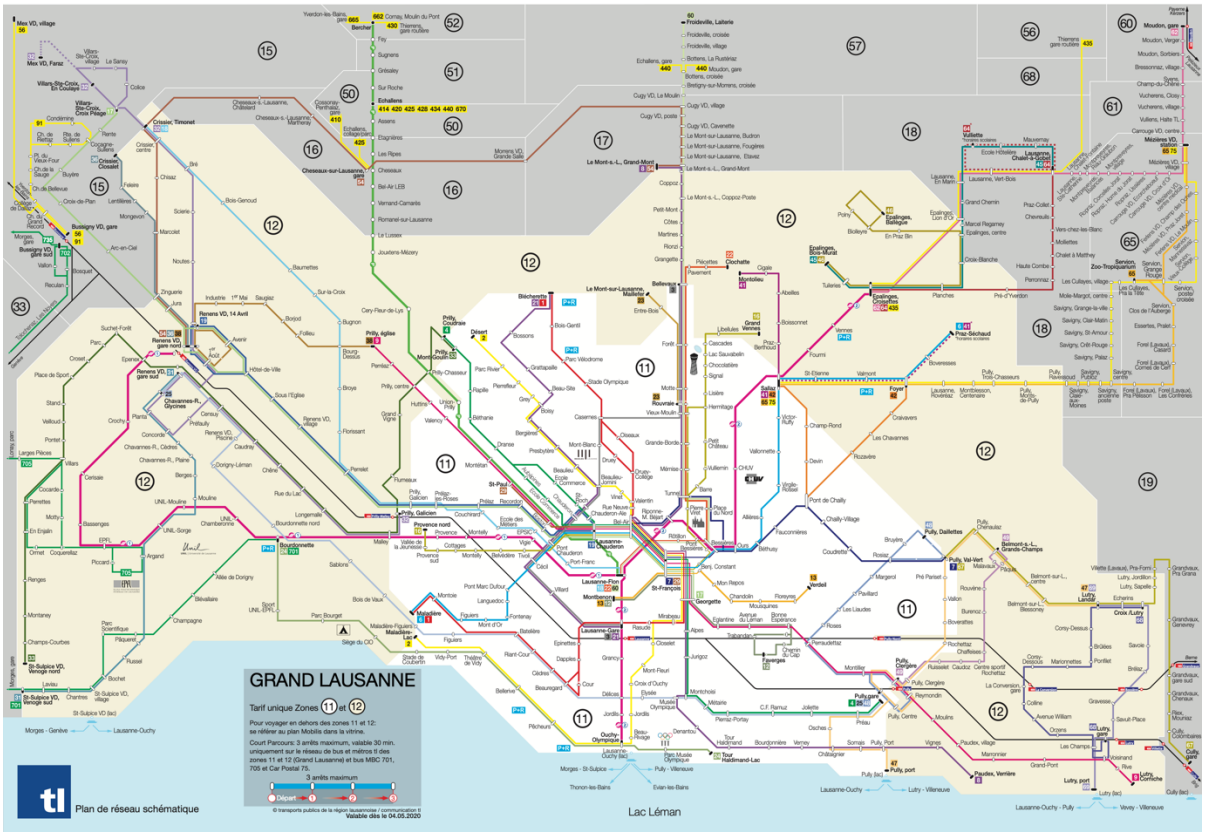


Figure 11 : Plan du réseau des Transports Lausannois (TL, 2020)

La mise en forme de données est nécessaire pour créer des visualisations et rendre ces dernières interprétables. En effet, un jeu de données est difficilement lisible sans une mise en forme conséquente. La figure 12 issue de Anscombe (1973) rassemble quatre jeux de données aux propriétés identiques. Les statistiques descriptives de la moyenne, de l'écart-type ainsi que de la droite de régression sont les mêmes. Par contre, une fois que les points sont distribués sur un graphique (Figure 13), les distributions s'avèrent très différentes. Cette illustration montre à quel point les visualisations sont importantes pour comprendre parfaitement un jeu de données, ses *outliers* et ses particularités.

SET A		SET B		SET C		SET D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.86	5	4.74	5	5.73	8	6.89

SUMMARY STATISTICS

μ_X = **9.0** σ_X = **3.317**

μ_Y = **7.5** σ_Y = **2.03**

LINEAR REGRESSION

Y = **3 + 0.5X**

R^2 = **0.67**

Figure 12 : Quatre jeux de données ayant les mêmes caractéristiques générales (Anscombe, 1973)

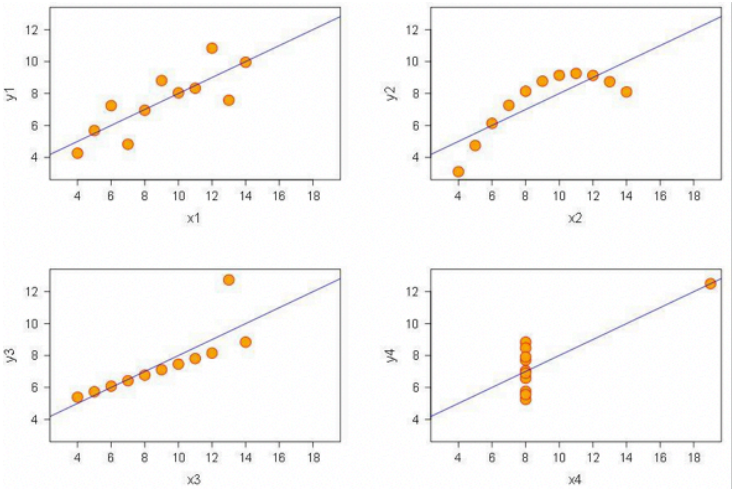


Figure 13 : Représentation graphique des quatre jeux de données d'Anscombe (1973)

La géovisualisation n'est donc pas une thématique séparée des ESDA mais en constitue plutôt sa colonne vertébrale (Bivand, 2010). Il s'agit en fait du prolongement des analyses statistiques projetées dans l'espace.

2.4.1. Dashboard

Il est possible de créer des visualisations de différents types. Le *dashboard* ou tableau de bord permet de regrouper plusieurs informations en une seule fenêtre. Un *dashboard* peut se définir comme suit : "*A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance*" (Few,

2006, p. 26). Un *dashboard* en informatique a finalement les mêmes caractéristiques que le tableau de bord d'une voiture : résumer les informations utiles en un seul coup d'œil.

Les tableaux de bord comportent plusieurs caractéristiques (Few, 2006) :

- ils doivent afficher l'information dans un but spécifique ;
- ils doivent tenir sur un écran d'ordinateur ;
- les objectifs de visualisation doivent définir le taux d'interactivité du *dashboard* ;
- ils sont utilisés pour exposer l'information en un seul regard.

La figure 14 est un bon exemple de *dashboard*. Il regroupe beaucoup d'informations par rapport au COVID-19, tant sur le plan des chiffres que de la localisation (ArcGIS, 2020).

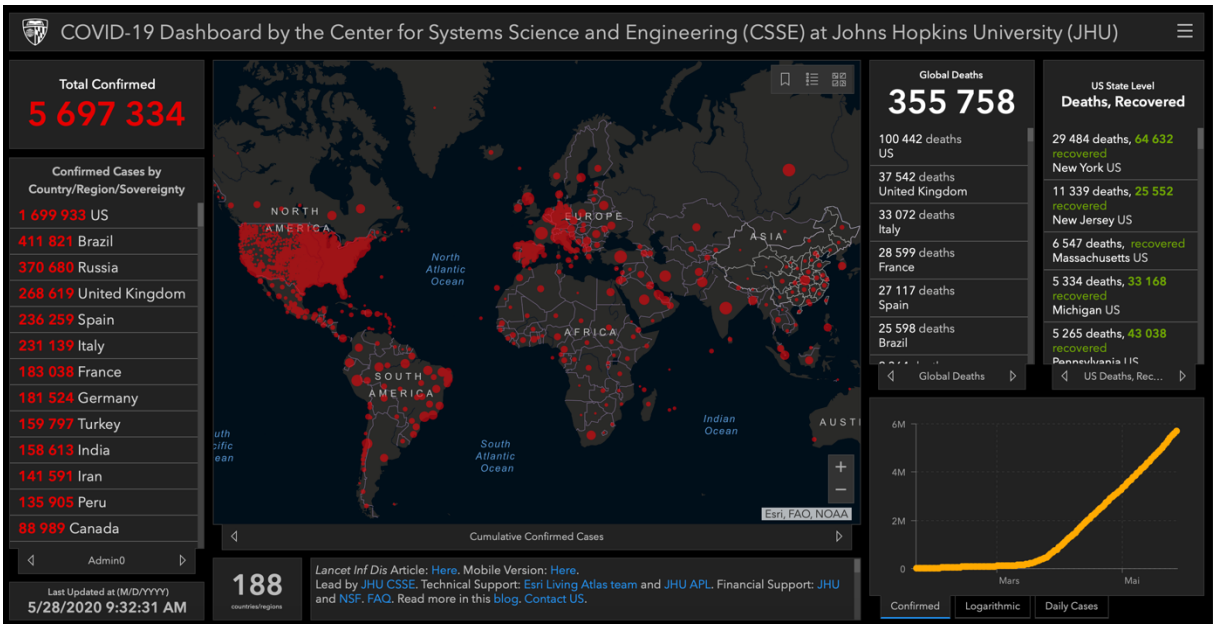


Figure 14 : Dashbord de la propagation du COVID-19 créé par ArcGIS, consulté le 28 mai 2020

2.5. Time series

"A time series is a sequence of observations taken sequentially in time", en français, une série temporelle est une séquence d'observations prises séquentiellement dans le temps (Box, Jenkins, Reinsel, & Ljung, 2016, p. 1). Beaucoup de jeux de données peuvent être décris comme une série temporelle. Il s'agit en fait de toutes les observations qui peuvent être décrites dans un laps de temps et qui sont organisées de façon chronologique, comme par exemple le nombre d'accidents de la route sur une année dans un pays, la quantité de précipitations lors d'une journée, mais aussi le nombre de passagers transportés par un bus dans une période donnée. Les exemples de séries

temporelles touchent de nombreux thèmes, comme l'économie, l'ingénierie, les sciences sociale ou encore la géographie. Une de leurs caractéristiques primordiales est la dépendance des observations par rapport au temps. L'analyse des séries temporelles est l'étude de ces dépendances.

Box et al. (2016) identifient cinq aires d'application importantes :

1. La prédiction de futures valeurs à partir de valeurs actuelles et antérieures.
2. La détermination d'une fonction de transfert qui montre un résultat par rapport aux séries temporelles.
3. L'utilisation d'indicateurs dans les fonctions de transfert pour représenter et évaluer les effets d'événements inhabituels sur le comportement d'une série temporelle.
4. L'examen des corrélations entre plusieurs variables de séries temporelles et la détermination de modèles dynamiques appropriés pour représenter ces relations au fil du temps.
5. La conception de schémas de commandes simples au moyen desquels les écarts potentiels du résultat du système par rapport à une cible souhaitée peuvent, dans la mesure du possible, être compensés par l'ajustement des valeurs d'entrée.

C'est grâce à ces procédés qu'il est possible de construire une série temporelle de N observations successives, ce qui forme un échantillon issu d'une population. Dans le cadre de ce travail, les observations sont distribuées sur l'année 2018, cela constitue l'échantillon de données.

3. Méthodologie

Ce chapitre décrit comment les données ont été utilisées, à quelles fins elles ont été traitées ainsi que par quels logiciels et langages de programmation. Il s'agit aussi dans cette partie de délimiter le territoire d'étude, en expliquer ses spécificités et les raisons pour lesquelles les lignes de bus numéros 2 et 18 ont été sélectionnées. Le processus de préparation de données est aussi brièvement expliqué dans ce chapitre. L'intégralité du code nécessaire au fonctionnement de cette application se trouve sur dans répertoire détaillé sur la plate-forme en ligne « GitHub » disponible à l'adresse suivante : https://github.com/romainloup/tlDataViewer. Les données des tl ne sont pas mises à disposition car elles ne sont pas en libre accès. La structure des tables est par contre ajoutée afin de pouvoir tester l'application avec d'autres données.

3.1. Délimitation du territoire

Les données et le cas d'étude proviennent des Transports Lausannois et les lignes de bus étudiées ont été choisies en accord avec les tl pour étudier certaines spécificités.

3.1.1. Réseau tl

Le réseau multimodal des tl est long de plus de 260 km et assure quotidiennement la mobilité de plus de 326'000 voyageurs dans l'agglomération lausannoise, sur 39 communes, par 40 lignes d'autobus et trolleybus en plus des deux lignes de métro (Transports Lausannois, 2020).

Les lignes choisies pour ce travail sont les lignes de bus numéros 2 et 18. Il s'agissait en effet de choisir des parcours qui étaient représentatifs du réseau des tl dont les irrégularités questionnent les dirigeants des Transports Lausannois. Ces lignes ne sont que partiellement en site propre et sont sujettes au trafic routier. Ces deux lignes ont des caractéristiques assez différentes, car la ligne 2 part du bord du lac et finit au nord de la ville, alors que la ligne 18 part de la ville et se termine à Crissier, dans la banlieue lausannoise.

Lors de la première réflexion à propos de ce travail, il a été établi que l'étude des lignes de métro n'apportait pas assez d'informations, car celles-ci sont complètement en site propre et les retards engendrés sont principalement occasionnels plutôt que systématiques, dus à des pannes ou des

problèmes exceptionnels. Le but de ce mémoire étant d'étudier des retards systématiques, le choix s'est alors porté uniquement sur des lignes de bus plutôt que de métro. Le plan géographique du réseau est disponible à l'annexe 1 (Transports Lausannois, 2020).

3.2. Lignes 2 et 18

Cette section décrit les caractéristiques des lignes de bus numéros 2 et 18 et leur tracé. Les tracés détaillés des deux lignes sont disponibles aux figures 15 et 16 (Etat de Vaud, 2020).

Les annexes 2 et 3 décrivent les positions ainsi que les noms et codes des arrêts pour les deux directions des lignes 2 et 18. Pour les codes des arrêts, les lettres N, E, S et O représentent respectivement les orientations nord, est, sud et ouest par rapport à la route sur laquelle est placé l'arrêt.

3.2.1. Ligne 2

La ligne numéro 2 a un tracé urbain et n'est que partiellement en site propre. Elle part de la Maladière, en direction d'Ouchy et monte jusqu'à St.-François, puis continue au nord-est de la ville pour finir à Lausanne, Désert. La partie sud de la ligne est fortement impactée par le trafic des pendulaires qui se rendent à l'autoroute via la Maladière. La section suivante, jusqu'à Georgette, traverse la partie sud de la ville et n'est pas en site propre. Cette partie est aussi sensible au trafic. La fin de la ligne jusqu'à Désert comporte plus de sections en site propre, ce qui rend les tronçons potentiellement moins sensibles au trafic bien que les giratoires doivent être parfois négociés avec le reste du trafic.

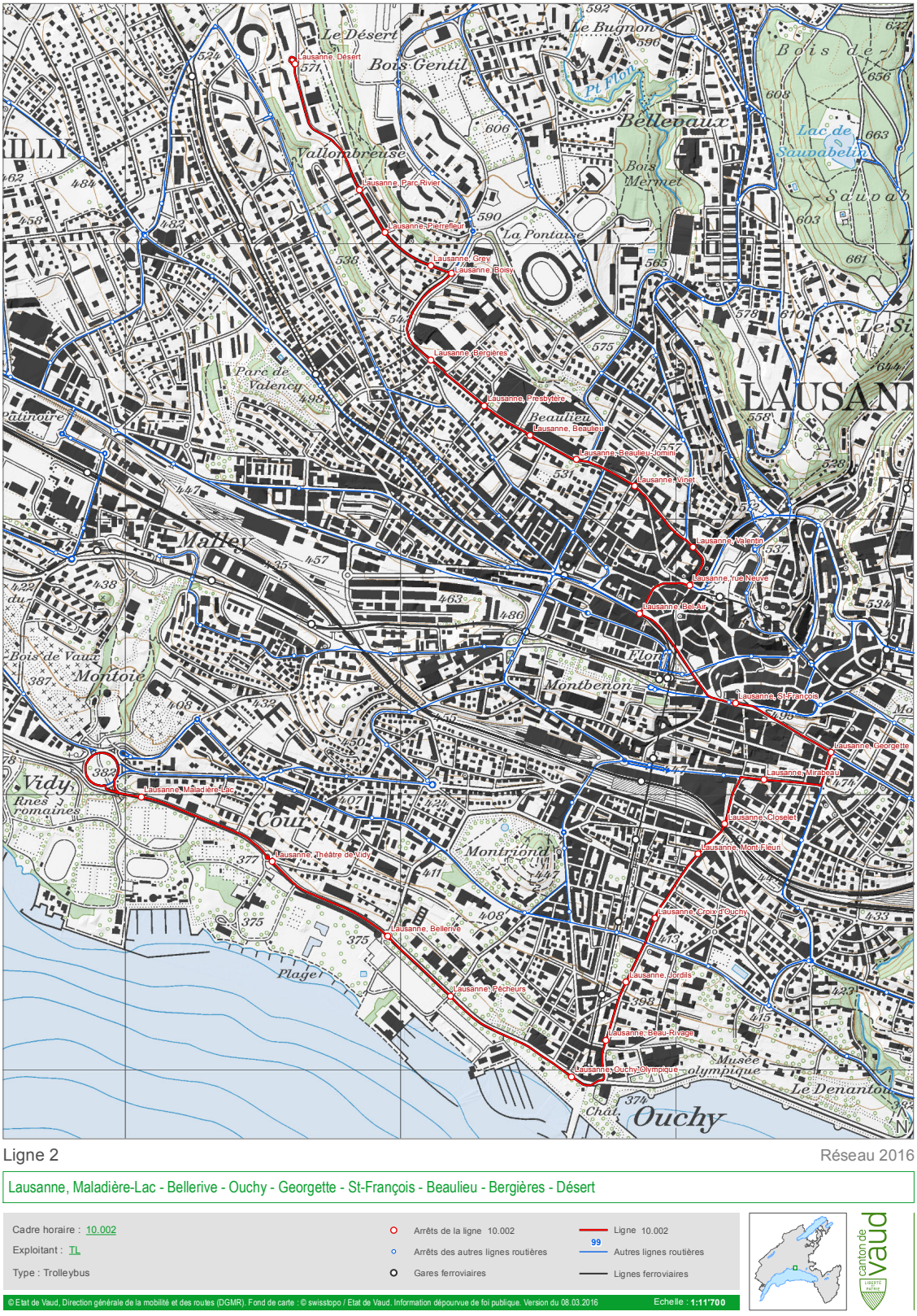


Figure 15 : Tracé de la ligne 2

3.2.2. Ligne 18

La ligne numéro 18 peut être caractérisée différemment. Elle commence à Lausanne-Flon, traverse Renens et se termine à Crisser, dans la banlieue ouest de Lausanne. Cette ligne passe aussi par un axe important de la région reliant le centre de la ville à l'ouest, traversant ainsi l'agglomération lausannoise. Les perturbations ne viennent pas des mêmes sources que pour la ligne 2, ce qui est intéressant à comparer.

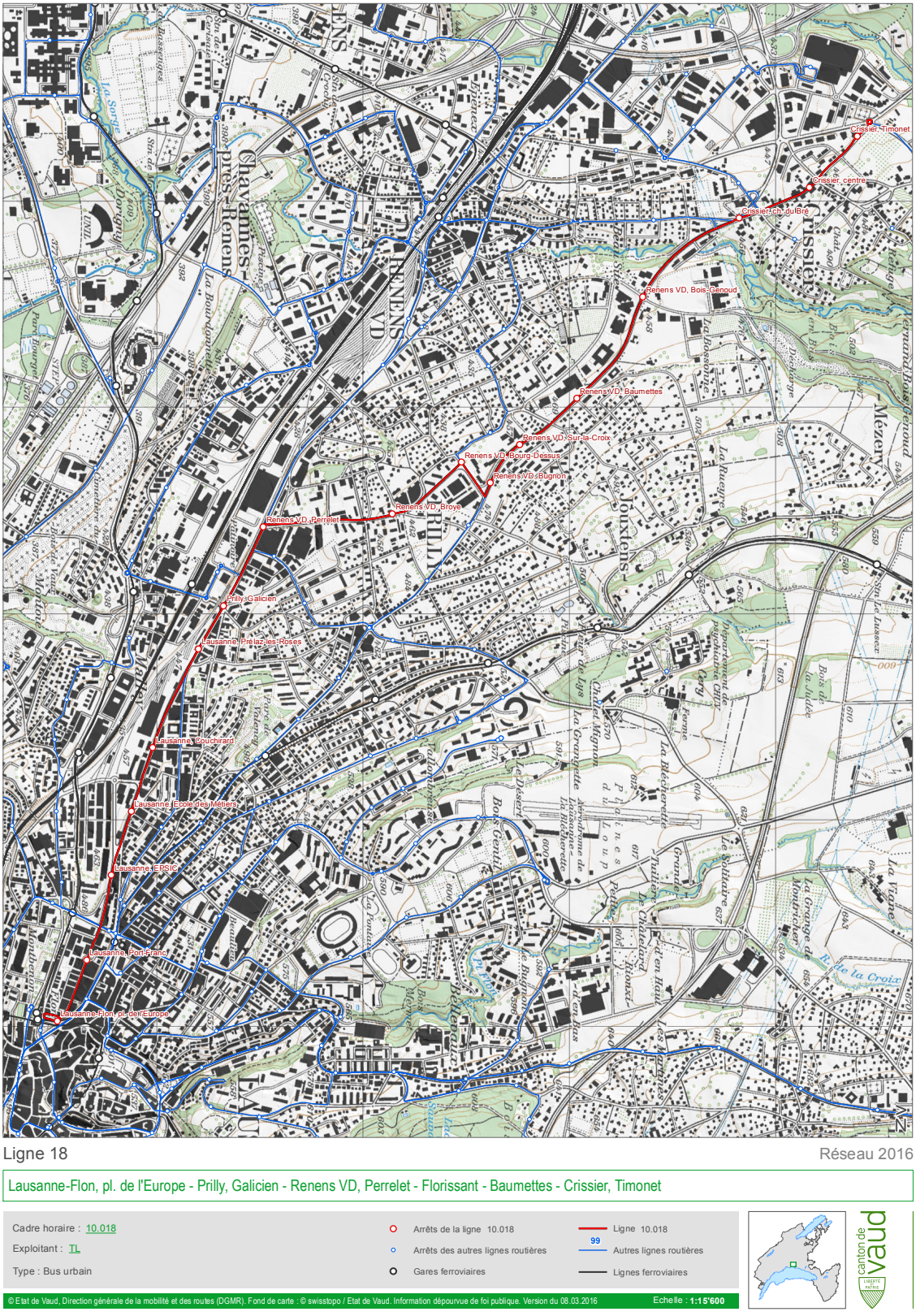


Figure 16 : Tracé de la ligne 18

3.3. Jeux de données

Plusieurs jeux de données ont été utilisés dans ce travail, mais la majorité des données proviennent des tl. Les données de météo proviennent de MétéoSuisse.

Les données des tl ne sont pas librement accessibles, un échantillon est donc disponible en annexe 4 afin de comprendre comment elles sont organisées.

Certaines autres données, comme le calendrier qui regroupe les plages de vacances, les numéros de semaine ou encore le nom des jours ainsi que les tracés des lignes ont été créées pour ce travail. Toutes les données couvrent l'années 2018.

3.3.1. Jeux de données tl

Trois jeux de données originaux ont été nécessaires afin de construire la base de données utile à ce travail, soit les fichiers trafic des passages (TPs), trafic journalier moyen (TJM) et horaires théoriques (h_theo).

TPs

Ce fichier contient les heures de passage (entrée et sortie) aux arrêts des voyages réalisés. Le retard est calculé d'après des jointures entre ce fichier et celui de l'horaire théorique. Ce type de fichier contient les colonnes suivantes (la terminologie des listes qui suivent est celle des tl) :

- Arret Cod – identifiant de quai (arrêt par direction) ;
- Arret Libelle – nom de l'arrêt ;
- Depart – heure théorique de départ du voyage depuis le terminus ;
- Distance Ligne – distance entre l'arrêt courant et l'arrêt précédent. Pour le terminus de départ, cette valeur est de 0 ;
- Jour de Date Exploit – date de voyage ;
- Ligne – numéro de ligne
- Ordre Ligne Sens – numéro d'arrêt dans voyage entier. Il ne commence pas à 0 pour les voyages partiels du matin
- Tronçon – tronçon avec la fin sur l'arrêt courant

- Voy Sens – direction de voyage : « A » pour aller et « R » pour retour
    - Ligne 2 : « A » est le sens Maladière à Désert, « R » est le sens Désert à Maladière
    - Ligne 18 : « A » est le sens Lausanne, Flon à Timonet, « R » est le sens Timonet à Lausanne, Flon
- voy_ID_unique – numéro unique de voyage dans la base de données de temps de parcours ;
- Max. Moment d'entrée à l'arrêt – heure réalisée d'entrée à l'arrêt ;
- Max. Moment de sortie à l'arrêt – heure réalisée de sortie de l'arrêt.

Les données de neuf jours sont manquantes, car ces données n'ont soit pas passé le test qualité des tl, soit un problème technique est survenu durant la prise de données. Il s'agit des 6, 7, 8, 9, 10, 11, 12 juin et 26, 27 juillet.

TJM

Ce fichier contient le « trafic moyen journalier ». Il s'agit des données relatives aux voyageurs montés et descendus par arrêt ainsi que les charges à bord pour les voyages réalisés. La charge peut être définie comme le nombre de passagers entrés dans un bus. Ce type de fichier contient les colonnes suivantes :

- Jour de Date Exploitation – date de voyage ;
- Ligne – numéro de la ligne ;
- Tranche hh:mm depart – heure théorique de départ de voyage du terminus ;
- Cont Voyage Id - numéro unique de voyage dans la base de données des voyageurs ;
- Sens – direction de voyage ;
- Max. de Ordre - numéro d'arrêt dans voyage entier, donc, il ne commence pas depuis 0 pour les voyages partiels du matin ;
- Arret Code – identifiant de quai (arrêt par direction) ;
- Arret Libelle – nom de l'arrêt ;
- Arrêt Heure Passage hh:mm:ss – heure réalisée de sortie de l'arrêt d'un bus, à utiliser pour une jointure entre les tableaux de TPs et TJM ;
- Descendu par jour – nombre de passagers descendus à l'arrêt « Arret Code » au moment entre l'heure réalisée d'entrée à l'arrêt et l'heure réalisée de sortie de l'arrêt ;
- Monte par jour – nombre de passagers montés à l'arrêt à l'arrêt « Arret Code » ;

- Charge par jour – nombre de voyageurs dans le bus pour le tronçon suivant l'arrêt « Arret Code ».

Remarques sur TPs et TJM

Ces types de fichiers sont agrégés par mois et par ligne de transport. Les noms d'arrêts ne sont pas constants, ils peuvent être changés en cours d'année pour des raison internes ou externes aux tl. L'utilisation du champ « Arret Libelle » dans une jointure des tableaux de temps de parcours et de trafic de voyageurs pourrait provoquer une erreur de jointure.

Les nombres de voyageurs ne sont pas toujours entiers. Si les données mesurées ne correspondent pas aux critères internes de qualité des tl, l'entreprise remplace les données des montées par des données estimées en fonction des voyages équivalents.

Les voyages partiels se produisent suivant trois cas :

- avant l'arrivée sur la ligne, le bus a effectué un voyage commercial depuis la sortie du dépôt et s'est joint à la ligne au milieu du trajet ;
- le bus rentre au dépôt depuis un arrêt au milieu du trajet ;
- un voyage n'a pas été complètement effectué pour des raisons d'exploitation.

h_theo

Le jeu de données h_theo contient les horaires théoriques de départ de chaque bus selon différents horaires : la semaine, le samedi, le dimanche et les vacances. L'horaire des vacances n'est valable que du lundi au vendredi, car durant les vacances, les horaires des weekends restent les mêmes.

Ce type de fichier contient les colonnes suivantes :

- Ligne – numéro de la ligne ;
- Voiture – numéro de la voiture ;
- Arrêt – identifiant de quai (arrêt par direction) ;
- Description – nom de l'arrêt ;
- Heure – heure de départ théorique depuis l'arrêt ;
- Position -A* – position de l'arrêt dans la ligne (commence à 1) ;

- Type Jr – type de jour (semaine, samedi ou dimanche) ;
- Direction – direction de voyage ;
- Voy. – numéro du voyage ;
- Ordinaire – 0 si horaire de vacances, 1 si hors des vacances.

La colonne « Ordinaire » a été rajoutée afin d'avoir une distinction entre les différents fichiers obtenus.

stops

Ce fichier contient le nom des arrêts, leur code, leur ordre dans la ligne ainsi que leurs coordonnées. Ce fichier est utile pour la géolocalisation des différents arrêts.

3.3.2. Jeu de données MétéoSuisse

Le jeu de données issu de MétéoSuisse contient les températures, les précipitations, le rayonnement, la date et l'heure pour la station de mesure de Pully. Une donnée est fournie par heure, ce qui fait 24 données produites par jour, sur 365 jours. Ces données ne sont pas disponibles sans en formuler la demande à MétéoSuisse.

3.3.3. Jeux de données construits

Un jeu de données de calendrier a été construit dans ce travail pour faciliter la jointure entre les différents fichiers. Il contient la date, le type de jour (semaine, samedi ou dimanche), les jours de vacances (0 ou 1), le jour de la semaine et le numéro de la semaine.

3.3.4. Structuration des données

Comme le dit Miller (2010), la difficulté avec les analyses de données est de jouer avec leur abondance. Les données ne sont pas récoltées pour répondre à des questions spécifiques et nécessitent donc une structuration plus ou moins profonde. Les données utilisées dans la base de données et pour l'application ont subi une profonde standardisation et plusieurs jointures afin d'être utilisables. Ces transformations sont expliquées au point 5.1.

3.4. Logiciels et langages de programmation

Dans les procédés de mise en forme, de stockage, d'exploitation et de visualisation des données, plusieurs logiciels et langages de programmation ont été nécessaires.

R

L'agrégation, la mise en forme et la première préparation des données a été effectuée avec R (version 3.6.3). R (https://www.R-project.org/) est un langage de programmation et un logiciel gratuit et *open source* destiné à la statistique et à la science de données.

SQL, PostgreSQL et Postgis

La création et la gestion de la base de données ainsi que le stockage des données ont été effectués avec PostgreSQL, version 12.2. PostgreSQL (https://www.postgresql.org) est un système de bases de données *open source* orienté objet, auquel il est possible d'ajouter des extensions. Pour le traitement des données spatiales, l'extension Postgis version 3.0.0 (https://www.postgis.net) a été utilisée. Les données stockées dans des bases de données offrent de nombreux avantages (Ramakrishnan & Gehrke, 2000) :

- Indépendance des données : les programmes qui interagissent avec une base de données doivent être aussi indépendants que possible ;
- Accès efficace aux données : les données peuvent être stockées et utilisées facilement ;
- Administration des données : les données sont centralisées et accessibles depuis plusieurs endroits ;
- Accès simultané et sauvegardes : des sauvegardes sont facilement effectuables ;
- Réduction du temps de calcul pour les applications : ce type de stockage permet un accès rapide à l'information, ce qui permet de stocker de nombreuses données et d'y avoir accès rapidement.

La base de données a été gérée à l'aide du langage Stuctured Query Language (SQL), en grande partie à l'aide de Postico (https://eggerapps.at/postico/), une application pour Mac qui permet de simplifier l'utilisation des bases de données PostgreSQL.

QGIS

QGIS version 3.12 (https://qgis.org/fr/site/) est un Système d'Information Géographique libre et *open source* qui a permis de créer et de visualiser le niveau géographique des données, c'est-à-dire la création des sections entre chaque arrêt de bus et leurs attributs. Ce logiciel permet aussi de se connecter directement à une base de données PostgreSQL.

Python

Python (version 3.7.7) est un langage de programmation interprété, orienté objet. Ce langage a été utilisé pour mettre en forme les données, mais il a aussi été très largement utilisé pour extraire les données de la base de données pour l'application de visualisation (voir chapitre 5.3).

HTML, CSS et JavaScript

Pour l'application de visualisation, ces trois langages de programmation ont été incontournables.

Le HTML (*Hypertext Markup Language*) est le langage utilisé pour créer le document d'une page web. Le HTML n'est pas un langage de programmation mais un langage de balisage (*markup language*) qui sert à identifier et à décrire les différentes composantes d'un document et sa structure.

Le CSS (*Cascading Style Sheets*) est utile pour soigner la présentation d'une page et quelques animations basiques.

Le JavaScript est un langage qui permet d'ajouter de l'interactivité et un certain comportement à une page web. Le JavaScript est utilisé pour manipuler des éléments dans une page et il est possible d'y ajouter des librairies qui permettent encore plus d'interactivité pour la page (Niederst Robbins, 2012). Ces trois langages sont illustrés à la figure 17.

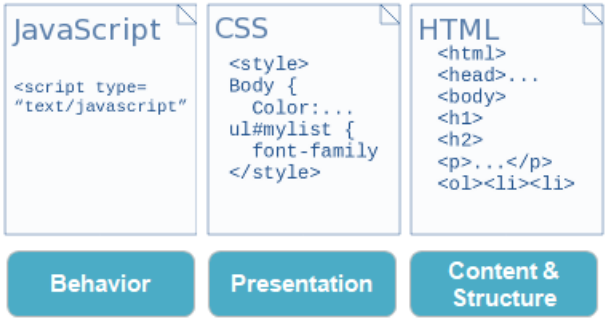


Figure 17 : Fonctions du JavaScript, CSS et HTML (repéré à https://www.listendata.com/2018/12/javascript-shiny-r.html)

Librairies JavaScript

Une librairie est une bibliothèque de fonctions qui facilitent le développement d'applications web (Khan Academy, 2018). Les principales librairies utilisées tout au long de ce travail sont multiples, en voici quelques exemples.

La librairie D3.js (https://d3js.org) a été utilisée pour la construction des graphiques principalement. Elle est très pratique pour la création de *dashboards* et de visualisations de données (Meeks, 2018).

Leaflet (https://leafletjs.com) permet l'intégration de cartes interactives et fonctionne facilement avec la plupart des navigateurs et aussi sur plateformes mobiles.

jQuery (https://jquery.com) permet la simplification de la manipulation transversale entre les diverses entités HTML, CSS et JavaScript.

4. Cadre opératoire

Cette partie décrit le processus de travail de ce mémoire en détaillant les étapes qui ont été nécessaires à la mise en forme des données et leur exploration ainsi qu'à leur visualisation.

L'objectif de ce projet étant de comprendre la structure des données des tl ainsi que de visualiser les occupations et retards des bus, les données à disposition ont dû être étudiées et modifiées pour avoir un format utilisable.

Les Transports Lausannois produisent quotidiennement des données grâce à des capteurs, notamment sur les marches des portes des bus afin de compter les passagers. Les déplacements sont également enregistrés dans le but d'obtenir les horaires réels de voyage.

Le public cible de ce travail sont les dirigeants des tl et dans une éventuelle version plus élaborée, les utilisateurs des transports publics.

4.1. Mise en forme des données

Objectif

La mise en forme des données a pour objectif de transformer des données quelconques dans un format utilisable et de les stocker sous forme de base de données. Les différents fichiers doivent pouvoir être joints pour être exploités ensemble.

Procédé

Les fichiers pris en compte étaient sous formes diverses, parfois en format Microsoft Excel, parfois en format CSV. Certains fichiers CSV ont des données qui comportent des virgules, alors que le séparateur est aussi la virgule. Un profond procédé de standardisation a été mis en place grâce des scripts R et Python dans le but de stocker tous les fichiers sous la même forme dans la base de données. Par exemple, il a fallu trouver un moyen pour joindre le fichier des horaires théoriques à celui les départs réalisés pour en sortir un calcul des retards. Pour ce faire, un script Python a été construit afin de passer en revue les horaires théoriques et de leur joindre un identifiant unique selon le type de jour ou s'il s'agissait de vacances par exemple.

Fonction

La mise en forme des données doit permettre d'interroger facilement et rapidement les données depuis une seule structure : la base de données. Cette base de données est ensuite interrogeable depuis différents programmes et langages de programmation, tels que R pour fournir des statistiques, QGIS pour la création de fichiers vectoriels ou encore Python et JavaScript pour l'application de visualisation.

4.2. Exploration des données

Objectif

L'exploration de données (EDA), a pour but de comprendre comment est articulé un jeu de données. Elle peut mettre en avant certaines irrégularités des données, des valeurs extrêmes ou des tendances.

Procédé

Une fois les données extraites et mises en forme, elles sont analysées avec R dans le but de produire des statistiques. Diverses tendances sont explorées ainsi que les valeurs extrêmes et les éventuelles corrélations.

Fonction

Une exploration du jeu de données permet de comprendre si les données sont correctement retranscrites et si elles sont utilisables, et quelles données doivent être exclues du jeu de données.

4.3. Visualisation

Objectif

L'objectif du *dashboard* est de visualiser des données et d'en tirer des informations claires, sans posséder de grandes notions en statistiques. Le tableau de bord permet aussi de télécharger les données mises en forme suivant les filtres désirés. L'application doit réunir les paramètres suivants :

- Choix de visualisation selon les jours ou selon les arrêts ;
- Visualisation des retards ;

- Visualisation de la charge ;
- Croisement des retards et charges de transport ;
- Choix des filtres.

Ces paramètres doivent être visibles clairement sur une page et les informations transmises par ces visualisations doivent être explicites, même pour les non-initiés à la thématique des transports.

Procédé

La construction de l'application est loin d'être la première étape de ce travail. L'exploration des données ainsi que leur mise en forme a permis préalablement de comprendre ce qu'il était possible d'élaborer comme visualisations. Il est aussi important de se baser sur les attentes du public cible pour conceptualiser la visualisation. Dans une première réflexion à propos de ce projet, il était d'abord question d'offrir un moyen de visualisation de l'état du trafic aux dirigeants, mais aussi aux utilisateurs des transports. Il a finalement été retenu de ne garder que l'agence de transport comme public cible. L'aide aux utilisateurs pourrait être une suite à ce travail.

Après la mise en forme des données, des premiers schémas ont été établis ainsi que quelques ébauches de graphiques illustrés à la figure 18. Le développement complet de l'application s'est fait de manière exploratoire et itérative grâce à l'exploration des données tout en se basant au mieux sur le cadre théorique.

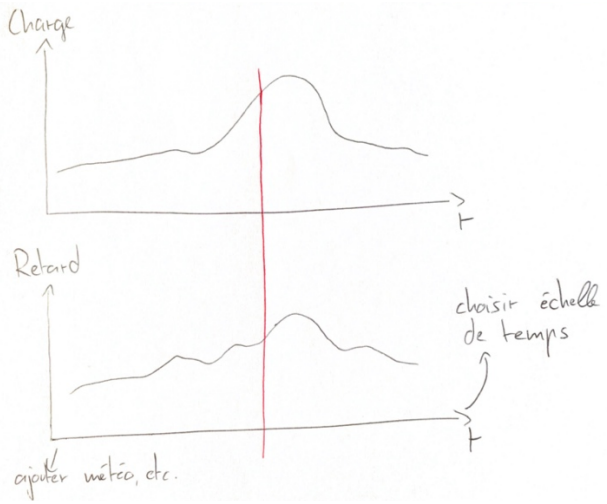


Figure 18 : Premières idées de visualisation

Google Chrome et Mozilla Firefox ont été utilisés pour tester le code de l'application, mais celle-ci peut être décodée par n'importe quel navigateur internet. De petites modifications d'affichage peuvent toutefois survenir. La structure de l'application repose sur un fichier HTML

(index.html), le style est défini dans le fichier CSS (style.css) et l'interactivité de la page est codée dans le fichier JavaScript (index.js). Un script Python (app.py) permet de faire la transition entre la base de données Postgres et la page de l'application. Ce script utilise le *framework* (ou infrastructure logicielle) Flask (https://palletsprojects.com/p/flask/).

Cette infrastructure permet de faciliter le développement d'applications en connectant des données à une page web via leur chemin d'accès. Le fonctionnement de l'application nécessite l'ouverture de plusieurs éléments : l'app.py connecte la base de données pour envoyer les données statiques utilisées par l'application sur une URL, qui sera `http://127.0.0.1:5000`, localisation par défaut sur un ordinateur personnel. Par exemple le chemin d'accès du serveur (localhost) suivi de `/delay/'2018-03-05'/'2018-03-22'/7/12/A/1/ma/0/30/18` permet d'accéder aux données des retards (`delay`) allant du 5 mars 2018 (`'2018-03-05'`) au 22 mars 2018 (`'2018-03-22'`), de 7 heures (`7`) à 12 heures (`12`), pour la direction « aller » (`A`), hors vacances (`1`), les mardi (`ma`), pour des précipitations journalières nulles ou plus abondantes (`0`), et une température journalière de 30°C ou moins (`30`), pour la ligne de transport n°18 (`18`). Les autres scripts permettent d'agencer la page et de choisir les différents filtres. C'est à partir des filtres de la page qu'il est possible de déterminer ces choix, comme le montre la figure 19. L'application a été codée et testée avec les données des lignes 2 et 18, mais le code est construit de telle manière à pouvoir ajouter des lignes de transports sans avoir à modifier le code. Seul le fichier index.html devra être modifié pour accueillir les boutons de sélection des éventuelles nouvelles lignes.

Figure 19 : Filtres de l'application, première version

Fonction

C'est grâce à ces divers filtres que les graphiques peuvent être construits et interagir entre eux. Les données sont aussi liées à la carte qui montre où passent les différentes lignes de bus ainsi que le retard à chaque arrêt, représenté par un cercle en symboles proportionnels. Le rayon du cercle est déterminé par la formule de Flannery ($r = k * N^{0.57}$), où k est une constante et N la valeur de l'indicateur (Flannery, 1956).

L'application et l'exploration de données ont pour but d'éclairer l'utilisateur sur la répartition du trafic et des retards ainsi que de fournir un outil de visualisation de données dynamique avec de multiples filtres permettant de choisir des périodes et des situations particulières à analyser. L'utilisateur peut donc choisir lui-même quels filtres appliquer. Il peut agir sur différents paramètres, comme les jours de la semaine, les heures, les vacances ou encore sur la météo par exemple. L'application a été nommée tlDataViewer.

La figure 20 résume la structure de travail de ce mémoire ainsi que les principales étapes de traitement des données. Elle offre un schéma de compréhension du processus de développement.

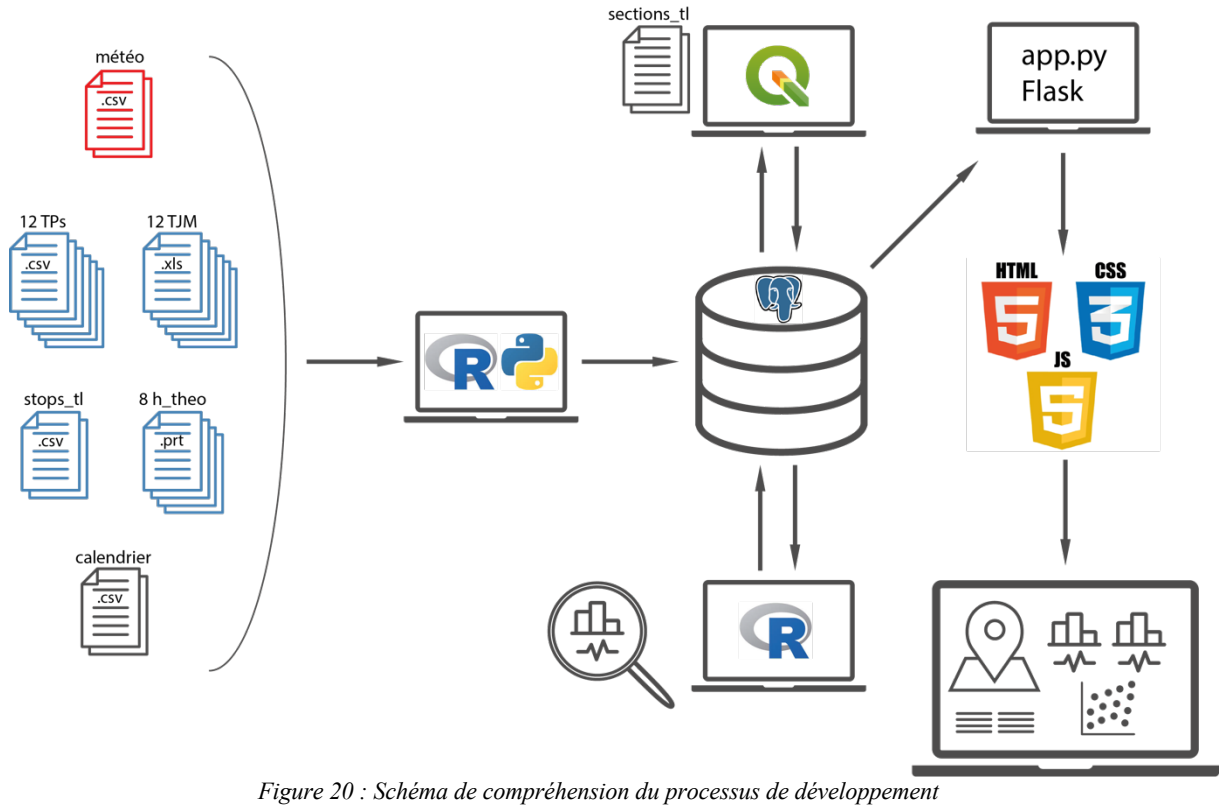


Figure 20 : Schéma de compréhension du processus de développement

5. Résultats

Ce chapitre étudie les différents traitements et analyses qui ont été effectués avec les données et met en exergue les principaux enseignements et résultats sous-jacents.

5.1. Mise en forme des données

5.1.1. Fichiers de base tl

Les données transmises par les tl ont dû largement être modifiées à des fins d'utilisation. Cette partie illustre, grâce à quelques exemples non exhaustifs, les manipulations qui ont été nécessaires pour que les données soient utilisables.

Les premières difficultés observées sont intervenues à la première lecture des fichiers de base issus des tl. Les jeux de données TJM et TPs n'ont pas la même numérotation des arrêts, il a donc été nécessaire de créer une nouvelle numérotation qui permette une correspondance entre les différentes tables. Les formats des tables sont par ailleurs différents (.xlsx et .csv), il n'est donc pas possible de les importer directement dans une base de données.

Les fichiers TPs ne comportent pas tous le même format pour les dates. Ces dernières sont écrites en français, et parfois avec des abréviations. Par exemple, il est possible de trouver « 1 janvier 2018 » ou « 22 mar. 18 ». Là aussi, une revue en détail des dates a été indispensable.

Les fichiers h_theo sont en format PRINT. Une refonte complète des données a été nécessaire pour une conversion en format CSV. Certains noms d'arrêts s'écrivent avec des virgules comme par exemple l'arrêt « Crissier, centre ». Il a fallu trouver un moyen pour différencier les virgules nécessaires à séparer les différentes colonnes des fichiers et celles présentes dans les noms d'arrêts.

Pour le fichier stops_tl, la numérotation des arrêts comporte un saut du numéro 23 au numéro 25. Cela a aussi dû être mis à jour pour permettre un bon fonctionnement de la base de données.

Toutes ces irrégularités montrent le besoin de rigueur nécessaire pour la mise en place de données utilisables pour diverses analyses, comme les statistiques ou la visualisation. Des identifiants

uniques pour chaque ligne de données permettent aussi de joindre plus facilement les diverses tables sans créer de doublons ou de fausses jointures.

Afin de créer une base solide utile à l'analyse, un modèle de données réfléchi, validé et documenté doit être établi. Les modèles de données de type relationnels ont déjà fait leur preuve. Ils sont formalisés et garantissent une utilisation des données dans des applications non définies à ce jour. Cependant, ces modèles doivent être documentés, car les données d'un domaine particulier comme les transports publics ont des spécificités qui lui sont propres. Le domaine des transports publics comporte de nombreuses particularités dont il faut tenir compte, comme par exemple des changements de lignes, des travaux sur la ligne, des courses supprimées ou encore des bus ajoutés ou remplacés. Ces particularités doivent être modélisées avec soin et documentées.

Certaines particularités semblent anodines mais ont une influence sur la validité des données, c'est pourquoi il est important de ne négliger aucun paramètre, de vérifier, de valider et de documenter un modèle de données. Ainsi, les heures de départ des bus ne peuvent être modélisées avec un type données « heure », car pour des raisons d'exploitation, les heures après minuit peuvent être représentées sous forme de 25h03 par exemple afin que les trajets débutés avant minuit finissent sur le même jour d'exploitation. En effet, le jour d'exploitation des transports publics commence aux alentours de 5h20 et se termine toujours sur le même jour d'exploitation.

5.1.2. SQL

Chaque fichier est entré dans la base de données via des commandes SQL. Toutes les tables, comme par exemple « tjm » et « tps », (Figure 21) qui contiennent respectivement les charges de passagers et les horaires réels réalisés, ont été créées de manière à pouvoir recevoir des données de nouvelles lignes de transport en tout temps sans changer la structure des requêtes.

```
---- CREATION DE TABLES

--Table "tjm" des charges des bus
CREATE TABLE "tjm" (
    "tjm_date" date NOT NULL,
    "tjm_line" int4 NOT NULL,
    "tjm_first_start" text NOT NULL,
    "count_trav_id" int8 NOT NULL,
    "tjm_direction" text NOT NULL,
    "tjm_position" int4 NOT NULL,
    "tjm_stop_code" text NOT NULL,
    "tjm_stop_name" text NOT NULL,
    "tjm_departure" text NOT NULL,
    "go_down" float4 NOT NULL,
    "go_up" float4 NOT NULL,
    "payload" float4 NOT NULL
);

-- Table "tps" des temps de passage réalisés.
CREATE TABLE "tps" (
    "tps_stop_code" text NOT NULL,
    "tps_stop_name" text NOT NULL,
    "tps_first_start" text NOT NULL,
    "distance" int4 NOT NULL,
    "tps_date" date NOT NULL,
    "tps_line" int4 NOT NULL,
    "tps_position" int4 NOT NULL,
    "section" text,
    "tps_direction" text NOT NULL,
    "trav_id" text NOT NULL,
    "tps_arrival" text NOT NULL,
    "tps_departure" text NOT NULL
);
```

Figure 21 : Création des tables « tjm » et « tps » en SQL

Un long procédé a été entrepris pour l'élaboration des calculs des retards. La figure 22 montre l'étape de création de la table contenant les retards. Il s'agit de l'utilisation de la colonne de l'horaire théorique qui a été préalablement ajoutée sur cette table et de la colonne du départ réel du bus, ceci pour créer une colonne du retard. Il a par ailleurs fallu modifier le format des données pour obtenir un retard en secondes. Ceci n'est qu'un exemple parmi d'autres.

```
281 create table tps_delay as
282 SELECT *, ((LEFT(theo_timetable, 2)::integer*3600)+
283     (RIGHT(theo_timetable, 2)::integer*60)) AS theo_timetable_s,
284 ((LEFT(tps_departure24, 2)::integer*3600)+(SUBSTR(tps_departure24, 4, 2)::integer*60)+
285     (RIGHT(tps_departure24, 2)::integer)) AS tps_departure_s,
286 ((LEFT(tps_departure24, 2)::integer*3600)+(SUBSTR(tps_departure24, 4, 2)::integer*60)+
287     (RIGHT(tps_departure24, 2)::integer))-((LEFT(theo_timetable, 2)::integer*3600)+
288     (RIGHT(theo_timetable, 2)::integer*60)) AS delay
289 FROM tps_theo;
```

Figure 22 : Une des étapes nécessaires à l'élaboration de la colonne du retard depuis la base « tps »

De nombreuses étapes de ce type ont été nécessaires à l'obtention d'une base de données utile pour l'élaboration de ce travail. Un échantillon des fichiers finals est disponible à l'annexe 4. Des index sur les tables les plus utilisées ont été créés afin d'optimiser la vitesse d'exécution de la base de données.

5.2. Exploration des données

L'exploration des données a pour but ici de comprendre leur nature, de savoir où se situeraient d'éventuelles anomalies et quelles sont les principales statistiques de bases du jeu de données. Ces analyses doivent aussi permettre de mieux pouvoir interpréter les visualisations produites par l'application. Les statistiques et graphiques sont produits avec le logiciel et langage de programmation R. Les couleurs utilisées pour les diverses visualisations sont les couleurs officielles des lignes des tl.

La première partie des tests est effectuée sur le jeu de données en entier par rapport aux indicateurs de la charge et du retard. Des statistiques descriptives ainsi qu'une série de graphiques sont effectués dans ce chapitre. Afin de comparer la charge et les retards, les données doivent être agrégées par une unité de temps comme le jour, la semaine ou le mois ou encore par arrêt pour que les données des deux tables puissent avoir un élément de comparaison.

5.2.1. Caractéristiques, charge et retard

Le nombre de passagers (fichier TJM) est un des indicateurs les plus importants du jeu de données. La deuxième valeur clé est le retard (fichier TPs). Il est primordial d'avoir un aperçu de la répartition de ces deux valeurs par ligne de transport, par période de temps ou encore par arrêt pour en tirer des premiers enseignements.

Ligne	**2**	**18**	**Total**
Nombre de données charge	1 771 568	1 360 447	3 132 015
Nombre de données retard	1 918 839	1 452 427	3 371 266
Distance parcourue en km	565 856	426 522	992 377
Distance ligne – Aller en km	7.764	5.593	13.357
Distance ligne – Retour en km	8.06	5.86	13.99
Charge annuelle en passagers	3 755 297	3 343 507	7 098 804
Retard moyen en secondes	164.67	127.15	148.38
Retard médian en seconde	101	86	93

Tableau 1 : Statistiques de base des deux lignes

Le tableau 1 regroupe les statistiques de bases des données. Il résume les principales caractéristiques des jeux de données empiriques des tl. Sur les deux tables de la charge (TJM) et du retard (TPs), la base de données comporte plus de 6 500 000 lignes de données récoltées pendant l'année 2018. Chaque ligne de donnée d'un des deux jeux représente le passage d'un bus à un arrêt ; les bus des lignes 2 et 18 ont passé environ 3 millions de fois à un arrêt en 2018. Les deux fichiers (TPs et TJM) ne comportent pas exactement le même nombre de données, car les capteurs placés dans les bus ne sont pas identiques et certaines données sont absentes parce qu'elles n'ont pas passé le test de qualité des tl. D'autres lignes de données ne représentent pas des trajets commerciaux qui transportent des passagers et n'ont pas été prises en compte pour les divers calculs. Les données manquantes ne représentent qu'environ 1% du total.

Rien que sur l'échantillon des lignes numéros 2 et 18, plus de 7 millions de passagers ont emprunté les lignes de transports tout au long des 27 kilomètres représentant les deux tracés. Première statistique intéressante : les retards moyens sont largement supérieurs aux retards médians, les valeurs extrêmes ont de l'influence sur les retards. Un choix a donc été fait de ne prendre en compte que 90% du centre des valeurs pour certains graphiques de répartition, car ce choix limite les valeurs extrêmes qui ont un effet sur les répartitions. N'utiliser que 90% des valeurs

pour l'application de visualisation n'est par contre pas justifié, car il s'agit dans cette partie d'avoir un accès complet aux données (développement au point 5.3 « Visualisation »).

5.2.2. Répartition du retard

Les figures 23, 24 et 25 mettent en évidence la répartition des retards et des avances sur 90% des données centrales, 5% des données ont été retirées, respectivement sur l'extrême gauche et l'extrême droite du jeu. Il est en effet plus efficace de supprimer les *outliers,* qui résultent de problèmes exceptionnels sur la ligne ou de mauvaise prise de données, pour comprendre la répartition des retards habituels. Que ce soit pour les données totales, pour les lignes 2 ou 18, ces graphiques ont tous la même architecture : les retards sont concentrés vers la gauche, la moyenne (traitillé gris) est plus grande que le mode[2]. La répartition des retards de la ligne 18 a une distribution statistiquement un peu plus symétrique. Cela indique que les aléas qui provoquent de grands retards sur la ligne 18 sont inférieurs à ceux de la ligne 2. La ligne 2 est probablement plus souvent perturbée par un fort trafic, ou par des manifestations par exemple. Malgré les valeurs centrales des 90%, le coefficient d'asymétrie est positif. Il faut s'attendre à ce que peu de grands retards influencent la répartition des données en général. Il n'y par contre que peu de trajets qui comportent de l'avance.

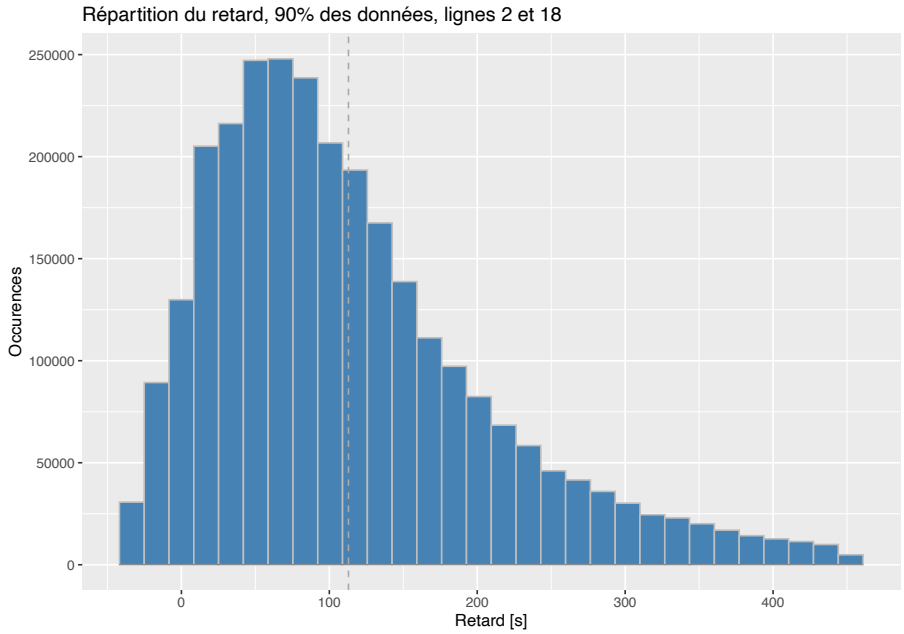


Figure 23 Répartition du retard pour 90% des données centrales, sur les lignes 2 et 18

[2] Mode : valeur la plus représentée dans la distribution

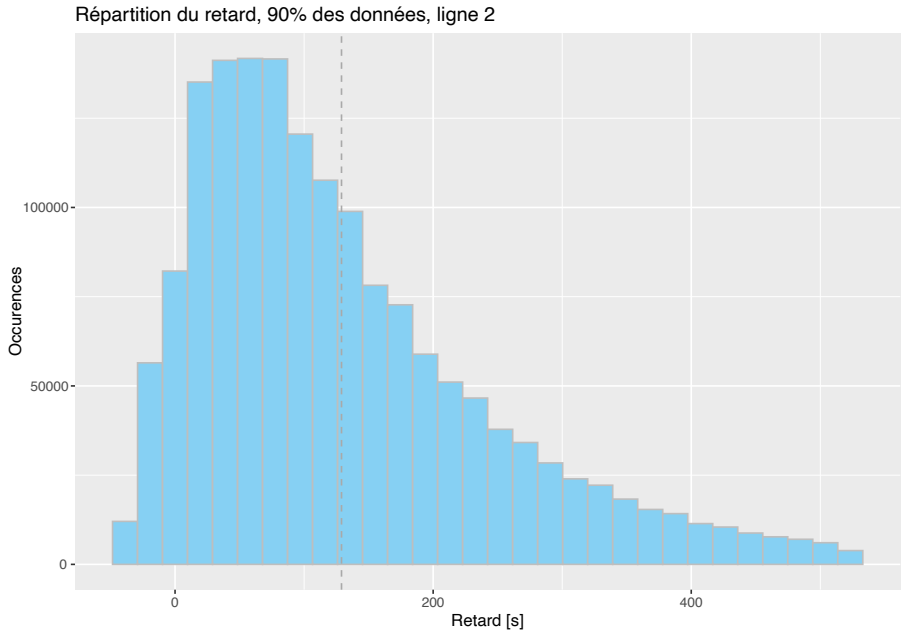


Figure 24 Répartition du retard pour 90% des données centrales, sur la ligne 2

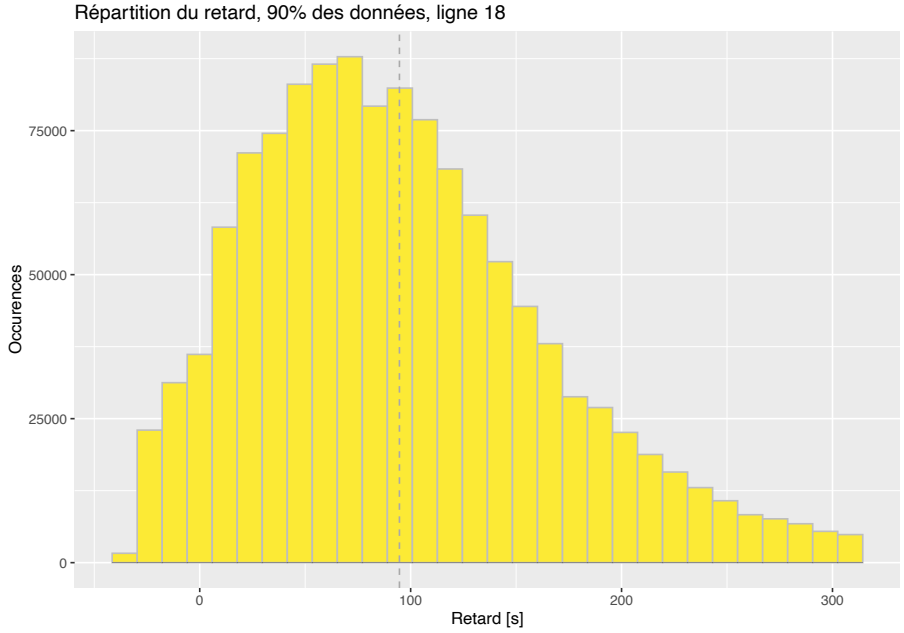


Figure 25 Répartition du retard pour 90% des données centrales, sur la ligne 18

5.2.3. Test de Student

Un test de Student a été effectué pour comparer les moyennes des échantillons pour les valeurs centrales du retard, afin de comprendre si les moyennes des retards sont les mêmes entre les lignes 2 et 18 et quels sont les intervalles de confiance à 95%. Le test du *t* de Student évalue la significativité de l'amplitude d'une distribution par sa tendance (Yue & Pilon, 2004). Sans surprise, les retards n'ont pas la même moyenne par jour pour les deux lignes. Pour l'échantillon au complet, les retards sont plus élevés de 36.9 à 38.1 secondes à 95% de fiabilité pour la ligne 2.

Pour l'échantillon comportant les 90% des données centrales, ces valeurs sont un peu plus faibles : l'intervalle est de 34.1 à 34.5 secondes à 95%.

Il est important de noter qu'un test de Student est calibré pour des distributions normale, mais donne tout de même des résultats satisfaisants pour des données non normales (Santiago, 2015). La normalité des données est testée plus loin dans le rapport.

5.2.4. Statistiques descriptives de la charge et du retard

Les tableaux 2 et 3 résument des statistiques descriptives concernant la charge et le retard par ligne en sommant les données sur les jours.

Charge journalière			
Ligne	2	18	Total
Minimum	2 924	2 670	5 594
Maximum	16 198	13 320	28 756
Moyenne	10 289.11	9 161.03	19 450.14
1er décile (10%)	5 662.54	4 013.66	9 723.06
Quartile inférieur (25%)	8 407.19	7 383.00	16 016.10
Médiane (50%)	11 448.39	9 771.60	22 035.23
Quartile supérieur (75%)	12 275.40	11 742.69	23 887.32
9ème décile (90%)	13 061.72	12 066.42	24 739.81
Écart-type	2 894.92	2 941.18	5 607.27

Tableau 2 : Statistiques descriptives de la charge journalière

Retard journalier moyen [s]			
Ligne	2	18	Total
Minimum	3.69	28.93	3.69
Maximum	484.05	765.66	765.66
Moyenne	163.63	125.83	144.73
1er décile (10%)	101.18	99.68	99.72
Quartile inférieur (25%)	121.50	107.98	111.84

Médiane (50%)	152.94	118.16	127.94
Quartile supérieur (75%)	191.73	133.07	164.07
9ème décile (90%)	237.22	155.07	205.68
Écart-type	62.22	46.38	58.00

Tableau 3 : Statistiques descriptives du retard journalier moyen

Les valeurs résumées dans ces deux tableaux permettent de comprendre plus facilement la nature des occupations et des retards des lignes étudiées, afin de pouvoir ensuite interpréter les visualisations en connaissant certaines caractéristiques des données.

Concernant tout d'abord la charge, la ligne numéro 2 est un peu plus fréquentée que la ligne 18, alors que son écart-type est moins important. Pourtant, l'intervalle interquartile est plus faible pour la ligne 2 que pour la ligne 18 (3868.21 contre 4359.69). La ligne 2 présente plus de valeurs extrêmes, mais la majorité des données sont regroupées autour de la médiane. Cette observation est illustrée par le boxplot de la figure 26.

Les retards moyens par jours sont répartis un peu différemment que les charges. Toutes les observations sont regroupées plus proche de la médiane, mais les *outliers* sont plus nombreux. En effet, les boxplots de la figure 27 montrent des boites plus compactes, avec des moustaches plus courtes. Par contre, de nombreux *outliers* influencent la valeur des moyennes vers le haut, que ce soit pour la ligne 2 ou 18. Ces points sont des jours isolés et représentent probablement des évènements isolés dus à des manifestations, des conditions météorologiques spéciales ou à des problèmes techniques généraux. Ces situations particulières peuvent être visualisées par l'application décrite au point 5.3 « Visualisation ». Les distributions de la charge et des retards sont plus symétriques pour la ligne 18 que la ligne 2.

Il serait intéressant d'avoir les données de toutes les lignes du réseau de transport afin de calculer des scores standardisés (centrés-réduits). Se baser sur les données totales permettrait de comparer plus facilement les lignes entre elles et de savoir quelles sont les lignes qui posent le plus de problèmes.

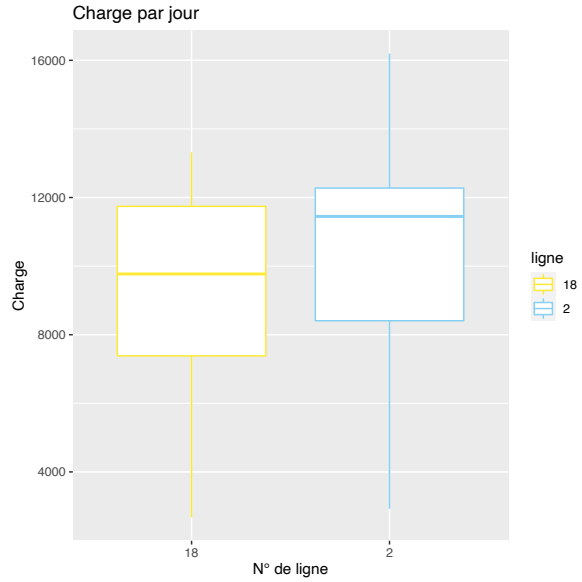


Figure 26 : Boxplot de la charge par jour

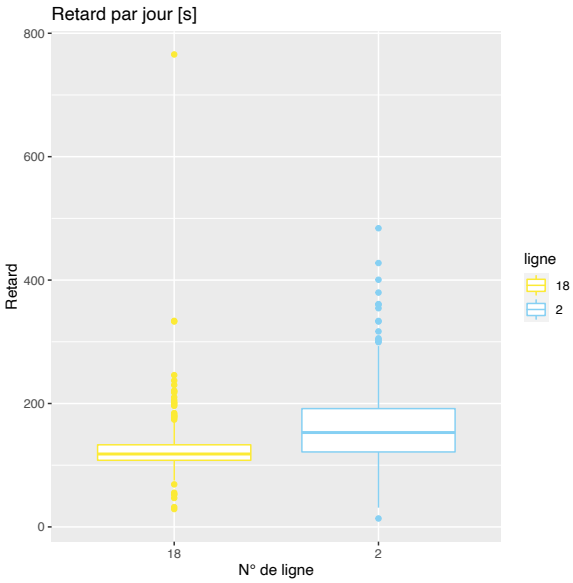


Figure 27 : Boxplot du retard par jour

5.2.5. Test de Shapiro-Wilk et normalité

Un test statistique de Shapiro-Wilk a été effectué pour tester la normalité des différentes distributions par jour. Pour la répartition de la charge ou du retard, les deux résultats du test indiquent une répartition non normale d'un point de vue statistique, car la valeur p est très petite ($p < 0.01$).

Il est aussi possible de vérifier aussi graphiquement la normalité de distributions de données. Les figures 28, 29, 30 et 31 illustrent ces propos. La figure 28 décrit la répartition de la charge et qui est principalement concentrée autour des 24 000 passagers par jour. La densité n'est pas normale car la distribution est asymétrique à gauche. La figure 29 montre l'intervalle (partie grisée) de valeurs suivant une loi normale standardisée. Les valeurs montrent une plus forte densité après la moyenne. Pour le retard, la distribution (Figure 30) tend cette fois vers la droite avec une forte concentration vers 150 secondes. Graphiquement, la distribution standardisée du retard (Figure 31) s'éloigne toujours d'une distribution normale plus l'attente est prolongée. Les valeurs extrêmes ont ici une grande influence sur la normalité de la distribution.

En général, les répartitions de la densité des deux indicateurs ne sont pas régulières et les illustrations du test de normalité montrent des profils différents de ceux d'une répartition suivant une loi normale.

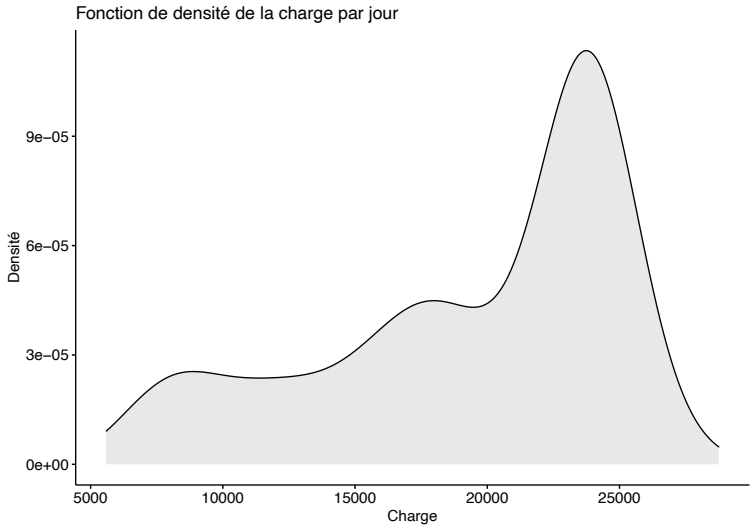


Figure 28 : Fonction de densité de la charge par jour

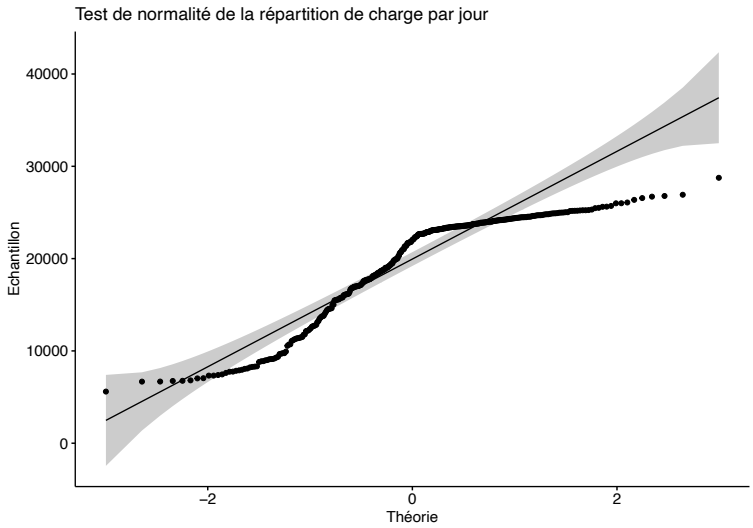


Figure 29 : Test de normalité de la répartition de la charge par jour

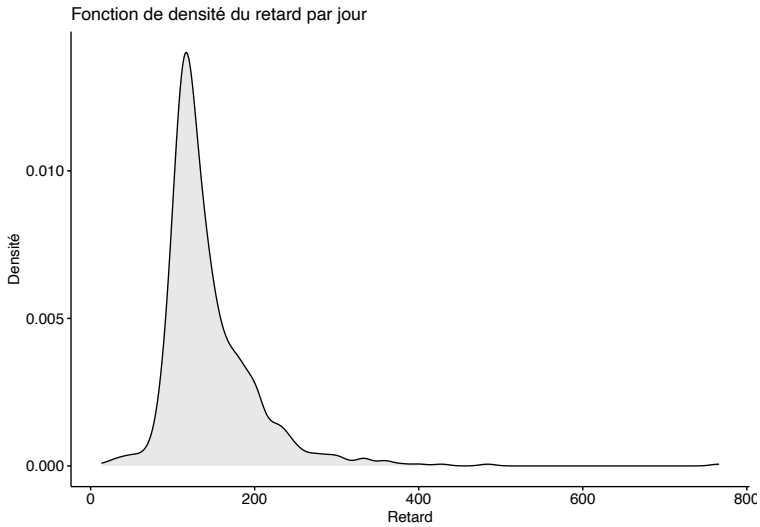


Figure 30 : Fonction de densité du retard par jour

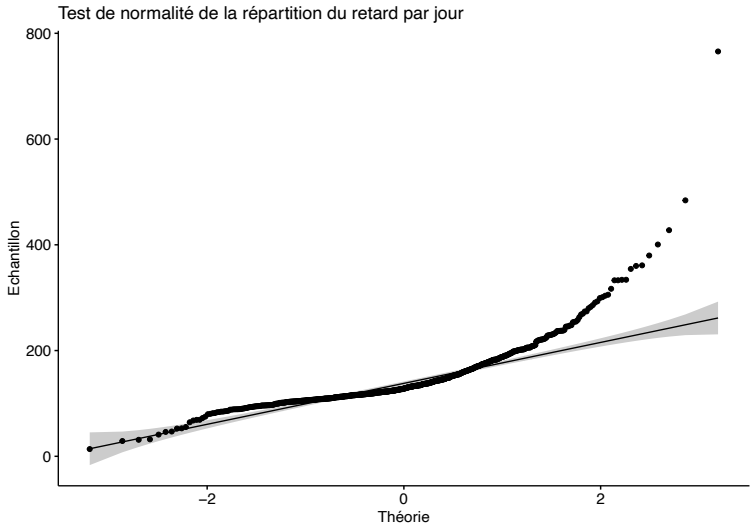


Figure 31 : Test de normalité de la répartition du retard par jour

5.2.6. Corrélations et comparaisons

Un indicateur simple comme la corrélation permet de comprendre si deux données sont liées (Bavaud, 1998). Plusieurs corrélations ont été effectuées pour se rendre compte si les charges et retards sont liés. Tout d'abord, une comparaison entre les deux lignes étudiées aide à comprendre l'articulation du réseau, afin de se rendre compte si une des caractéristiques statistiques dépend de la ligne ou non. La corrélation de la charge par jour entre les lignes 2 et 18 est de 0.85. Le lien ici est indéniable, l'occupation du réseau ne dépend pas de la ligne pour cet échantillon. Par contre, la corrélation entre les retards journaliers pour ces deux lignes ne suit pas la même tendance : la corrélation est de 0.36, donc seulement faiblement positive. Cette observation est similaire quant à la comparaison entre la charge et le retard par jour, la corrélation est de 0.34, donc aussi seulement faiblement positive. La significativité est acceptable pour chaque corrélation.

Pour ce type d'analyse, il n'est donc pas possible d'affirmer que le retard est dû aux fortes charges, même si la tendance n'est pas complètement nulle pour ces indicateurs.

Les graphiques des figures 32 et 33 permettent de comparer visuellement les charges et les retards sur douze mois pour les lignes 2 et 18. La ligne 2 présente de plus grandes différences inter mensuelles que la ligne 18. Comme la ligne 2 a la caractéristique de transporter les utilisateurs de la ville vers le lac, elle est probablement plus sujette aux différences saisonnières que la ligne 18 qui a un tracé est-ouest de la ville à la banlieue. La fréquentation de la ligne 2 est croissante depuis le mois d'avril jusqu'au mois de juin. Il est possible d'imaginer que les utilisateurs utilisent les

transports publics pour leurs loisirs et se rendre au bord du lac pour le début de l'été. La fréquentation chute par contre pendant la pause estivale. La ligne 18 a une fréquentation plus régulière, mais une baisse de passagers est à observer durant les mois de juillet août. Le graphique des retards médians (Figure 33) montre une plus grande différence entre la ligne 2 et la ligne 18 (cf. corrélation). La répartition du retard dans le temps est plutôt régulière pour la ligne 18, alors qu'il culmine en juin pour la ligne 2. Il y a clairement une influence de l'été dans les retards pour la ligne 2, ce qui pourrait être testé dans une suite à ce travail.

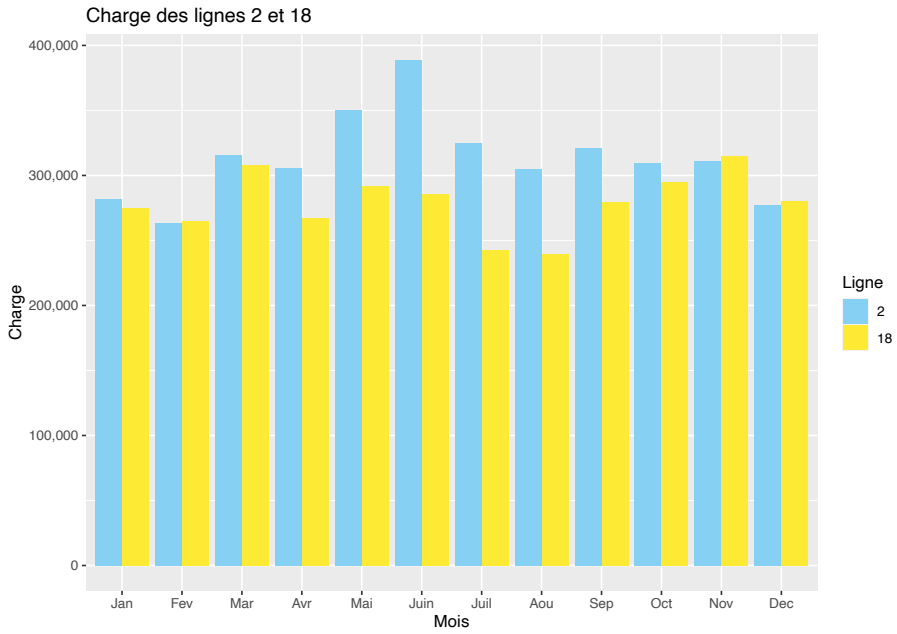


Figure 32 : Charge des lignes 2 et 18

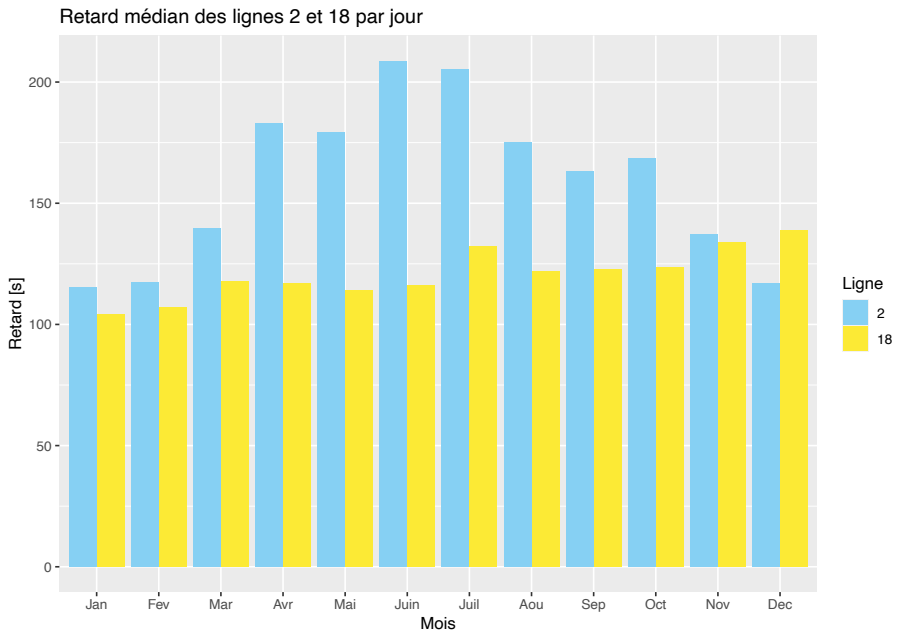


Figure 33 : Retard médian des lignes 2 et 18

Pour clore cette partie, des graphiques en nuages de points (Figures 34 et 35) ont été effectués afin de comparer la charge de transport avec le retard pour les deux lignes étudiées. Ce type de graphique est aussi disponible avec l'application. Le premier graphique (Figure 34) fait état des données agrégées par jour des lignes 2 et 18, alors que le second graphique (Figure 35) individualise la comparaison pour chaque ligne. Une courbe de régression ajustée avec un intervalle de confiance à 95% a été ajoutée afin de comprendre la tendance locale qu'il y a entre charge et retard.

Pour des faibles valeurs de charge, le retard est aussi assez faible, que ce soit pour la ligne 2 ou la ligne 18. Pour les valeurs centrales, il est plus difficile d'articuler une tendance, car la majorité des points est groupée vers la courbe de tendance, mais quelques *outliers* avec de plus forts retards sont à noter. Lorsque les charges commencent à être plus importantes, à partir du troisième tiers des graphiques, la courbe de tendance part à la hausse. Cela montre que lors de fortes affluences, il est plus difficile pour les lignes de bus de garder la cadence et de suivre l'horaire. Une comparaison intéressante est à faire entre la ligne 2 et la ligne 18 : les deux courbes de tendance se suivent tout au long des observations. La principale différence entre la ligne 2 et la ligne 18 est que la ligne 2 est plus fréquentée. C'est à partir de là où la courbe de tendance de la ligne 18 s'arrête que celle de la ligne 2 part à la hausse. Il est imaginable que si la ligne 18 avait la même fréquentation que la ligne 2, elle suivrait probablement cette tendance de retard à la hausse pour une fréquentation à la hausse.

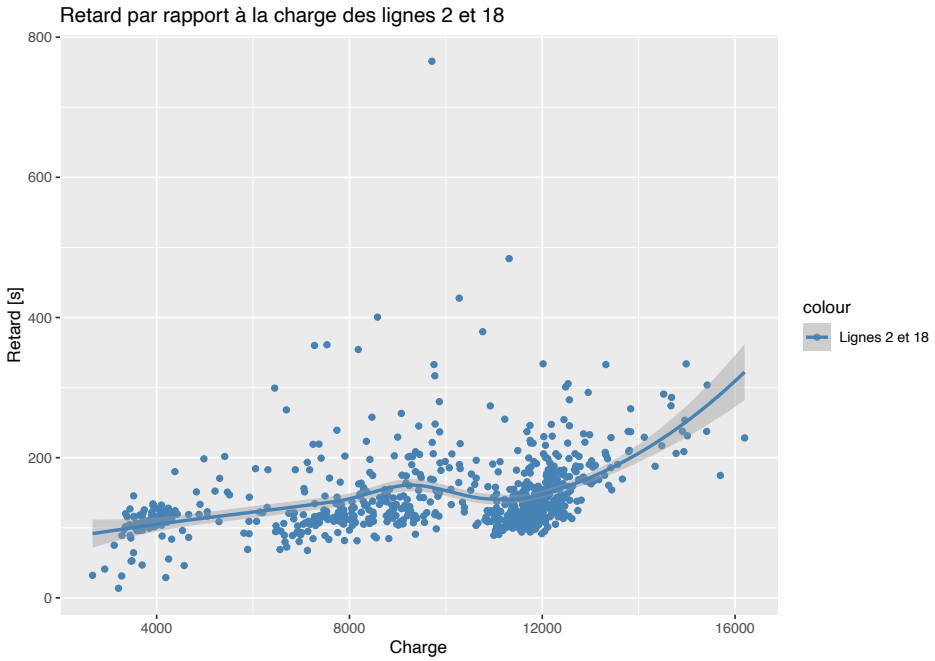


Figure 34 : Retard par rapport à la charge, courbe de tendance sur le total

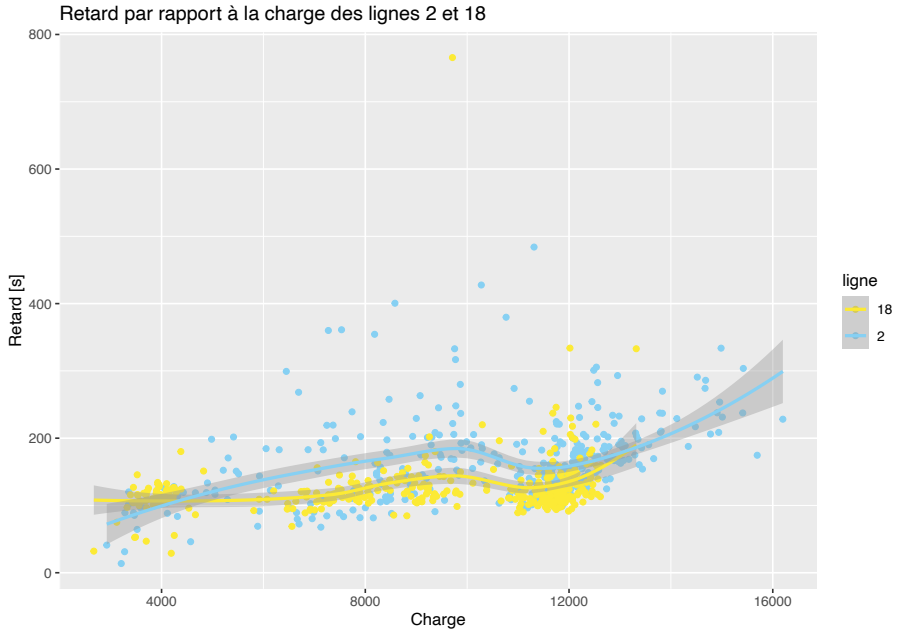


Figure 35 : Retard par rapport à la charge, courbe de tendance entre les lignes 2 et 18

5.2.7. Météo

Le facteur de la météo peut jouer un rôle sur l'utilisation des transports publics et donc une influence sur leurs occupation et retards. Selon l'Organisation Météorologique Mondiale (2017), une précipitation journalière est comptabilisée si elle dépasse 1 mm par jour. Il existe cependant plusieurs seuils de précipitation à 5 et 10 mm par exemple. Il a été décidé d'utiliser des seuils de 1, 5 et 10 mm de précipitations journalières pour comparer les charges et retards des deux lignes. Un seuil de températures journalières moyenne de 3°C a été choisi comme jour de risque de neige afin d'identifier d'éventuelles perturbations sur les temps de trajets des bus.

Individuellement, des précipitations ou une température faible n'ont pas d'influence sur le retard journalier annuel, comme illustré sur le graphique de la figure 36 et par les valeurs du tableau 4. Par contre, lorsque les précipitations peuvent tomber sous forme de neige, les retards sont de plus en plus importants suivant le cumul en une journée. Sans surprise, les retards sont influencés par la neige. Ce type de visualisation est aussi possible via l'application décrite dans la partie 4.3.

Retard annuel moyen [s]	**Sans filtre de température**		**Température < 3 °C**	
Ligne	2	18	2	18
Sans contrainte de précipitations	164.67	127.16	130.70	130.2
Précipitations > 1mm jour	153.68	137.88	195.35	215.13
Précipitations > 5 mm par jour	152.02	144.42	300.73	377.85
Précipitations > 10 mm par jour	161.81	145.36	484.05	765.66

Tableau 4 : Caractéristiques des retards liés à la météo

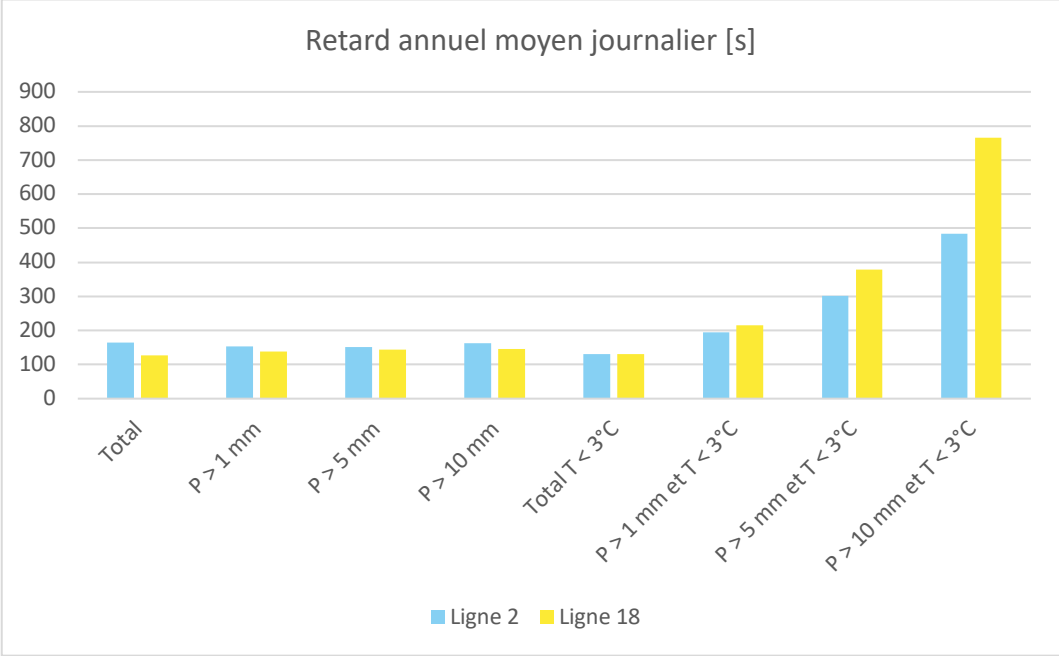


Figure 36 : Retard annuel moyen selon la météo

5.3. Visualisation

5.3.1. Fonctionnement tlDataViewer

L'application interactive de visualisation a été créée dans le but de lire facilement des données sur un seul écran. Elle se compose de deux variantes dotées de fonctions similaires, mais de filtres de visualisation différents : la première permet de visualiser les données par jour (Figure 37), alors que la seconde offre à peu près les mêmes fonctions, mais par arrêt (Figure 38). L'application est, pour ce projet, uniquement disponible sur l'ordinateur qui a servi à sa conception, un MacBook Pro 15". L'interface de la page est composée de quatre parties principales :

- une carte interactive en haut à gauche ;
- une série de filtres dans un menu latéral déroulant ;

- les visualisations sur la partie droite de l'écran ;
- des statistiques de base en bas à gauche.

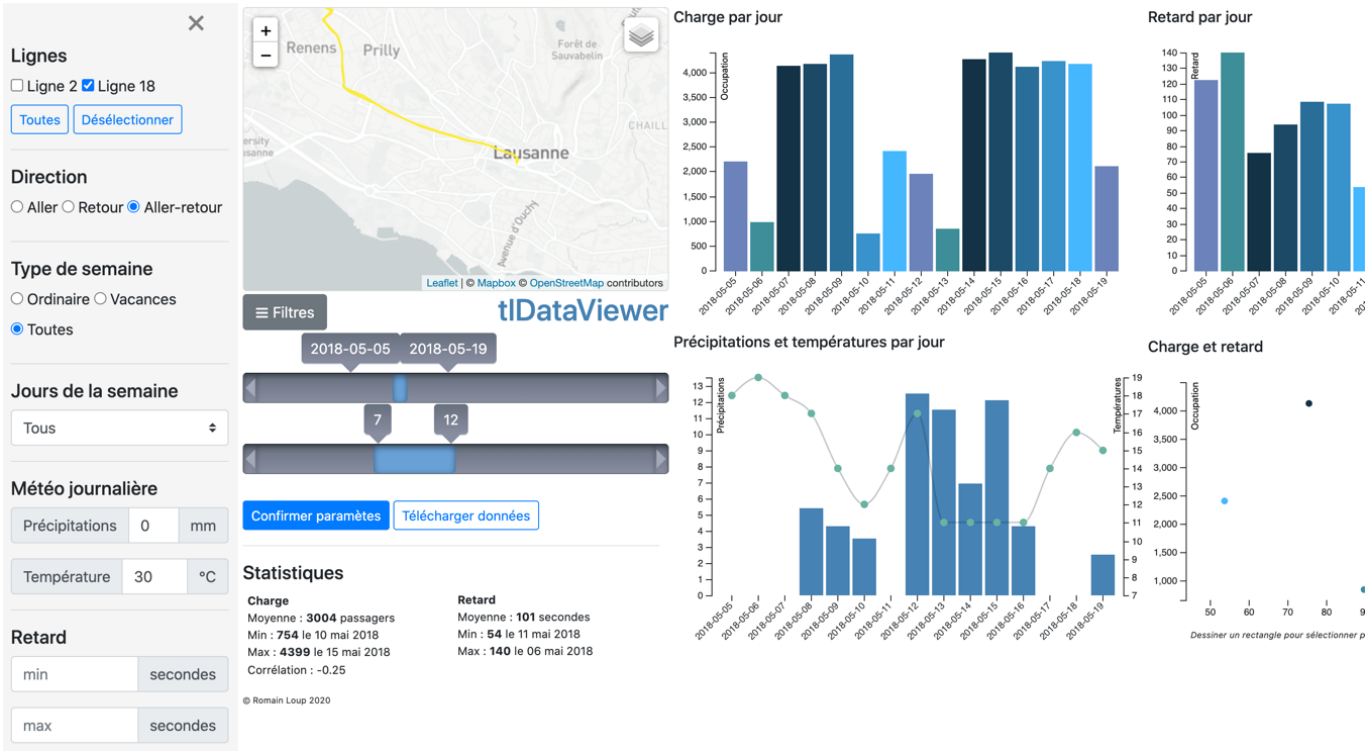


Figure 37 : Page d'accueil de la version par jour, avec le menu des filtres ouvert

Figure 38 : Page d'accueil de la version par arrêt, version 1

En se rendant sur la page, l'utilisateur arrive sur une visualisation par défaut. Cette première visualisation lui permet d'avoir directement un aperçu du fonctionnement de l'application et de découvrir les diverses possibilités qu'offrent les filtres qu'il peut activer. Il lui est ensuite possible de jouer avec ces différents paramètres.

En cliquant sur le bouton « Confirmer paramètres » (Figure 39), il peut lancer les visualisations de son choix. Les transitions des barres et des points après une sélection se font en douceur sans rupture grâce à la librairie JavaScript D3. La carte affiche les différentes lignes de bus sélectionnées. En déplaçant la souris sur les divers tronçons affichés sur la carte, il verra s'afficher le nom de ces derniers. Une autre variante illustre les statistiques par arrêt : il est aussi possible de passer la souris sur les cercles représentants les retards aux arrêts pour en voir les caractéristiques.

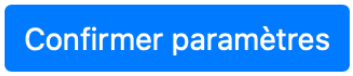


Figure 39 : Lancement de la visualisation

Les filtres sont les suivants (Figure 37) :

- numéros de lignes : 2, 18 ou les deux dans cette version ;
- direction du trajet ;
- type de semaine ;
- jour de la semaine ;
- météo journalière, précipitations et températures ;
- intervalle des dates ;
- intervalle des heures ;
- retard minimal et maximal en secondes (Pour la version « jours »).

Les codes couleurs sont les suivants : les jours de la semaine sont affichés selon une échelle de bleus allant de foncé à clair du lundi au vendredi, le samedi est en violet et le dimanche en turquoise. Ces couleurs se retrouvent sur la sélection du menu déroulant des jours (Figure 40) (cela dépend du navigateur) et sur les différents graphiques.

Figure 40 : Jours et couleurs correspondantes

Des statistiques de bases sont alors produites ainsi que les graphiques de l'occupation, du retard, de la météo journalière et un nuage de points, avec comme entrées la charge de transport et le retard. Chaque graphique est affiché selon les filtres sélectionnés. Une fois les graphiques affichés, il est possible de déplacer la souris sur les différentes barres et points, afin de voir la correspondance des données entre les différents graphiques (Figure 41). Les caractéristiques de l'élément sélectionné sont alors visibles dans un *tooltip*, une petite fenêtre qui s'ouvre à côté de la souris.

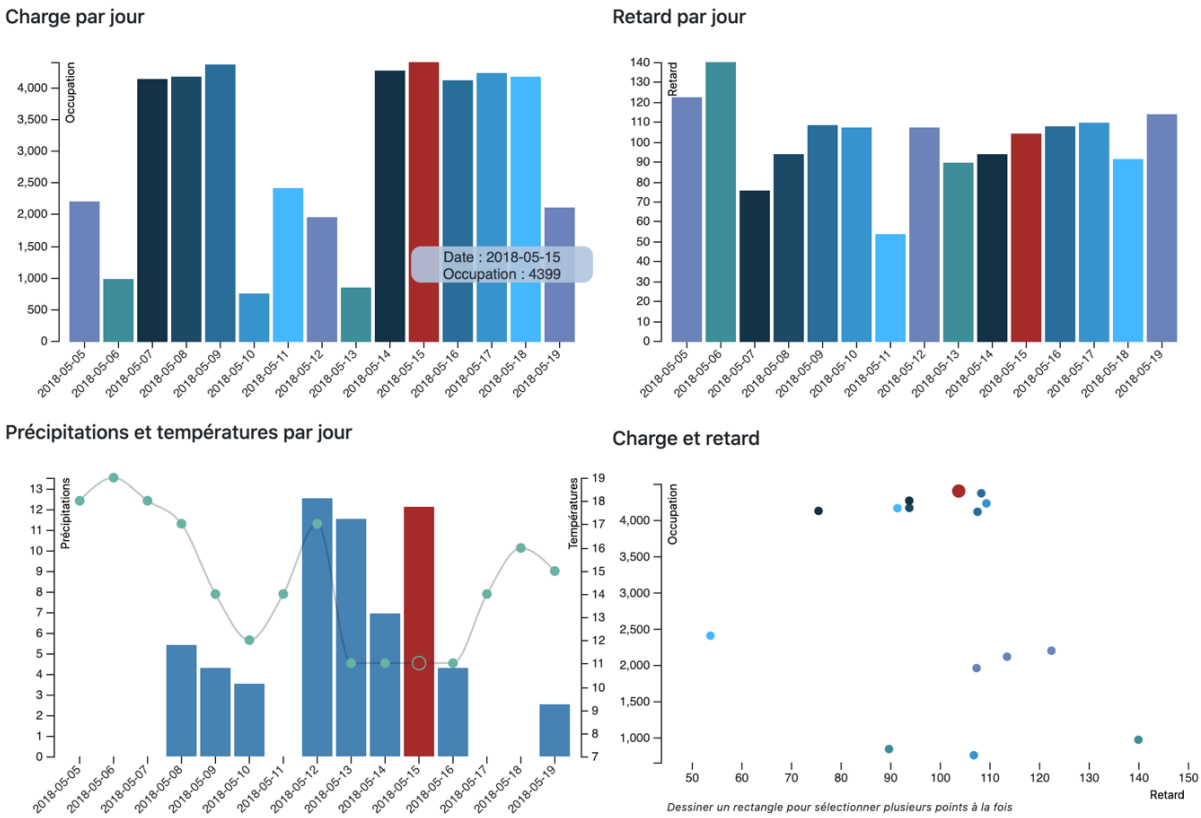


Figure 41 : Correspondance entre les différents graphiques et tooltip

Il est aussi possible de créer un rectangle sur le graphique du nuage de points, afin de sélectionner plusieurs points et de voir où ces derniers se situent sur les différents graphiques (Figure 42).

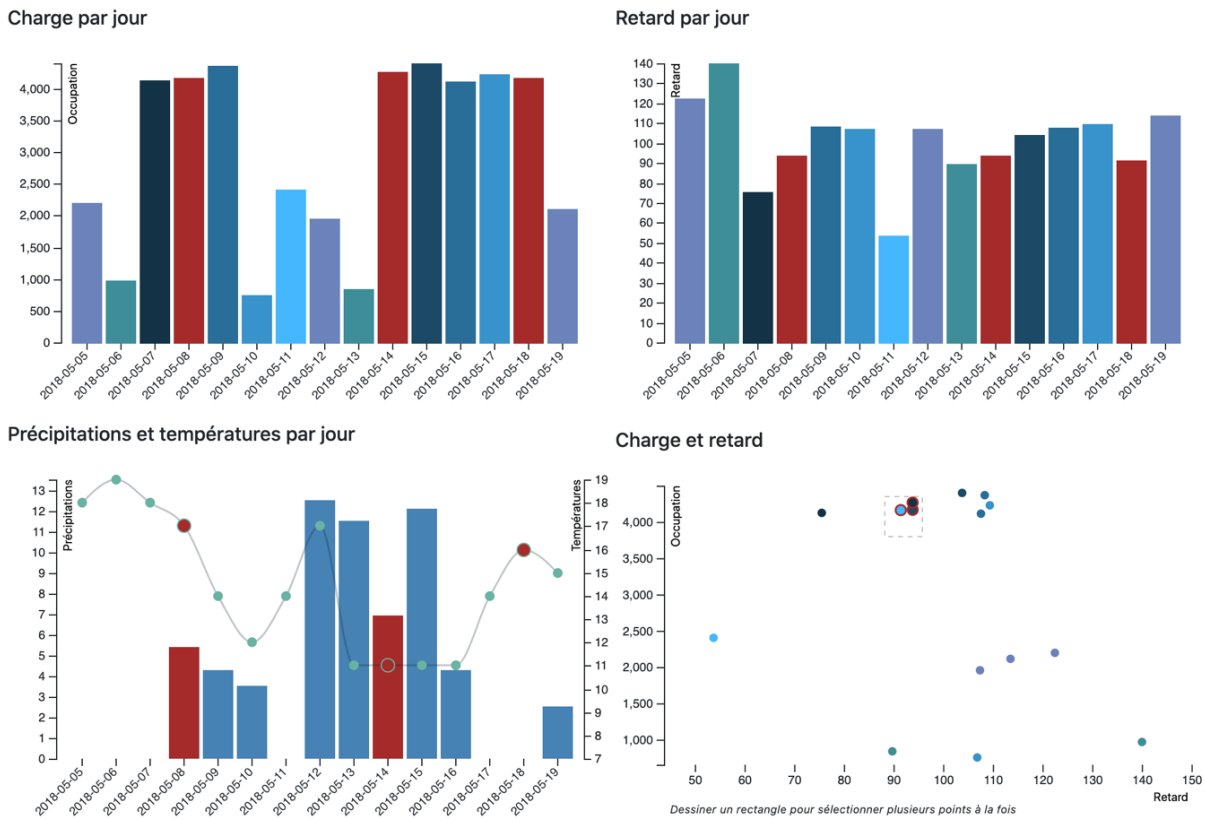


Figure 42 : Rectangle de sélection

Les données qui ont été utiles pour la création des diverses visualisations sont téléchargeables au format CSV via un bouton « Télécharger données ». Le fichier est nommé grâce aux différents filtres utilisés. La figure 43 illustre la présentation de ce fichier.

date	charge	retard
2018-05-05	2196	122.49710144927536
2018-05-06	967.80005	140.03630363036302
2018-05-07	4124.599	75.55886850152905
2018-05-08	4166.4	93.86383601756954

Figure 43 : Exemple de fichier téléchargé via l'application

Sur la carte *Leaflet*, il est possible de changer de fond de carte en choisissant une version dépouillée ou une version d'images satellites. Le niveau de zoom peut aussi être changé. Un fond de carte simple permet de faire mieux ressortir les tracés des lignes de bus. Ces préférences sont directement activables depuis la carte interactive.

Pour la version agrégée par jour, une sélection de plus de deux mois de données engendre un empilement des barres des jours sur les graphiques pour alléger la lecture. L'axe X n'affiche alors

plus les jours, mais les numéros de semaine. Une semaine non complète n'affiche que les jours choisis dans la sélection. La figure 44 illustre une situation où le nombre de jours sélectionnés dépasse 2 mois de données.

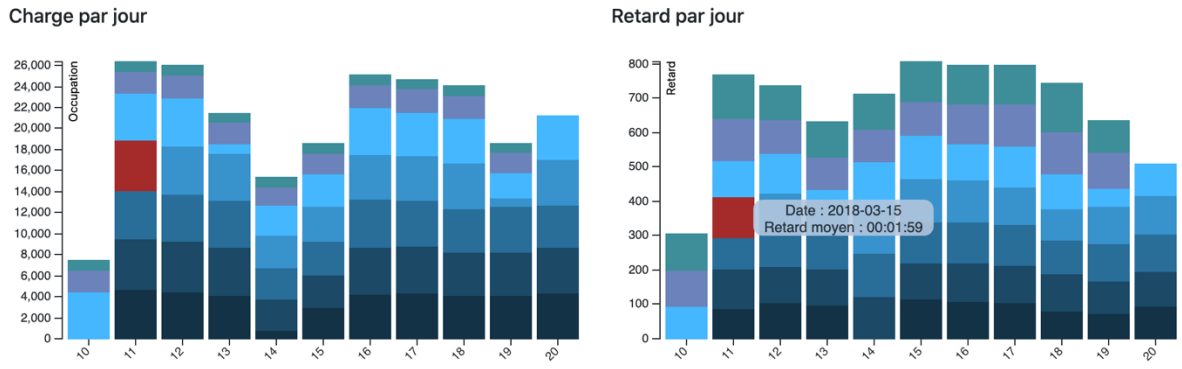


Figure 44 : Données de charge et retard sur plus de deux mois

Pour la version agrégée par arrêt, l'application montre le retard à un arrêt sous forme de symbole proportionnel. En passant la souris par-dessus un symbole, le retard ainsi que le nom de l'arrêt apparaissent sur un *tooltip* (Figure 45).

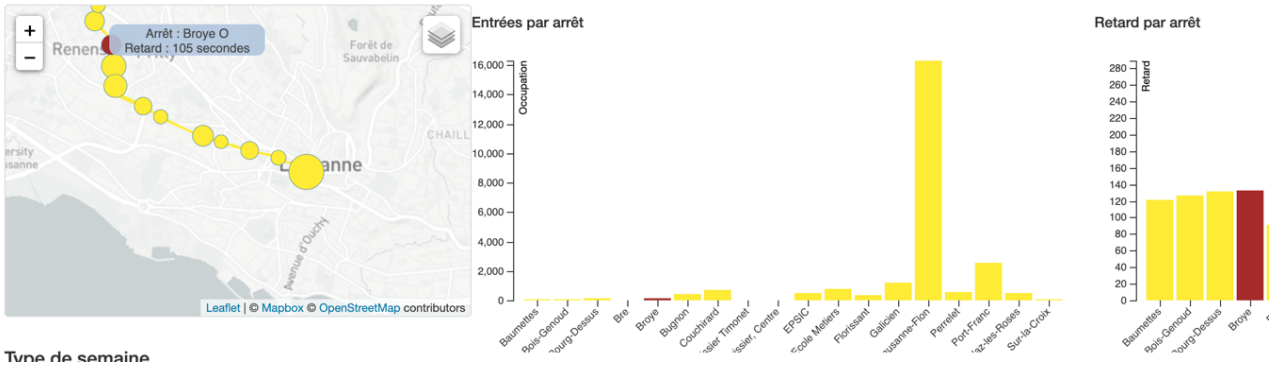


Figure 45 : Retard représenté par des symboles proportionnels sur chaque arrêt

Cette application permet de visualiser l'ensemble des données sans tri préalable. Certaines valeurs extrêmes peuvent être mises en avant et masquer la grande partie des données. C'est pour cela que les filtres permettent de définir plus précisément les événements à visualiser. La partie suivante illustre le fonctionnement de l'application grâce à des cas concrets.

5.3.2. Exemples d'utilisation

Cas général

Figure 46 : Page d'accueil de tlDataViewer

La figure 46 montre des données de la ligne 18, le matin de 7h à 12h pour les dates allant du 5 au 19 mai 2018. L'occupation de la ligne, les jours de semaine (bleu foncé au bleu clair) du 14 au 18 mai, est assez régulière. Par contre, le jeudi 10 montre une occupation de seulement 724 personnes, car il s'agit du jeudi de l'Ascension. Les charges sont d'ailleurs comparables aux deux dimanches illustrés via l'application. Il est par contre difficile de tirer une conclusion sur les différents retards pour cette période. Une période horaire de 7h à 12h montre l'évolution principale du matin, mais il est tout à fait possible de se concentrer sur les heures de pointes, comme de 7h à 9h ou de 17h à 19h pour évaluer la question du retard et de l'occupation à ces heures. La corrélation entre la charge et le retard est affichée dans les statistiques en bas à gauche.

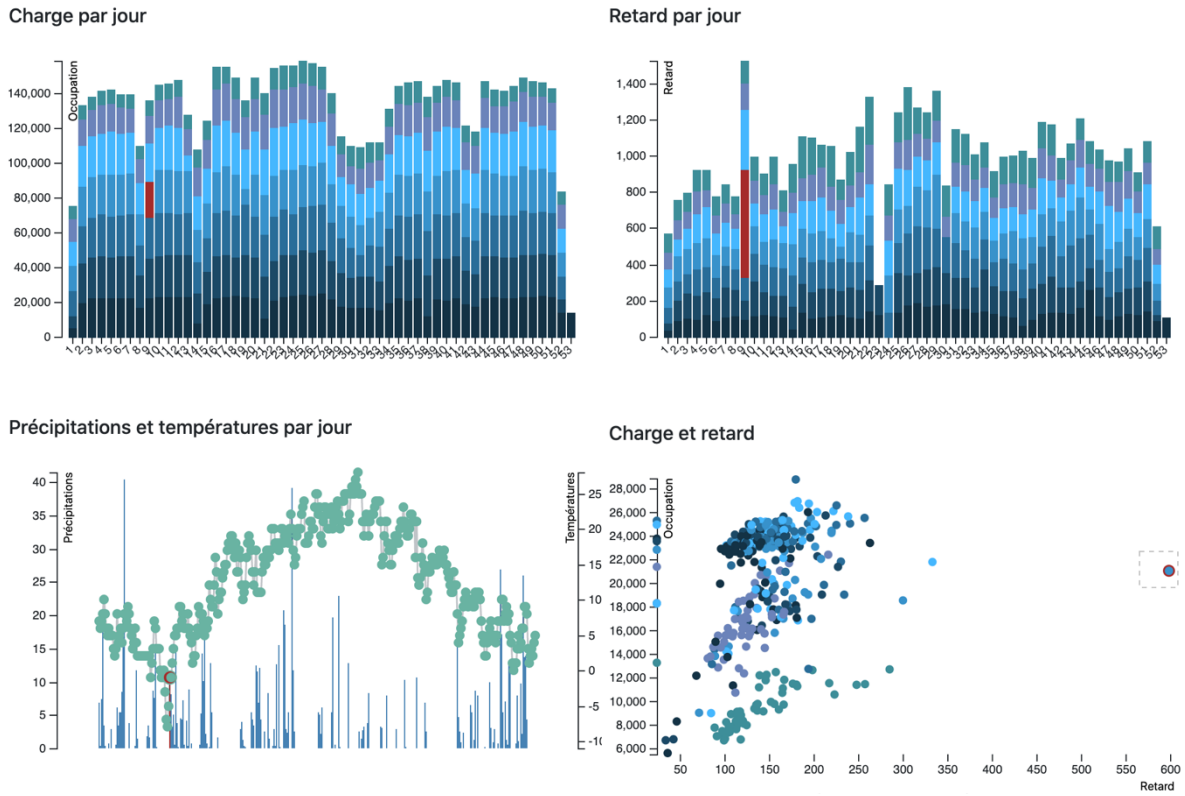


Figure 47 : Données complètes de l'année 2018 avec la sélection d'un outlier

La figure 47 illustre l'occupation et les retards des deux lignes pour toute l'année 2018, durant toutes les heures de la journée.

Plusieurs observations sont possibles sur ces différents graphiques. Tout d'abord, les valeurs manquantes du fichier TPs des horaires réalisés sont visibles de par le « trou » qu'elles laissent aux semaines 23 et 24 dans le graphique « Retard par jour ». Les points qui sont directement sur l'axe Y du graphique « Charge et retard » sont aussi le résultat des données manquantes. Il est tout de même possible de lire la charge sur cet axe pour les données non présentes en X.

Le graphique « Charge et retard » montre qu'il y a trois groupes de données : les dimanches (turquoise), les samedis (violet) et le reste de la semaine (nuances de bleu). En effet, même s'il est difficile d'alléguer un lien entre charge et retard, il est clairement visible que les dimanches sont moins occupés et dans une moindre mesure, les samedis aussi. Les quatre points qui sont en bas à gauche de ce graphique et qui sont des jours de la semaine sont tous des jours fériés. Il s'agit du 1er janvier et 2 janvier, du lundi de Pâques et de Noël.

Un point particulièrement extrême figure sur le graphique « Charge et retard ». Il s'agit du 1er mars. Le retard moyen est de près de 10 minutes pour les deux lignes lors de cette journée. En regardant la correspondance de ce point sur le graphique de la météo, il est possible de remarquer

qu'il s'agit d'un jour froid aux précipitations abondantes. Le journal 24 heures titrait en cette période « Le canton tourne au ralenti sous la neige » (Maendly, 2018). Le deuxième point comportant le plus de retard est d'ailleurs le 2 mars, jour de neige également.

Afin de visualiser les données de l'année 2018 d'une façon plus générale, il est judicieux de modifier les filtres et de supprimer cette valeur extrême en excluant un tel retard. La visualisation ci-dessous comporte les mêmes filtres que précédemment, mais en limitant les retards maxima par arrêt à 300 secondes, soit 5 minutes. Le retard maximum journalier est d'environ 120 secondes sur ce graphique une fois les données extrêmes retirées. La répartition des points est alors plus facile à analyser car ces derniers sont mieux répartis dans l'espace. Les trois groupes de points sont plus facilement identifiables lorsque les valeurs extrêmes n'apparaissent plus sur le graphique. La prochaine étape d'analyse est de séparer les trois groupes afin d'observer si diverses corrélations sont notables ou non.

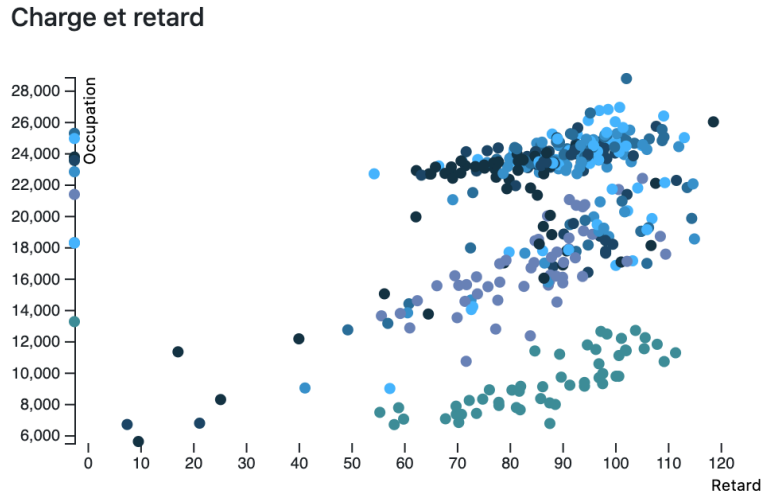


Figure 48 : Charge et retard, retards maxima de 5 minutes

Les figures 49 et 50 ci-dessous représentent respectivement les jours ouvrables et les jours ouvrables dont le retard par arrêt est de moins de 5 minutes. Les corrélations entre la charge et le retard deviennent significatives : la corrélation est de 0.52 pour le premier graphique et monte à 0.77 pour le second. Pour les jours ouvrables, il y a donc un lien entre la charge et les retards. Il est aussi à noter qu'une augmentation de l'occupation des bus est probablement liée avec une augmentation générale de la circulation routière. Pour des soucis de justesse, les données manquantes dans une des deux tables de charge ou de retard ne sont pas prises en compte pour le calcul de la corrélation dans l'application tlDataViewer.

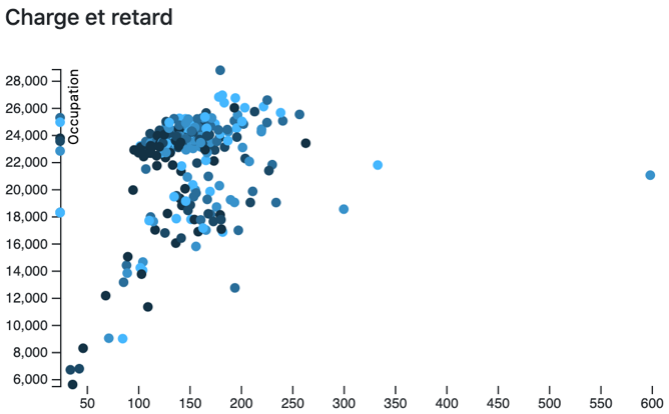


Figure 49 : Charge et retard, jours ouvrables

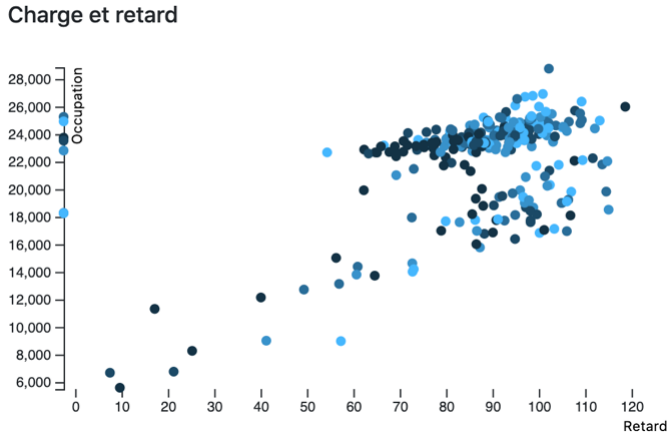


Figure 50 : Charge et retard, jours ouvrables dont le retard par arrêt est de moins de 5 minutes

Comme deux groupes de points apparaissent sur le graphique de la Figure 50, il est possible d'émettre l'hypothèse que les vacances scolaires sont le groupe de points du dessous. La fFigure 51 ci-dessous présente alors les points utilisant les filtres suivants : jours ouvrables hors des vacances scolaires dont le retard aux arrêts est de moins de 5 minutes. Suivant ces caractéristiques, la corrélation est de 0.91. Il est donc possible d'observer que lors de conditions normales pour des jours ouvrables, un lien est remarquable entre la charge et la ponctualité des bus. De plus, les trois *outliers* dans la partie basse du graphique sont le jeudi de l'Ascension, le lundi de Pentecôte et le lundi du Jeûne. L'application ne permet pas de prendre en compte les jours fériés, élément qui pourrait être ajouté pour les prochaines versions. Les corrélations des samedis et des dimanches pour moins de 5 minutes de retard sont elles aussi très élevées : elles figurent à 0.85 et 0.89. Pour ce genre d'analyse, des données externes aux tl comme le trafic automobile pourraient améliorer la compréhension du phénomène.

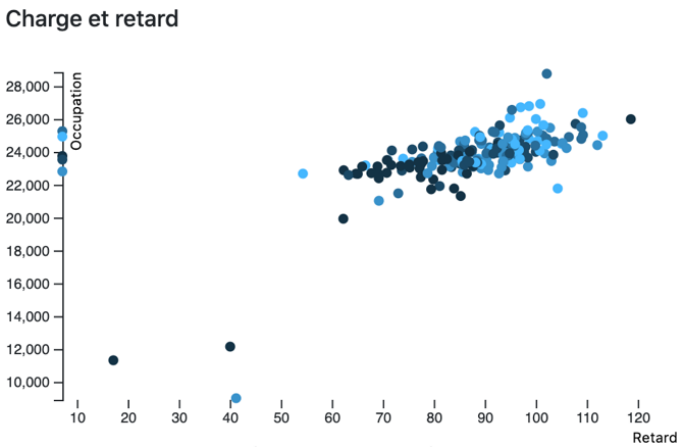


Figure 51 : Charge et retard, jours ouvrables hors des vacances scolaires dont le retard aux arrêts est de moins de 5 minutes

Météo

Un autre indicateur à prendre en compte est la météo journalière qu'il est possible d'ajouter dans les filtres. Des précipitations journalières de plus de 5 mm sont donc ajoutées aux données totales pour comprendre les répartitions des retards compte tenu de la météo via l'application tlDataViewer. La corrélation est exactement la même en incluant les précipitations que pour les données totales, soit de 0.46.

tlDataViewer « par arrêt »

La deuxième version de l'application qui a comme dénominateur les arrêts comme base plutôt que les jours, utilise les mêmes données que la version « par jour ». Par contre, il est possible de visualiser les retards directement sur la carte grâce à des symboles proportionnels. Il est ici intéressant de sélectionner une seule ligne de transport afin d'augmenter la visibilité. Le but de ces visualisations est de comprendre à quel(s) arrêt(s) les utilisateurs montent principalement. La figure 52 illustre un exemple sur la ligne 18. Pour ce tracé, la plupart des passagers montent au premier arrêt qui est le Flon. Il est intéressant de noter que l'arrêt du Flon est celui qui accumule le plus de retard, mais aussi de passagers. Cette visualisation montre que cette ligne est utilisée pour amener les passagers de la ville vers la banlieue principalement.

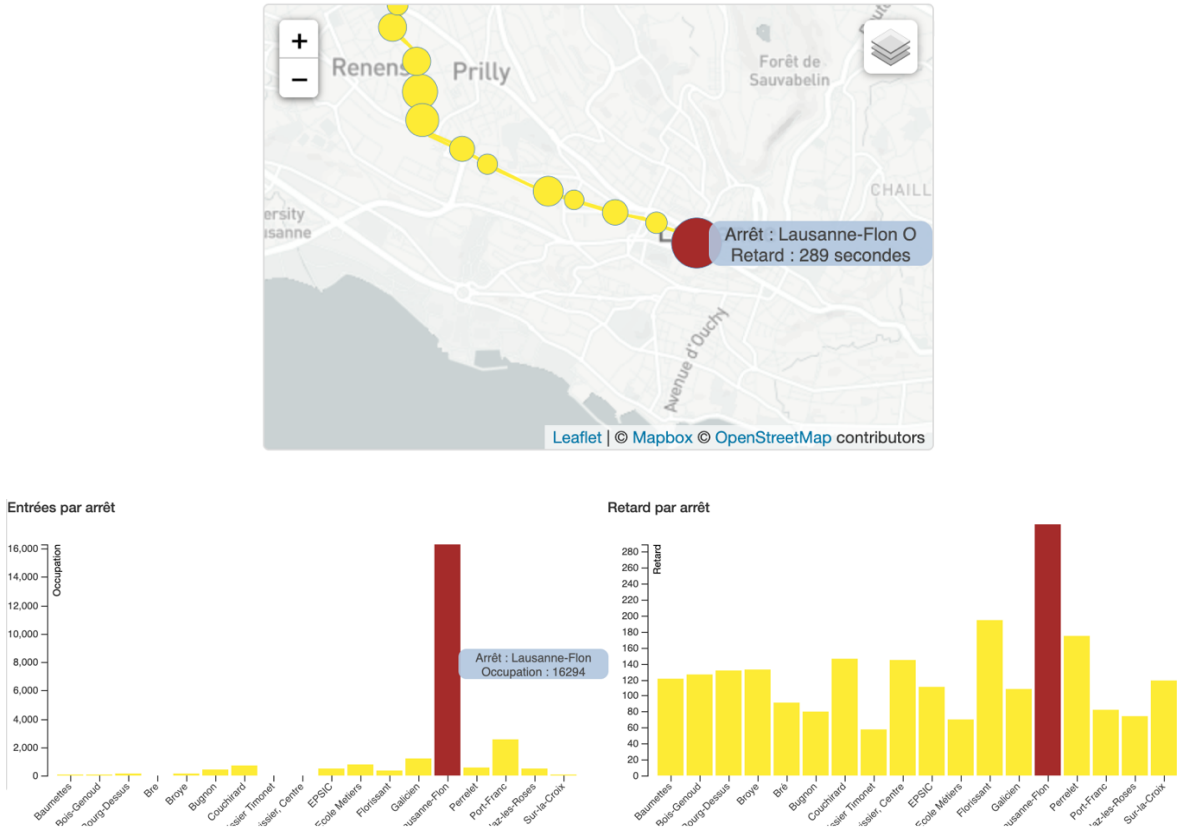


Figure 52 : Charge et retard pour l'arrêt du Flon, Ligne 18

6. Discussion

Les objectifs de base de ce travail étaient de trouver un moyen d'utiliser des données sans structure standardisée afin d'en tirer des statistiques, ainsi que de visualiser facilement les retards des bus et d'en identifier certains facteurs. La première idée de cette recherche était d'utiliser des données des tl couplées à des traces de déplacement issues du réseau téléphonique Swisscom. Pour des raisons de prix et de droits sur le réseau Swisscom, il a été retenu d'utiliser uniquement les données internes aux tl. L'analyse des retards a donc dû être réalisée sans données externes de déplacement, ce qui a légèrement changé la problématique en cours de route. Le processus de travail de ce mémoire a été exploratoire et évolutif : il était difficile d'évaluer directement ce qu'il était possible d'entreprendre avec les données disponibles et dans quelle mesure il était envisageable de répondre aux hypothèses dressées préalablement. Créer un outil d'aide à la décision était l'objectif idéal de ce travail, mais de nombreux imprévus et étapes de mise en forme pouvaient en tout temps modifier les objectifs initiaux. Il a finalement été possible de répondre aux principales questions dressées lors de la problématique grâce à des statistiques menées avec R et à l'application de visualisation.

L'appréciation des retards des bus, mais aussi de leur occupation, peut être estimée via cette nouvelle application tlDataViewer. Cet outil de visualisation aide à comprendre plus précisément la répartition des retards. tlDataViewer est aussi un outil interactif qui permet à chaque utilisateur de choisir quelles situations seraient pertinentes à étudier, que ce soit les heures de pointe de la semaine ou la période plus creuse du dimanche par exemple. En suivant les principes de l'exploration de données (EDA) et de la visualisation décrits dans le cadre théorique, l'utilisateur peut développer ses connaissances sur la ponctualité et l'occupation des bus des deux lignes choisies.

Les graphiques et statistiques produits par l'application permettent aux utilisateurs de naviguer dans les données en mettant en évidence visuellement certaines caractéristiques remarquables. En effet, il est possible d'affiner une visualisation en excluant par exemple un *outlier* via les filtres. Cela permet alors de recentrer l'échelle du graphique sur les données principales et d'en comprendre la tendance. Cet outil est fonctionnel dans un processus itératif pour des analyses qui pourraient être utiles aux tl.

A son niveau de développement actuel, l'application n'en est qu'au début de l'utilisation de son potentiel et fournit une première exploration sur les jeux de données des tl. Un test utilisateur

serait un moyen efficace pour comprendre dans quelle direction l'application pourrait être améliorée et quelles fonctionnalités seraient utiles pour une suite et réelle utilisation de tlDataViewer. En effet, la situation du début d'année 2020 et la crise sanitaire du COVID-19 n'a pas permis d'aller faire des tests utilisateurs, que ce soit sur le campus de l'Université de Lausanne ou dans les bureaux des tl. L'expérience utilisateur, *user experience (UX)* en anglais est une méthode d'évaluation où l'utilisateur est impliqué. Ce dernier est directement mis en interaction avec le produit et son évaluation est intéressante et observable ou mesurable (Albert, Tullis, & Safari, 2013). L'intérêt pour ce type de pratique est qu'elle regroupe la pratique empirique et la recherche dans le domaine des produits interactifs comme les pages web (Hassenzahl & Tractinsky, 2006). Cette étape serait essentielle pour une éventuelle continuation de ce travail.

De plus, l'application pourrait être mise en ligne directement sur un serveur pour qu'un large panel d'utilisateurs puisse la tester. Il n'était pas possible de faire tourner l'application directement sur un service d'hébergement web comme GitHub car la base de données est très volumineuse. Si l'application était mise en service, il serait intéressant d'optimiser encore plus les différentes tables et index afin d'accélérer le temps de calcul des diverses requêtes.

D'un point de vue des données, des pistes d'amélioration seraient aussi possibles. Une standardisation des divers jeux de données pourrait améliorer la facilité de traitement et la mise en forme d'une base de données. Des métadonnées précises ainsi qu'une uniformisation des diverses tables issues des tl simplifieraient les jointures entre les différents fichiers. Des identifiants uniques pour chaque ligne de données sont aussi d'une grande aide pour la vérification des résultats de traitement au sein d'une base de données, ce qui n'était pas le cas dans la plupart des fichiers. Un travail important de suppression des trajets non commerciaux a été effectué afin de ne garder que les trajets utiles dans la base de données. Il serait intéressant pour optimiser de futures analyses que les tl notifient une différence entre « trajets commerciaux » et « trajets hors service » dans les différentes tables de leurs données. L'ajout d'une indication des jours fériés pourrait être utile dans l'automatisation des processus et dans la visualisation. Tous les processus manuels de jointure de fichiers et plus particulièrement la comparaison des horaires théoriques et réalisés devraient être automatisés. Par exemple, rassembler sur une même table les horaires théoriques et réalisés pourrait améliorer l'exactitude des calculs et grandement les simplifier.

Un travail manuel de création des tronçons des lignes a été effectué afin de les afficher sur la carte interactive. L'étude du réseau ainsi que le tracé des diverses lignes aurait pu être plus développée. Il aurait été intéressant d'avoir une typologie précise du réseau comme par exemple un répertoire des tronçons en site propre, des feux de signalisation activables ou non, des pertes de

priorités ou encore des passages pour piétons afin d'intégrer ces paramètres dans l'analyse des retards.

Il existe évidemment un certain nombre de biais dans la conception de ce projet, que ce soit dans les données elles-mêmes, dans le choix des lignes de transports ou encore dans l'application tlDataviewer. Les deux lignes (lignes 2 et 18) étudiées ont été suggérées par les tl car ces tracés présentent des caractéristiques multiples et parfois compliquées. Pour une suite de ce travail, des données d'autres lignes de transport dites « faciles » pourraient être ajoutées pour pondérer les données. En effet, une base de données comprenant plus de lignes de bus permettrait de mieux évaluer la moyenne du réseau et de tirer des enseignements globaux sur le réseau en entier. De plus, cela permettrait de standardiser les observations et d'ainsi comparer les données avec un retard moyen 0 et un écart-type de 1 pour chaque donnée. Afin de simplifier des ajouts de données, l'application a déjà été codée de telle façon à ce qu'il n'y ait pas besoin de changer le code si de nouvelles lignes de transport étaient ajoutées. Seules de minimes modifications dans le ficher HTML seraient nécessaires pour ajouter le choix de la nouvelle ligne de transport. La puissance de calcul devrait alors être augmentée afin de traiter plus facilement la grande masse de données.

Pour revenir sur les objectifs de base qui ont été posés pour le travail et donc pour l'application, il est tout à fait possible de visualiser les données et d'en tirer certaines conclusions. Ce travail fournit un outil évolutif exploratoire qu'il serait possible d'améliorer à des fins d'analyse plus profonde, de prédiction des retards et de modélisation de ces derniers grâce aux données récoltées. Il est déjà possible de visualiser que la ligne 2 accumule plus de retard lorsque le nombre de passagers dépasse un certain seuil. Les retards sur la ligne 18 sont plus faibles, probablement parce que ce seuil d'occupation n'est pas atteint. D'autres données et techniques de prédictions plus poussées pourraient mettre en avant cette tendance même si après le nettoyage des données et la sélection de jours qui ont la même caractéristique, l'application montre que la charge et les retards sont corrélés sous certaines conditions.

Ces pistes permettent à l'application d'avoir un potentiel d'amélioration futur et de pouvoir évoluer d'un outil de *visualisation* à un outil de *prédiction*. Ce prolongement de l'étude pourrait être envisageable tant dans le milieu académique que dans le milieu professionnel auprès les tl.

Conclusion

Le traitement des données et leur visualisation font partie des méthodes émergentes dans le domaine de la géographie. L'interactivité et l'implication de l'utilisateur s'avèrent être des éléments cruciaux dans la géovisualisation de larges bases de données. En effet, cette approche laisse aux utilisateurs le choix de cibler directement les phénomènes qu'ils souhaitent étudier. Néanmoins, bien que la géovisualisation prenne de l'ampleur grâce à l'informatisation croissante de la société, beaucoup de données récoltées ne sont pas utilisées car elles nécessitent un trop grand travail d'adaptation pour être exploitées à des fins de compréhension d'un sujet.

L'objectif de ce travail de recherche était de mettre en place un outil de visualisation utile pour valoriser des données empiriques issues des tl. Il s'agissait de construire une application d'utilisation simplifiée pour mettre en avant les retards des bus suivant différents choix opérés via des filtres. La mise en forme des données et l'implémentation de l'application a permis de mettre en place un outil qui apporte, au moins en partie, une réponse à l'objectif de la recherche.

L'évolution continue des outils de la géoinformatisation permet de sans cesse dynamiser le traitement de données et la mise en valeur de ces dernières. Cette évolution est encore plus importante du fait que les supports numériques sont en constante amélioration eux aussi et permettent d'accueillir des visualisations très complexes via différentes plateformes. Comme l'informatique, cet outil est perfectible à l'infini, mais il faut toujours garder en tête les objectifs de base afin de ne pas complexifier plus que nécessaire son fonctionnement et garder au premier plan les besoins des utilisateurs.

Les résultats des observations qui ont été menées grâce à l'application ont montré qu'il était possible d'isoler des phénomènes en supprimant les valeurs extrêmes et en analysant uniquement les groupes de données qui avaient une même caractéristique visuelle. Cela a par exemple permis de mettre en avant la corrélation qu'il y a entre la charge et le retard des bus dans certaines situations. Cette observation n'était pas évidente lors de l'étape de l'exploration de données alors que l'analyse et la compréhension de ces mêmes données grâce à tlDataViewer se sont révélées plus faciles. La manipulation de l'application a été conçue pour être agréable à utiliser.

Pour conclure, ce travail a consisté en trois grandes étapes importantes : la mise en forme de données non structurées, l'analyse exploratoire de ces données et finalement la conception d'une application interactive de géovisualisation. Dans une vision prospective d'une continuation de

ce projet, l'objectif serait de développer l'application de manière à ce qu'elle devienne un outil performant pour permettre aux responsables de la mobilité d'analyser, de prédire et d'améliorer la planification du réseau et des horaires des tl.

Bibliographie

Abenoza, R. F., Cats, O., & Susilo, Y. O. (2017). Travel satisfaction with public transport : Determinants, user classes, regional disparities and their evolution. *Transportation Research Part A: Policy and Practice*, *95*, 64-84. https://doi.org/10.1016/j.tra.2016.11.011

Aggarwal, C. C. (2015). Outlier Analysis. Dans C. C. Aggarwal, *Data Mining* (pp. 237-263). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8_8

Albert, W., Tullis, T., & Safari, an O. M. C. (2013). *Measuring the User Experience, 2nd Edition*. (S.l.): (s.n.). Repéré à https://www.safaribooksonline.com/complete/auth0oauth2/&state=/library/view//9780124157811/?ar

Aldenderfer, M. S., & Maschner, H. D. G. (Éds). (1996). *Anthropology, space, and geographic information systems*. New York: Oxford University Press.

Anderson, C. (2008). The end of theory : The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7). Repéré à http://www.uvm.edu/pdodds/files/papers/others/2008/anderson2008a.pdf

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966

Anselin, L. (1989). What is Special About Spatial Data ? Alternative Perspectives on Spatial Data Analysis (89-4). *UC Santa Barbara : National Center for Geographic Information and Analysis*. Repéré à https://escholarship.org/uc/item/3ph5k0d4

Anselin, L. (1996). Interactive Techniques and Exploratory Spatial Data Analysis. *Regional Research Institute Publications and Working Papers*, *200*, 253-266.

Anselin, L., & Bao, S. (1997). Exploratory Spatial Data Analysis Linking SpaceStat and ArcView. Dans M. M. Fischer & A. Getis (Éds), *Recent Developments in Spatial Analysis* (pp. 35-59). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-03499-6_3

ArcGIS. (2020). COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). *COVID-19 Dashboard*. Repéré à https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Bavaud, F. (1998). *Modèles et données : Une introduction à la statistique uni-, bi- et trivariée*. Paris: L'Harmattan.

Bertin, J. (1983). *Semiology of graphics*. Madison, Wis: University of Wisconsin Press.

Bhattacharya, S., Phithakkitnukoon, S., Nurmi, P., Klami, A., Veloso, M., & Bento, C. (2013). Gaussian process-based predictive modeling for bus ridership. Dans *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication—UbiComp '13 Adjunct* (pp. 1189-1198). Zurich, Switzerland: ACM Press. https://doi.org/10.1145/2494091.2497349

Bielli, M., Caramia, M., & Carotenuto, P. (2002). Genetic algorithms in bus network optimization. *Transportation Research Part C: Emerging Technologies*, *10*(1), 19-34. https://doi.org/10.1016/S0968-090X(00)00048-6

Bivand, R. S. (2010). Exploratory Spatial Data Analysis. Dans M. M. Fischer & A. Getis (Éds), *Handbook of Applied Spatial Analysis* (pp. 219-254). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03647-7_13

Bougheas, S., Demetriades, P. O., & Morgenroth, E. L. W. (1999). Infrastructure, transport costs and trade. *Journal of International Economics*, *47*(1), 169-189. https://doi.org/10.1016/S0022-1996(98)00008-7

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis : Forecasting and control* (Fifth edition). Hoboken, New Jersey: John Wiley & Sons, Inc.

Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA : Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Brunet, R. (Éd.). (2006). *Les mots de la géographie : Dictionnaire critique* (3. éd. rev. et augmentée). Paris: La Documentation Française.

Chambers, J. M. (2008). *Software for data analysis : Programming with R*. New York: Springer.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection : A survey. *ACM Computing Surveys*, *41*(3), 1-58. https://doi.org/10.1145/1541880.1541882

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed). New York: Wiley.

Edgeworth, F. Y. (1887). XLI. *On discordant observations. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *23*(143), 364-375. https://doi.org/10.1080/14786448708628471

Etat de Vaud. (2020). Etat de Vaud, mobilité. *Etat de Vaud*. Repéré à https://www.vd.ch/themes/mobilite/

Few, S. (2006). *Information dashboard design : The effective visual communication of data* (1st ed). Beijing ; Cambride [MA]: O'Reilly.

Flannery, J. J. (1956). *The graduated circle : A description, analysis, and evaluation of a quantitative map symbol* (University of Wisconsin--Madison). USA: (s.n.).

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography : Perspectives on spatial data analysis*. London ; Thousand Oaks, Calif: Sage Publications.

Gastner, M. T., & Newman, M. E. J. (2006). Optimal design of spatial distribution networks. *Physical Review E*, *74*(1), 016117. https://doi.org/10.1103/PhysRevE.74.016117

Goodchild, M. F. (1992). Geographical data modeling. *Computers & Geosciences*, *18*(4), 401-408. https://doi.org/10.1016/0098-3004(92)90069-4

Goodchild, M. F., & Haining, R. P. (2004). GIS and spatial data analysis : Converging perspectives. Dans R. J. G. M. Florax & D. A. Plane (Éds), *Fifty Years of Regional Science* (pp. 363-385). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-07223-3_16

Hassenzahl, M., & Tractinsky, N. (2006). User experience—A research agenda. *Behaviour & Information Technology*, *25*(2), 91-97. https://doi.org/10.1080/01449290500330331

Heilporn, G., De Giovanni, L., & Labbé, M. (2008). Optimization models for the single delay management problem in public transportation. *European Journal of Operational Research*, *189*(3), 762-774. https://doi.org/10.1016/j.ejor.2006.10.065

Hey, A. J. G. (Éd.). (2009). *The fourth paradigm : Data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

Jemelin, C. (2008). *Transports publics dans les villes : Leur retour en force en Suisse*. Lausanne: Presses polytechniques et universitaires romandes.

Johnston, R. (1997). W(H)ITHER spatial science and spatial analysis. *Futures*, *29*(4-5), 323-336. https://doi.org/10.1016/S0016-3287(97)00014-1

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1-8. https://doi.org/10.1109/2945.981847

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science : A New Paradigm for Biodiversity Studies. *BioScience*, *59*(7), 613-620. https://doi.org/10.1525/bio.2009.59.7.12

Khan Academy. (2018). What is a library? *Khan Academy*. Repéré à https://www.khanacademy.org/computing/computer-programming/html-css-js/using-js-libraries-in-your-webpage/a/whats-a-js-library

Kitchin, R. (2013). Big data and human geography : Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262-267. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Kraak, M. J., & Ormeling, F. (2010). *Cartography : Visualization of geospatial data* (3rd ed). Harlow ; New York: Prentice Hall.

Kuhn, T. S., & Hacking, I. (2012). *The structure of scientific revolutions* (Fourth edition). Chicago ; London: The University of Chicago Press.

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2014). Traffic Flow Prediction With Big Data : A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1-9. https://doi.org/10.1109/TITS.2014.2345663

Maendly, V. (2018, 1 mars). Le canton toure au ralenti sous la neige. *24 Heures*. Repéré à https://www.24heures.ch/vaud-regions/photos-canton-tourne-ralenti-neige/story/29153699

Magnanti, T. L., & Wong, R. T. (1984). Network Design and Transportation Planning : Models and Algorithms. *Transportation Science*, *18*(1), 1-55. https://doi.org/10.1287/trsc.18.1.1

Mandl, C. E. (1980). Evaluation and optimization of urban public transportation networks. *European Journal of Operational Research*, *5*(6), 396-404. https://doi.org/10.1016/0377-2217(80)90126-5

Meeks, E. (2018). *D3.js in action : Data visualization with JavaScript* (Second edition). Shelter Island, NY: Manning.

Mérenne, E. (2013). *Géographie des transports contraintes et enjeux*. Rennes: Presses universitaires de Rennes.

Miller, H. J. (2010). The Data Avalanche Is Here. Shouldn't We Be Digging? *Journal of Regional Science*, *50*(1), 181-201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

National Science Board. (2005). Long-Lived Digital Data Collections : Enabling Research and Education in the 21st Century. *National Science Foundation*, *Technical Report NSB-05-40*. Repéré à www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Niederst Robbins, J. (2012). *Learning Web design : A beginner's guide to HTML, CSS, JavaScript, and web graphics* (Fourth edition). Beijing: O'Reilly.

Office fédéral des transports. (2019). *Transports publics – pour la Suisse Stratégie de l'OFT 2019*. Ittigen, Suisse: Département fédéral de l'environnement, des transports, de l'énergie et de la communication DETEC. Repéré à https://www.bav.admin.ch/dam/bav/fr/dokumente/das-bav/eine_strategie_fuerdiezukunftdesoeffentlichenverkehrs.pdf.download.pdf/une_strategie_pourlavenirdestransportspublics.pdf

Organisation Météorologique Mondiale. (2017). *Directives de l'OMM pour le calcul des normales climatiques* (Rapport No. 1203). Genève: Organisation Météorologique Mondiale. Repéré à https://library.wmo.int/doc_num.php?explnum_id=4220

Park, R. E., Burgess, E., W., & McKenzie, R. D. (1984). *The city : Suggestions for investigation of human behavior in the urban environment*. (S.l.): (s.n.). Repéré à http://public.eblib.com/choice/publicfullrecord.aspx?p=3563067

Pointet, A. (2007). Rencontre de la science de l'information géographique et de l'anthropologie culturelle:modélisation spatiale et représentation de phénomènes culturels. https://doi.org/10.5075/EPFL-THESIS-3789

Ramakrishnan, R., & Gehrke, J. (2000). *Database management systems* (2nd ed). Boston: McGraw-Hill.

Santiago, E. (2015, 27 avril). What Should I Do If My Data Is Not Normal? *The Minitab Blog*. Repéré à https://blog.minitab.com/blog/understanding-statistics-and-its-application/what-should-i-do-if-my-data-is-not-normal-v2

Spence, R. (2014). *Information visualization : An introduction* (Third edition). Cham Heidelberg New York Dordrecht London: Springer.

Tobler, W. R. (1979). Cellular Geography. Dans S. Gale & G. Olsson (Éds), *Philosophy in Geography* (pp. 379-386). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9394-5_18

Transportation Research Board of the National Academies. (1999). *TCRP Report 47*. Washington, D.C.

Transports Lausannois. (2020). Réseau et flotte. *A propos des tl*. Repéré à https://www.t-l.ch/a-propos-des-tl/entreprise/reseau-et-flotte

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub. Co.

Tukey, J. W. (1992). The Future of Data Analysis. Dans S. Kotz & N. L. Johnson (Éds), *Breakthroughs in Statistics* (pp. 408-452). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_31

Vuillemot, R. (2010). *Un cadre de conception pour la Visualisation d'Information Interactive*. Ecole doctorale Infomaths sciences et technologies de l'information et de la communication, France.

Washington, S., Karlaftis, M. G., & Mannering, F. L. (2011). *Statistical and econometric methods for transportation data analysis* (2nd ed). Boca Raton, FL: CRC Press.

Xu, H., & Zheng, M. (2012). Impact of Bus-Only Lane Location on the Development and Performance of the Logic Rule-Based Bus Rapid Transit Signal Priority. *Journal of Transportation Engineering*, *138*(3), 293-314. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000325

Yue, S., & Pilon, P. (2004). A comparison of the power of the t test, Mann-Kendall and bootstrap tests for trend detection / Une comparaison de la puissance des tests t de Student, de Mann-Kendall et du bootstrap pour la détection de tendance. *Hydrological Sciences Journal*, *49*(1), 21-37. https://doi.org/10.1623/hysj.49.1.21.53996

Zhong, C., Huang, X., Müller Arisona, S., Schmitt, G., & Batty, M. (2014). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems*, *48*, 124-137. https://doi.org/10.1016/j.compenvurbsys.2014.07.004

Annexes

Annexe 1 : plan géographique du réseau des tl

Annexe 2 : tracé détaillé de la ligne 2

Annexe 3 : tracé détaillé de la ligne 18

Annexe 4 : exemple des tables de données TPs, TJM

Annexe 1 : plan géographique du réseau des tl

Annexe 2 : arrêts, positions et codes de la ligne 2

Aller – direction Désert			Retour – direction Maladière		
Position	Arrêt code	Arrêt	Position	Arrêt code	Arrêt
0	MALL_E	Maladière-Lac	0	DST_E	Désert
1	TVIDY_E	Théâtre de Vidy	1	P-RIV_E	Parc Rivier
2	BVE_E	Bellerive	2	PIERF_E	Pierrefleur
3	PECH_E	Pecheurs	3	GREY_E	Grey
4	OUCHY_E	Ouchy-Olympique	4	BOISY_E	Boisy
5	B-RIV_N	Beau-Rivage	5	BERG_E	Bergières
6	JDILS_N	Jordils	6	PRBYT_E	Presbytère
7	CX-OY_N	Croix d'Ouchy	7	BLIEU_E	Beaulieu
8	MTFLI_N	Mont-Fleuri	8	BLIJO_E	Beaulieu-Jomini
9	CLOSL_N	Closelet	9	VINET_E	Vinet
10	MIRAB_N	Mirabeau	10	VAL_S	Valentin
11	GTE_N	Georgette	11	RNEUV_S	Rue Neuve
12	SF_O	St-François	12	B-AIR_D	Bel-Air
13	B-AIR_N	Bel-Air	13	SF_S	St-François
14	RNEUV_N	Rue Neuve	14	GTE_D	Georgette
15	VAL_N	Valentin	15	CLOSL_D	Closelet
16	VINET_O	Vinet	16	MTFLI_S	Mont-Fleuri
17	BLIJO_O	Beaulieu-Jomini	17	CX-OY_S	Croix d'Ouchy
18	BLIEU_O	Beaulieu	18	JDILS_S	Jordils
19	PRBYT_O	Presbytère	19	B-RIV_S	Beau-Rivage
20	BERG_O	Bergières	20	OUCHY_O	Ouchy-Olympique
21	GREY_O	Grey	21	PECH_O	Pecheurs
22	PIERF_O	Pierrefleur	22	BVE_O	Bellerive
23	P-RIV_O	Parc Rivier	23	TVIDY_O	Théâtre de Vidy
24	DST_O	Désert	24	MALL_O	Maladière-Lac
25	DST_E	Désert	25	MALL_E	Maladière-Lac

Annexe 3 : arrêts, positions et codes de la ligne 18

Aller – direction Crissier Timonet			Retour – direction Lausanne-Flon		
Position	Arrêt code	Position	Arrêt code	Position	Arrêt code
0	FLON_O	Lausanne-Flon	0	TIMON_T	Crissier Timonet
1	PFRAN_O	Port-Franc	0	TIMON_E	Crissier Timonet
2	EPSIC_O	EPSIC	1	CRISS_S	Crissier, Centre
3	ECMET_O	Ecole Métiers	2	BRE_S	Bré
4	CRARD_O	Couchirard	3	GENOU_E	Bois-Genoud
5	PZROS_O	Prélaz-les-Roses	4	BAUM_S	Baumettes
6	GALIC_O	Galicien	5	SURLC_E	Sur-la-Croix
7	PERRL_N	Perrelet	6	BGNON_E	Bugnon
8	FLORI_O	Florissant	7	BDESS_D	Bourg-Dessus
9	BROYE_O	Broye	8	BROYE_E	Broye
10	BDESS_M	Bourg-Dessus	9	FLORI_E	Florissant
11	BGNON_O	Bugnon	10	PERRL_E	Perrelet
12	SURLC_O	Sur-la-Croix	11	GALIC_E	Galicien
13	BAUM_N	Baumettes	12	PZROS_E	Prélaz-les-Roses
14	GENOU_O	Bois-Genoud	13	CRARD_E	Couchirard
15	BRE_N	Bré	14	ECMET_E	Ecole Metiérs
16	CRISS_N	Crissier, Centre	15	EPSIC_E	EPSIC
17	TIMON_T	Crissier Timonet	16	PFRAN_E	Port-Franc
			17	FLON_E	Lausanne-Flon
			18	FLON_O	Lausanne-Flon

Annexe 4 : exemple des tables de données TPs, TJM

TPs

tps_stop_code	tps_stop_name	tps_first_start	distance	tps_date	tps_line	tps_position	section	tps_direction	trav_id	tps_arrival	tps_departure	cal_day_type	day_name	cal_regular	week_no	car
BAUM_S	Baumettes	06:58	593	2018-05-30	18	5	Bois-Genoud -> Baumettes	R	83887-11-07 00:00:00	07:04:58	07:05:36	se	me	1	22	18_3
BAUM_S	Baumettes	06:58	593	2018-05-29	18	5	Bois-Genoud -> Baumettes	R	83865-11-29 00:00:00	07:04:30	07:04:59	se	ma	1	22	18_3
BAUM_S	Baumettes	06:58	593	2018-06-22	18	5	Bois-Genoud -> Baumettes	R	84390-02-12 00:00:00	07:04:19	07:04:49	se	ve	1	25	18_3
BAUM_S	Baumettes	06:58	593	2018-06-20	18	5	Bois-Genoud -> Baumettes	R	84345-06-20 00:00:00	07:06:02	07:06:38	se	me	1	25	18_3
BAUM_S	Baumettes	06:58	593	2018-04-26	18	5	Bois-Genoud -> Baumettes	R	83193-10-24 00:00:00	07:03:50	07:04:23	se	je	1	17	18_3
BAUM_S	Baumettes	06:58	593	2018-04-24	18	5	Bois-Genoud -> Baumettes	R	83143-08-19 00:00:00	07:04:18	07:04:52	se	ma	1	17	18_3
BAUM_S	Baumettes	06:58	593	2018-04-23	18	5	Bois-Genoud -> Baumettes	R	83120-05-24 00:00:00	07:04:10	07:04:42	se	lu	1	17	18_3
BAUM_S	Baumettes	06:58	593	2018-04-25	18	5	Bois-Genoud -> Baumettes	R	83167-01-16 00:00:00	07:04:37	07:05:07	se	me	1	17	18_3
BAUM_S	Baumettes	06:58	593	2018-05-16	18	5	Bois-Genoud -> Baumettes	R	83599-02-21 00:00:00	07:04:20	07:04:50	se	me	1	20	18_3
BAUM_S	Baumettes	06:58	593	2018-05-17	18	5	Bois-Genoud -> Baumettes	R	83622-11-07 00:00:00	07:03:13	07:03:47	se	je	1	20	18_3

theo_timetable	theo_position	theo_trav_id2	last_stop	tl_position	geom	theo_first_start	tps_departure24	tps_arrival24	theo_timetable_s	tps_departure_s	delay	weather_date	p	t	r	nb_data
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:05:36	07:04:58	25440	25536	96	2018-05-30	0	17.762500087420147	869	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:59	07:04:30	25440	25499	59	2018-05-29	6	17.995833198229473	893	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:49	07:04:19	25440	25489	49	2018-06-22	0	17.787500103314716	947	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:06:38	07:06:02	25440	25598	158	2018-06-20	0	22.891666571299236	899	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:23	07:03:50	25440	25463	23	2018-04-26	0	14.3125	898	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:52	07:04:18	25440	25492	52	2018-04-24	0	16.041666666666668	848	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:42	07:04:10	25440	25482	42	2018-04-23	0	16.68749992052714	690	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:05:07	07:04:37	25440	25507	67	2018-04-25	0	18.204166491826374	794	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:04:50	07:04:20	25440	25490	50	2018-05-16	4.4	11.408333341280619	351	24
07:04	6	1348	Lausanne-Flon	4	0101000020E6100000C8847485325D1A407F6F0D4B87454740	06:58	07:03:47	07:03:13	25440	25427	-13	17.05.18	0	14.099999944368998	935	24

TJM

tjm_date	tjm_line	tjm_first_start	count_trav_id	tjm_direction	tjm_position	tjm_stop_code	tjm_stop_name	tjm_departure	go_down	go_up
2018-08-01	18	00:05:00	2,0181E+12	R	4	BAUM_S	Baumettes	00:12:00	0	0
2018-08-01	18	00:20:00	2,0181E+12	R	4	BAUM_S	Baumettes	00:26:04	0	0
2018-08-01	18	00:35:00	2,0181E+12	R	4	BAUM_S	Baumettes	00:39:18	0	0
2018-08-01	18	00:50:00	2,0181E+12	R	4	BAUM_S	Baumettes	00:54:01	0	0
2018-08-01	18	01:05:00	2,0181E+12	R	4	BAUM_S	Baumettes	01:10:57	0	0
2018-08-01	18	01:20:00	2,0181E+12	R	4	BAUM_S	Baumettes	01:23:51	1	0
2018-08-01	18	01:35:00	2,0181E+12	R	4	BAUM_S	Baumettes	01:38:56	0	0
2018-08-01	18	05:35:00	2,0181E+12	R	4	BAUM_S	Baumettes	05:40:23	0	0
2018-08-01	18	05:50:00	2,0181E+12	R	4	BAUM_S	Baumettes	05:58:18	0	0
2018-08-01	18	06:05:00	2,0181E+12	R	4	BAUM_S	Baumettes	06:09:40	0	1

payload	cal_date	cal_day_type	cal_regular	day_name	week_no	weather_date	p	t	r	nb_data
1	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
0	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
0	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
1	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
0	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
0	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
1	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
2	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
0	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24
2	2018-08-01	se	0	me	31	2018-08-01	0	25.09999982515971	840	24