Auteur du rapport :

Escudié Tom

Tuteur IUT :

Cabanac Guillaume

Dates du stage :

11/04/2022 au 22/06/2022

RAPPORT DE STAGE

TRANSITION-VELO ELABORATION D'UN WEBSIG

MAÎTRE DE STAGE

Jouve Bertrand

LISST - UMR 5193
CNRS, UT2J, EHESS, ENSFEA
Université de Toulouse Jean Jaurès
Maison de la Recherche
5, Allée Antonio Machado
31058 Toulouse Cedex 9

Auteur du rapport :

Escudié Tom

Tuteur IUT :

Cabanac Guillaume

Dates du stage :

11/04/2022 au 22/06/2022

RAPPORT DE STAGE

TRANSITION-VELO ELABORATION D'UN WEBSIG

MAÎTRE DE STAGE

Jouve Bertrand

Remerciements

Au terme de ce stage, je tenais particulièrement à remercier notre tuteur de stage M. Bertrand Jouve de nous avoir accueilli, de nous avoir accompagné et d'avoir fait en sorte que ce stage se déroule dans les meilleures conditions possibles.

Je tenais également à remercier toute l'équipe travaillant sur le projet TRANSITION-VÉLO de nous avoir grandement aidé lors des réunions notamment grâce à leurs idées et leurs différents retours.

Je remercie également le chercheur M. Laurent Jégou de nous avoir prêté son serveur personnel pour nous avoir permis de mettre en place une base de données lors de la phase de tests.

Et enfin, je remercie notre tuteur IUT M. Guillaume Cabanac, qui nous a fourni d'excellents conseils pour mener à bien notre stage et qui nous serviront également pour la suite de notre future carrière professionnelle ainsi que mon camarade de promotion Mattéo Camin, avec qui j'ai réalisé ce stage, et qui, grâce à son sérieux a fait en sorte que nous avons pu remplir avec succès les missions qui nous ont été confiées.

Table des matières

Figures et des tableaux

Figures

Tableaux

Extraits de code

Annexes

Introduction

Le but du projet pluridisciplinaire TRANSITION-VELO, coordonné par Bertrand Jouve (directeur de recherche au CNRS, mathématicien et également notre maître de stage) est d'étudier et de caractériser l'impact de la pandémie de COVID-19 sur l'utilisation des VLS (Vélo en Libre-Service) en comparant les situations de Toulouse et de Lyon. L'objectif est d'aider les opérateurs et pouvoirs publics à se saisir rapidement de ces transformations d'usage pour à la fois répondre au mieux à une situation d'urgence créée par la pandémie et se préparer aux crises futures, comprendre les opportunités actuelles en termes d'usage de vélos et mettre en place les conditions d'une pérennisation de nouvelles pratiques cyclables.

C'est pour cela que le LISST (Laboratoire Interdisciplinaire Solidarités, Sociétés, Territoires) nous a embauché en tant que stagiaire pour 10 semaines, moi et un camarade de promotion de l'IUT Informatique UT3 Mattéo Camin afin de mettre en place un webSIG (Système d'Information Géographique) permettant l'exploration et l'exploitation des différentes données de transport urbain des villes de Lyon.

Mon objectif premier lors de ce stage est d'arriver, au terme des 10 semaines à fournir aux chercheurs du LISST une application fonctionnelle, permettant d'une part l'exploitation et la compréhension des données, et d'autre part une certaine adaptabilité pour les éventuelles personnes qui amélioreraient l'application dans le futur.

Dans ce rapport nous verrons donc le cheminement vers cet objectif, en commençant par une présentation du lieu d'accueil et des différents collaborateurs, puis de l'analyse du problème jusqu'à aux méthodes de résolution, en passant par les difficultés rencontrées, afin de terminer par une présentation des résultats obtenus ainsi que du bilan de ce que j'ai tiré personnellement de cette expérience.

I) Le laboratoire LISST

1) Présentation du LISST

a) Le LISST

Le LISST (Laboratoire Interdisciplinaire Solidarités, Sociétés, Territoires) est une Unité Mixte de Recherche en Sciences Humaines et Sociales à large couverture thématique qui relève des sections 36, 38 et 39 du CNRS. Il est situé à la Maison de la Recherche sur le campus du Mirail de l'Université Toulouse Jean-Jaurès. Le LISST a pour tutelles l'Université Toulouse Jean Jaurès, le CNRS, l'EHESS et l'ENSFEA.

Les recherches sont portées par 4 équipes :

- LISST-CAS (Centre d'Anthropologie Sociale)
- LISST-CERS (Collectif : Expériences Réseaux et Sociétés)
- LISST-CIEU (Centre Interdisciplinaire d'Études Urbaines)
- LISST-Dynamiques Rurales

Le travail réalisé par chaque équipe est complété au sein de 5 axes transversaux : Environnement et Sociétés, Mondialisations, Différenciations territoriales et action collective, Innovations et société, Parcours de vie et inégalités dont les terrains d'enquêtes se situent dans l'espace européen, les Amériques, l'Afrique et l'Asie.

b) Les différents collaborateurs

Chercheurs titulaires

- Philippe Dugot (géographie aménagement) – Professeur – Université Toulouse Jean Jaurès – LISST – Toulouse
- Nour-Eddin El Faouzi (mathématicien) – Professeur – Université Gustave Eiffel et Ecole Nationale des Travaux Publics de l'Etat – LICIT – Lyon
- Fabrice Escaffre (géographie aménagement) – Maître de Conférences – Université Toulouse Jean Jaurès – LISST – Toulouse
- Angelo Furno (mathématicien) – Chercheur – Université Gustave Eiffel et Ecole Nationale des Travaux Publics de l'Etat – LICIT – Lyon
- Synda Haouès-Jouve (géographie aménagement) – Maîtresse de Conférences – Université Toulouse Jean Jaurès – LISST – Toulouse
- Marc Ivaldi (économiste) – Directeur d'Etude – Ecole des Hautes Etudes en Sciences Sociales et Ecole d'Economie de Toulouse – Toulouse
- Bertrand Jouve (mathématicien, coordonnateur) – Directeur de Recherche CNRS – LISST – Toulouse
- Bruno Revelli (géographie aménagement) – Maître de Conférences – Université Toulouse Jean Jaurès – LISST – Toulouse
- Paul Rochet (mathématicien) – Enseignant chercheur – Ecole Nationale de l'Aviation Civile – Toulouse
- Najla Touati (géomaticienne) – Ingénieure – Université Toulouse Jean Jaurès – LISST – Toulouse

2) Conditions de travail

a) Encadrement

Comme cité plus haut, le LISST est avant tout un laboratoire de sciences humaines, ce qui implique plusieurs choses, d'une part, le LISST ne dispose pas d'un service informatique à proprement parler, même si les personnes travaillant sur le projet TRANSITION-VELO sont chacun plus ou moins familier avec l'outil informatique, ils n'ont pas forcément le temps ou les compétences pour nous éclairer sur des problèmes techniques, ce qui a demandé une bonne autonomie de notre part pour résoudre les différents problèmes auxquels nous étions confrontés. D'autre part, les chercheurs peuvent mal évaluer la difficulté ou la faisabilité d'une fonctionnalité à implanter dans l'application lors de leurs demandes.

Notre tuteur de stage M. Bertrand Jouve se trouve être le coordonnateur du projet transition-vélo. Donc au début du stage, il a pu nous briefer et était disponible pour répondre à toutes nos questions afin de bien nous lancer.

b) Environnement de travail

La première semaine, nous avons été affectés dans une salle avec 3 autres stagiaires, elle disposait d'ordinateurs de travail mais nous avons préféré travailler sur nos machines personnelles, plus performantes. Cependant, la semaine d'après, des stagiaires qui étaient en déplacement sont revenus dans leur bureau, et dû au manque de place nous avons été affectés dans une salle de visioconférence, cette fois-ci sans ordinateur, et avec un seul câble Ethernet pour deux ordinateurs.

Nous disposions cependant d'une connexion très haut débit (≈600 Mo/s descendant), un plus étant donné le volume de données à traiter. Il nous a été également fourni de quoi écrire et prendre des notes, et plus tard un moniteur externe 27" nous a été fourni pour nous permettre de mieux visualiser les cartes.

c) Horaires et accès

L'accès à la maison de la recherche est libre la journée, cependant, un badge qu'on nous a remis le premier jour de travail est nécessaire pour pouvoir accéder à notre salle. Les horaires étaient flexibles, cependant le règlement intérieur préconisait des horaires de référence (08h00 - 18h30), ainsi qu'une pause méridienne d'au moins 45 min et d'une durée maximale de 2 heures. Seul le travail effectif est pris en compte, c'est-à-dire que le temps de travail hebdomadaire s'élève à 35 heures pauses déjeuner exclues. En tant que stagiaires, nous devions obligatoirement travailler en présentiel, sauf lorsque l'université était fermée où là on avait le choix, notamment lors des vacances scolaires.

II) Travail à réaliser et outils de mise en œuvre

1) Travail à réaliser

a) Basé sur un travail existant

Comme expliqué plus haut, notre mission lors de ce stage est de réaliser un webSIG (système d'information géographique) à partir de données brutes. Cependant, nous ne partons pas de zéro, en effet, un groupe de master 1 et un groupe de master 2 *Sigma* ont déjà travaillé sur ce sujet et nous devions nous baser sur leurs travaux déjà effectué, ils ont effectué un travail permettant de mettre en évidence les données qu'ils allaient utiliser pour le webSIG ainsi que les outils et moyens de représentations de celles-ci.

D'après leur rapport de stage, les M2 ont déjà effectué un travail de formatage et d'insertion sur base de données, mais ils ont tout effectué à la main sur *Microsoft Excel*, en plus sur des tout petits fragments de données test, le temps de calcul n'était pas du tout pris en compte.

Ils ont également élaboré une maquette de site elle-même issue de précédents travaux des M1 géomatique. Ils ont également programmé un prototype sur *R Shiny* mais il servait uniquement de test pour l'affichage, là aussi le temps de calcul n'était pas pris en compte. Donc notre objectif lors de ce stage est d'améliorer leur travail déjà effectué notamment en :

- automatisant le traitement des données qu'ils ont fait manuellement
- créant l'application *R Shiny* de visualisation des cartes
- mettant le tout disponible d'accès pour les chercheurs, sur un serveur

Nous les avons donc contactés pour obtenir leurs dossiers, rapport de stage, soutenances, et quelques informations supplémentaires.

b) Base de travail

Voici quelques extraits des informations qu'ils nous ont envoyé :

Informations sur les données brutes

Dans leur rapport de stage, les M2 ont détaillé les fichiers qu'ils ont utilisé ainsi que toutes informations qui pourraient être utiles pour faciliter le travail des personnes qui mettraient en place l'application, par exemple les données inutilisables (stations de test, valeurs aberrantes, spécificités…). Voici un exemple du type d'informations et du niveau de détails auquel nous avons eu accès.

KHLIFI Ali, FAUCHER Paul, & HADDOUCHE Chihab. (2022). *Web-SIG permettant de visualiser la comparaison des mobilités en période COVID et hors période COVID à Toulouse et à Lyon.*

Vélo : Ces données VLS sont des données protégées fournies par JCDecaux. Ces données brutes nous ont été transmises sous la fome de fichier csv (séparé par pointvirgule) de la forme :

```
Code borne sortie";"Borne sortie";"Date sortie CNIL (sans secondes)";"Code borne retour";"retour";"Date retour CNIL (sans secondes)"
```

```
6005;"6005 - PLACE EDGAR QUINET";"2019/01/01 00:00:00";2024;"2024 - R…PUBLIQUE / MAUPIN";"2019/01/01 00:20:00"
```

Cette structure de la donnée brute est la même pour les deux villes. On met en évidence que chaque enregistrement correspond à un flux avec le lieu et la date de sortie et de retour pour un même vélo. La précision de ces enregistrements est à la minute. On a aussi accès avec cette donnée à la relation entre le code de la borne et le nom de la borne, mais nous n'avons pas directement accès à la localisation géographique de ces bornes. Cette information de localisation des bornes est cependant facilement accessible en open source sur les sites data.Toulouse-métropole pour Toulouse et data.grandlyon pour Lyon. On remarque que les identifiants des bornes présents dans les données open sources correspondent exactement à ceux présents dans les enregistrements des trajets VLS fournis par JC Decaux. Après une étude plus approfondie des données, nous avons remarqué que les codes des bornes de Lyon vont de 201 à 34002 et ceux de Toulouse de 0 à 288. Nous avons ensuite essayé de déterminer si des bornes de Toulouse et de Lyon avait le même 'code borne', ce qui pourrait poser un problème lors d'un éventuel regroupement. C'est le cas uniquement pour le code 201. Ce code correspond toutefois à une station de test dans les données de Lyon. Cela nous a permis de mettre en évidence que certaines stations n'étaient en réalité pas de vraies stations, mais des stations de tests. Ces stations seront donc à supprimer lors de notre traitement de la donnée brute.

Schéma de la base de données

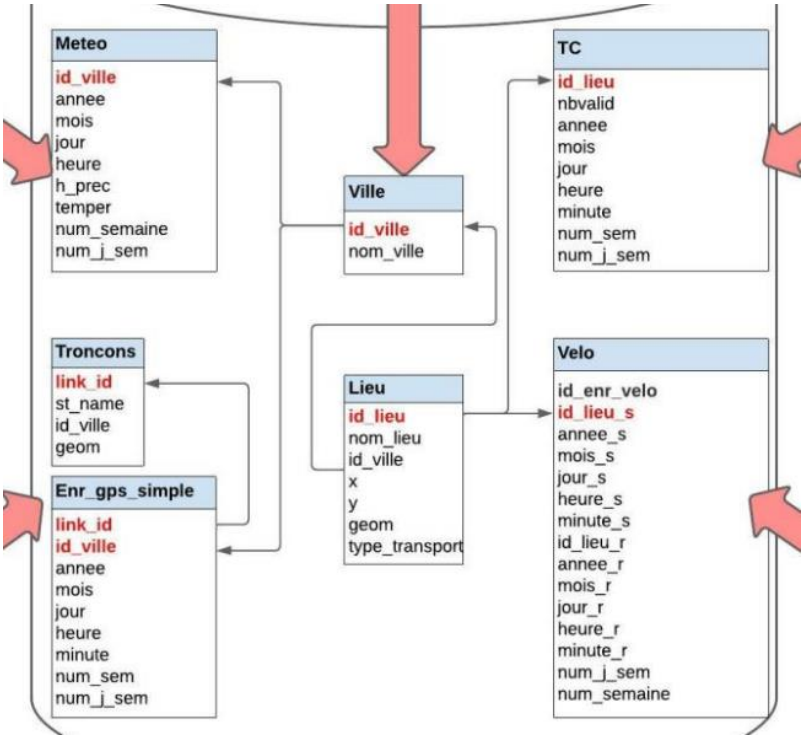


Figure 1 : MLD base de données issue du rapport de stage des M2

Les M2 ont dans leur rapport également laissé une trace de la base de données qu'ils ont conçu, il s'agit d'un MLD et nous n'avions pas d'informations supplémentaires sur les types ou encore la taille. Cependant, il était tout à fait compréhensible grâce aux compléments d'informations associées.

Maquette du site

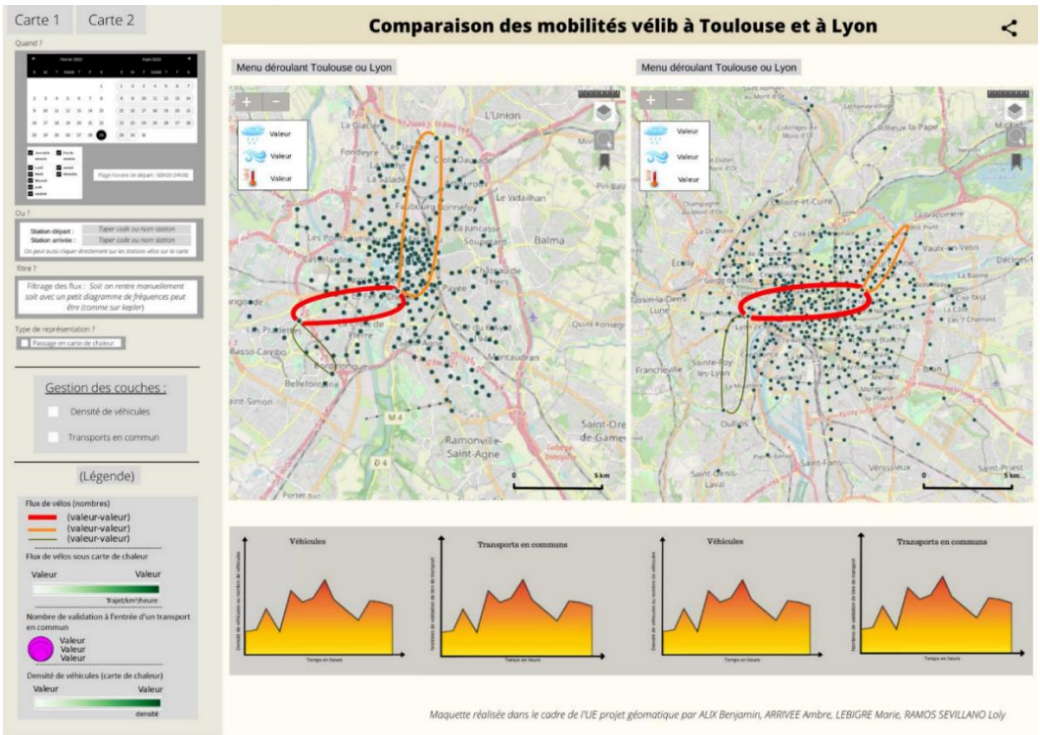


Figure 2 : Maquette de l'application issue du rapport des M1

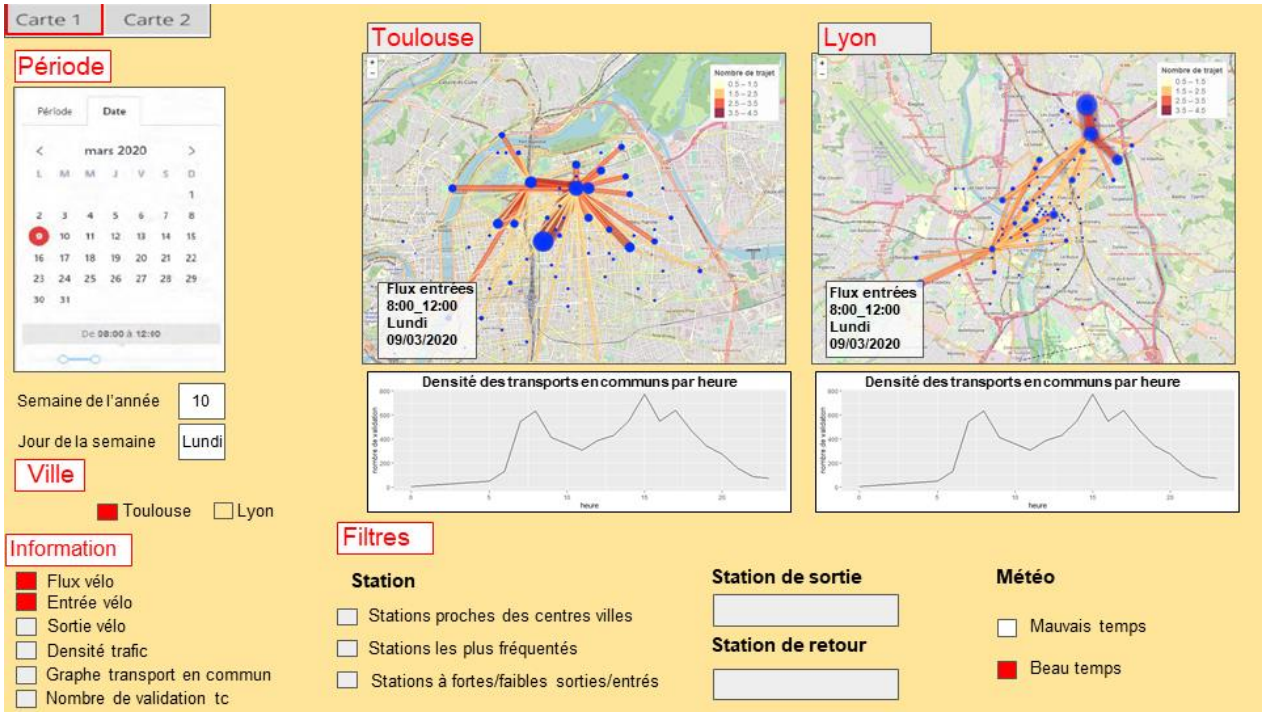


Figure 3 : Maquette de l'application issue du rapport des M2

Ces maquettes étaient ce à quoi le site devait ressembler. Après discussion avec le tuteur, nous avons estimé qu'il fallait apporter pas mal de modification à ces maquettes, voici quelques problèmes que nous avons pu souligner et par la suite résoudre au **III) 5)**.

- Que signifie le sélecteur Période/Date ?
- La comparaison se fait à Toulouse et Lyon
- Qu'est-ce que le beau temps ?
- A quoi correspond le centre-ville, notamment à Lyon
- …

2) Outils de mise en œuvre

Pour le choix des outils, nous avons décidé d'utiliser les mêmes outils que les M2 ont utilisés lors de leur précédent travail, le langage de programmation R uniquement (les M2 avaient également utilisé *Python* pour certains traitements car ils n'avaient pas réussi sur R, mais nous avons préféré ne pas tout mélanger), avec la bibliothèque *R Shiny* pour l'interface graphique ainsi que *Leaflet* pour la génération de cartes, le tout relié à une base de données *PostgreSQL*.

a) Langages de programmation

Le langage *R*

R est un langage de programmation et un logiciel libre (sous licence *GPL*) destiné aux statistiques et à la science des données. Dû à sa popularité parmi la communauté des data scientists, *R* dispose de nombreuses bibliothèques permettant d'avoir accès à des fonctions spécifiques très utiles pour traiter une grande quantité de données. *R* n'est cependant pas reconnu pour sa rapidité.

R Shiny

Shiny est un paquet de *R* qui facilite la création d'applications web interactives directement à partir de *R*.

b) La base de données

PostgreSQL

PosgreSQL est un SGBD lui aussi libre, il a été choisi car il possède l'extension *PostGIS*.

c) Outils spécifiques à la géomatique

PostGIS

PostGIS est une extension du SGBD *PostgreSQL*, qui active la manipulation d'informations géographiques sous forme de géométries, conformément aux standards établis par l'*OGS* (*Open Geospatial Consortium*). Il permet à *PostgreSQL* d'être un SGBD spatial (SGBDs) pour pouvoir être utilisé par les systèmes d'informations géographiques.

Leaflet

Leaflet est une bibliothèque *JavaScript* open source de cartographie en ligne. Elle est très complète et inclue toutes les fonctionnalités de mapping dont un développeur a besoin.

Type de données géométrique

C'est ce type de données qui est utilisé en géomatique voici quelques exemples de données que l'on peut représenter, ce sont celles qu'on a utilisées dans ce projet mais il en existe bien d'autres :

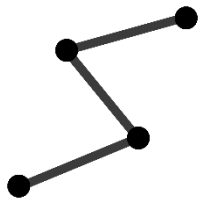
Points

Format données	Représentation	Explications
`(x , y)`		Un point est représenté par une coordonnée x et y

Lignes

Format données	Représentation	Explications
`[( x1 , y1 ) , (x2 , y2 ) ]`		Une ligne est représentée par deux points.

Chemins (tronçons)

Format données	Représentation	Explications
`[ ( x1, y1 ) , … , ( x2 , y2 ) ]`		Un tronçon est représenté par au moins deux points reliés entre eux, nous utilisons ce format pour représenter une portion routière.

d) *GeoJSON* et *ShapeFile*

Le format *GeoJSON* est un format ouvert d'encodage d'ensemble de données géospatiales simples utilisant la norme *JSON*. Le format est similaire qu'un simple format *JSON*, à l'exception de la représentation des informations géographiques.

Les données que nous ont fourni *Autoroutes-trafic* sont au format *ShapeFile*, le fonctionnement est similaire au format *GeoJSON* mais est moins pratique d'utilisation.

III) Analyse et méthodologie

1) Planification du travail

Nous avons consacré le premier jour de stage à l'analyse du travail à effectuer, ainsi qu'à la prévision des tâches à réaliser. Nous en avons profité pour lire les documents (annonce du stage, AAP, anciens rapports de stages…) afin de mesurer l'ampleur du travail et de réaliser le premier un diagramme de GANTT. Cependant, au fil du stage, certaines tâches ont été plus ou moins rapide à compléter, et d'autres se sont ajoutées. Voici le diagramme de GANTT final qui représente l'ensemble des tâches accomplies durant ce stage.

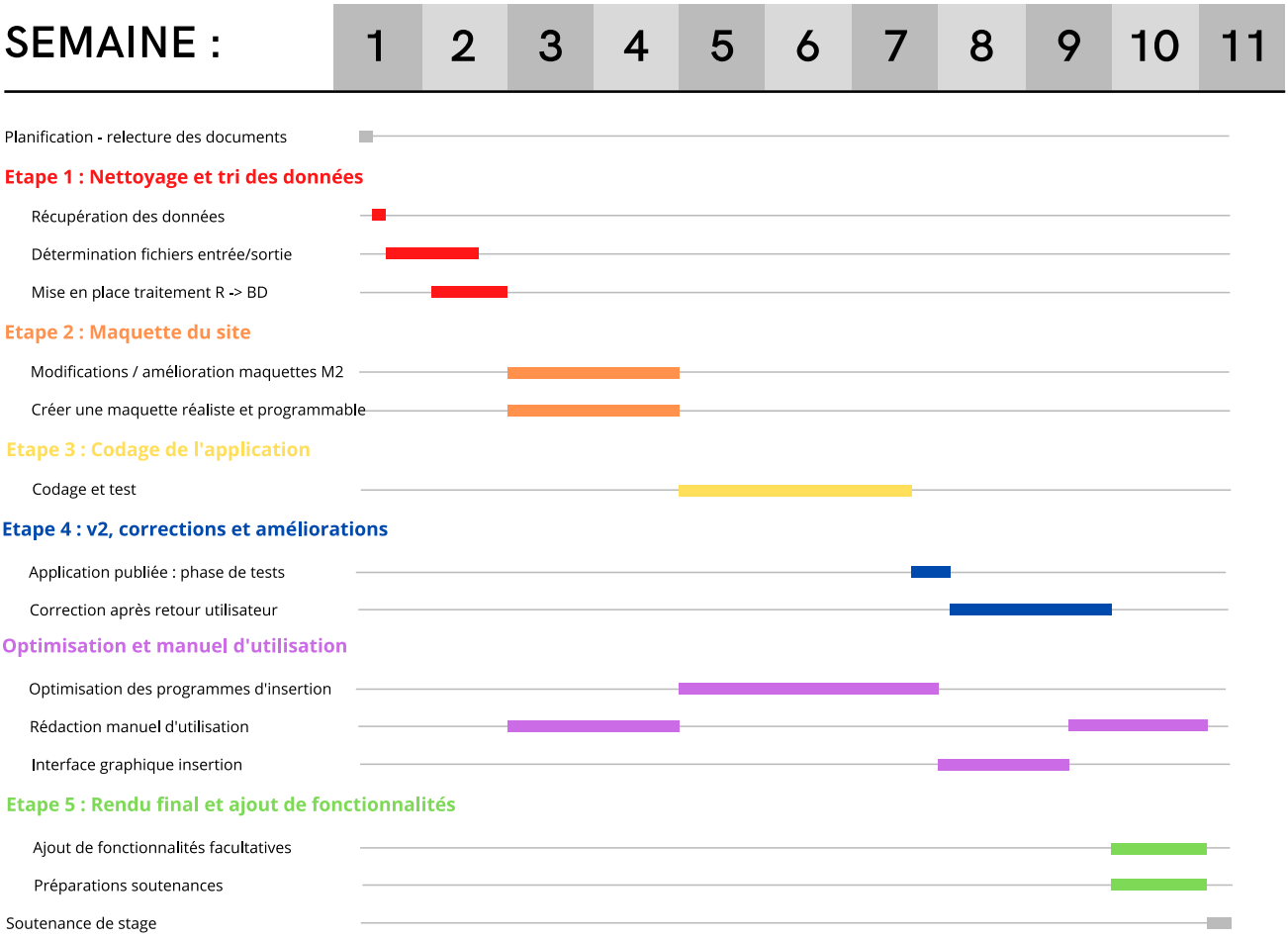


Figure 4 : Diagramme de GANTT du stage

Etape 1 : Nettoyage et tri des données

Cette étape est la première que nous avons dû réaliser, elle consistait à d'abord faire le tri parmi toutes les données que nous avions à notre disposition, et d'ensuite les nettoyer pour les rendre exploitables.

Etape 2 : Maquette du site

Cette partie consistait à discuter avec le tuteur et les membres de transition-vélo sur les maquettes des M2 afin de créer une maquette réaliste qui leur conviendrait le mieux.

Etape 3 : Codage de l'application

Mattéo a effectué seul cette partie, elle consiste à programmer sur *R Shiny* la maquette précédemment établie.

Etape 4 : v2, correction et améliorations

Nous avons mis à disposition l'application durant les 4 jours du pont de l'ascension afin que les chercheurs puissent tester l'application. A l'issue de cette phase, nous avons recueilli les avis afin d'améliorer l'application.

Optimisation et manuel d'utilisation

J'ai effectué cette partie seul pendant que Mattéo s'occupait de la visualisation des données. J'ai en premier lieu rédigé un manuel d'utilisation, mais au fil de l'expérience acquise sur *R* et des nouvelles fonctionnalités découvertes lors de l'optimisation des scripts, le manuel s'est retrouvé obsolète, et j'ai attendu la fin afin d'en rédiger un nouveau. J'ai également programmé une interface graphique afin de faciliter l'insertion des données brutes dans la base.

Etape 5 : Rendu final et ajout de fonctionnalités

Cette étape consiste à mettre en place sur le serveur final la base de données avec les données insérées et l'application. Ainsi que de préparer la fin du stage notamment en rédigeant tous les documents, et en préparant tous les rendus finaux. Cette étape ne sera pas présente dans le rapport car ultérieure à sa date de rendu.

Autres

Au début du stage, nous avions convenu d'au moins une réunion chaque semaine avec le tuteur chaque début de semaine, et lorsque c'était nécessaire, nous organisions une réunion avec les participants du projet transition-vélo sur le logiciel de communication *Zoom*.

2) Tri et nettoyage des données

Afin de pouvoir exploiter les données, il est nécessaire de les formater afin qu'on puisse les explorer depuis un programme. C'est d'autant plus nécessaire car les données mises à notre disposition sont d'une part issue de sources hétérogènes, mais en plus de formats hétérogènes (parfois plusieurs formats dans le même fichier), et surtout de taille conséquente. Décompressé, il y a en effet plus de 80 Go de données à traiter, ce qui est conséquent.

Les données

Les données ont été mises à disposition sur la plateforme *ODS My Core*, on outil du *CNRS* permettant le stockage et le partage de données jusqu'à 100 Go.

Ville	Catégorie	Fournisseur	Détails	Taille
Toulouse	VLS	*JCDecaux*	Déplacements 2019	400 Mo
			Déplacements 2020	300 Mo
			Emplacements bornes vélo	
			Nombre de locations 2019-2020	
		Open Data Toulouse	Emplacements bornes vélo	157 Ko
	TC	*TISSÉO*	Déplacements (bus + métro) 2019	674 Mo
			Déplacements (bus + métro) 2020	525 Mo
			Emplacements stations bus/métro	172 Ko
	TRAFIC	*Autoroutes Trafic*	Enregistrements GPS 2019	15 Go
			Enregistrements GPS 2020	13 Go
			Coordonnées troncons routiers	15 Mo
Lyon	VLS	*JCDecaux*	Déplacements 2019	884 Mo
			Déplacements 2020	747 Mo
		Data Grand Lyon	Emplacements bornes vélo	185 Ko
	TC	*SYTRAL*	Déplacements (métro) 2019	94 Mo
			Déplacements (métro) 2020	95 Mo
		Data Grand Lyon	Emplacements stations bus/métro	1 604 Ko
	TRAFIC	*Autoroutes Trafic*	Enregistrements GPS 2019	30 Go
			Enregistrements GPS 2020	27 Go
			Coordonnées troncons routiers	19 Mo
Toulouse Lyon	METEO	*MÉTÉO-FRANCE*	Météo Toulouse & Lyon 2019-2020	14 Mo
Données utilisées			Total : 88 Go	
Données partiellement utilisées				
Données non-utilisées				

Tableau 1 : Ensemble des données mises à disposition

Le tri des données et l'insertion dans la base de données est un travail qui a été précédemment effectué par les M2, cependant, ils n'ont pas automatisé cette tâche, ils ont tout fait à la main avec Microsoft Excel. Notre mission était d'automatiser cette tâche. De

plus, leur formation n'est pas spécialisée en informatique. Nous en avons profité pour apporter des modifications au schéma de la BD, notamment en appliquant entre autres une nomenclature plus rigoureuse, en choisissant des types de données plus adaptés et en ajoutant des index. L'état de ce schéma a bien sur évolué tout au long du stage pour arriver à l'état actuel.

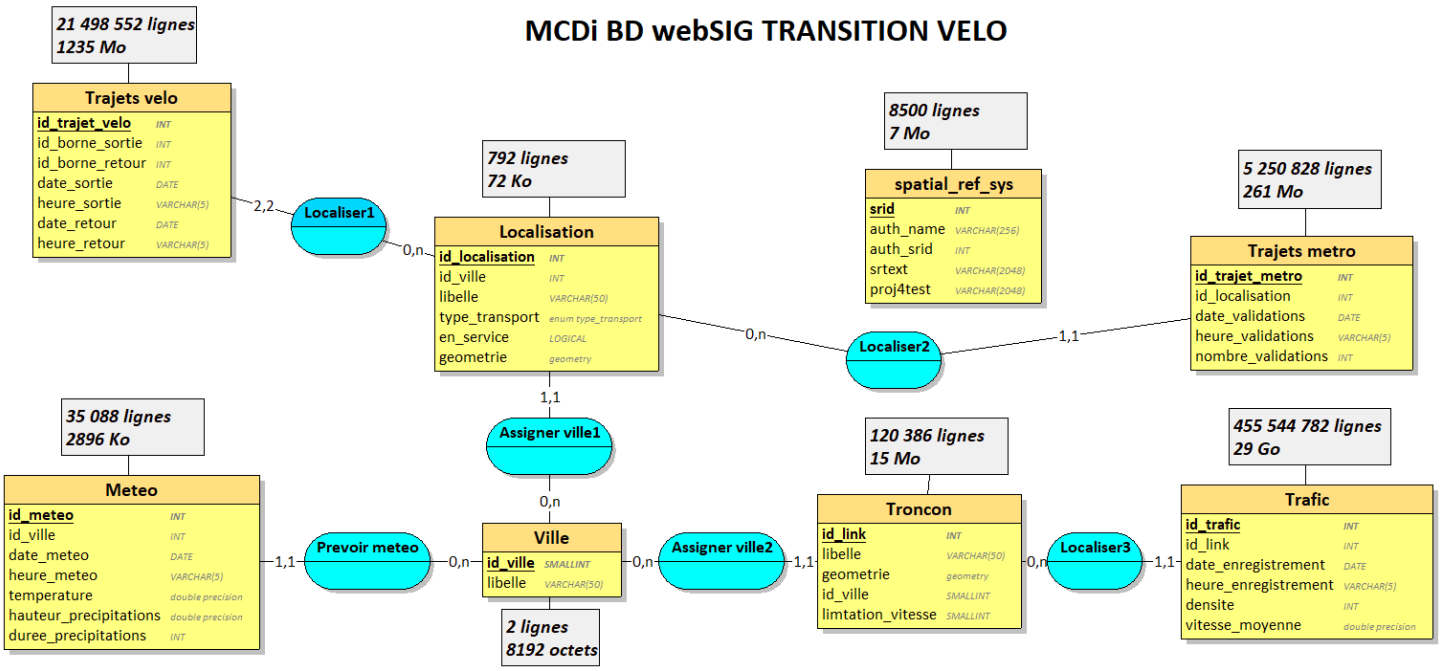
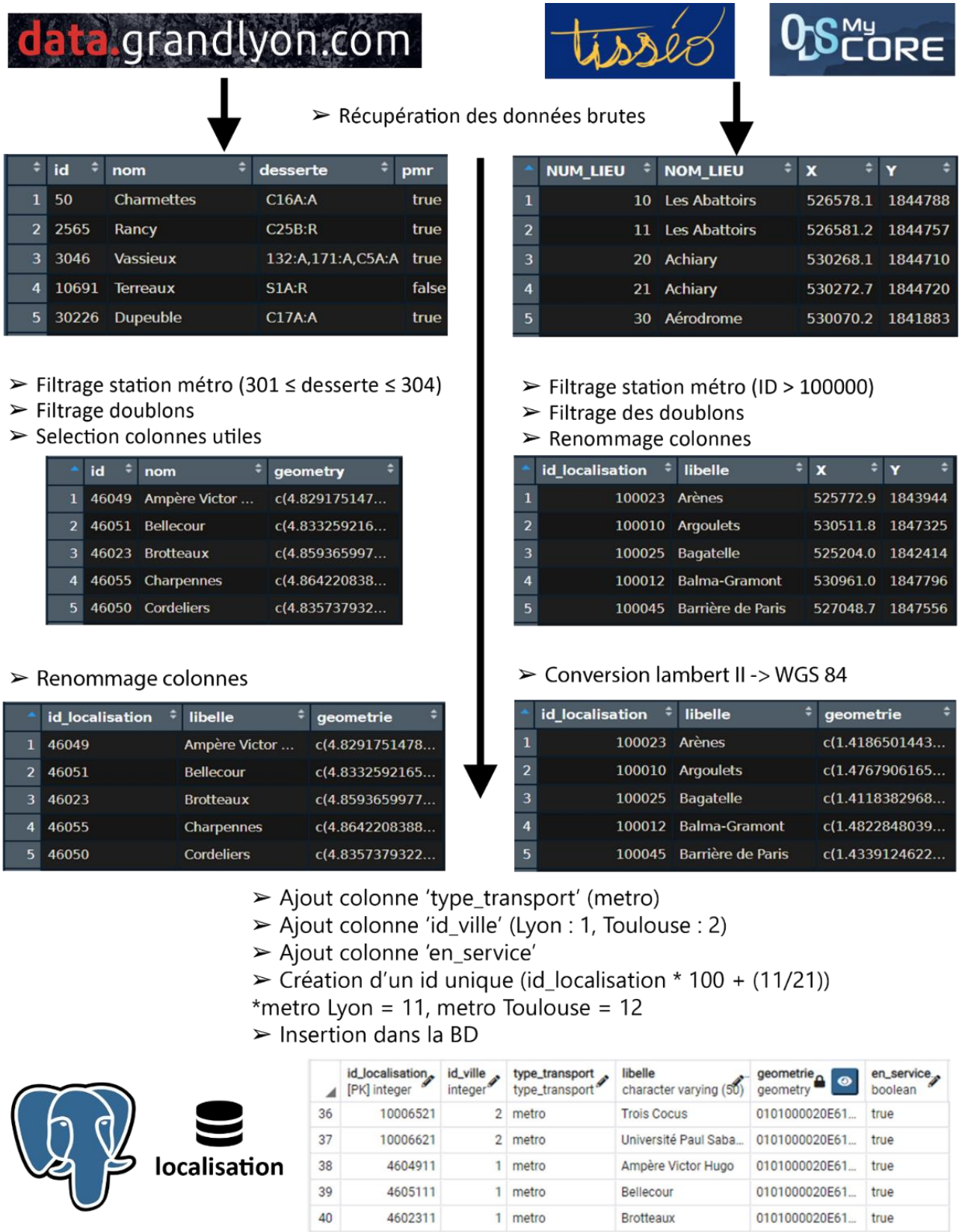


Figure 5 : MCDi de la base de données avec des informations sur la taille des tables

Formatage et insertion des données

Je vais maintenant montrer en vulgarisant le cheminement que vous avons emprunté pour passer des données brutes à un format adapté à la base de donnes. Je n'ai pas pu sélectionner tous les traitements mais seulement les plus pertinents.

Format stations de métro



	id	nom	desserte	pmr
1	50	Charmettes	C16A:A	true
2	2565	Rancy	C25B:R	true
3	3046	Vassieux	132:A,171:A,C5A:A	true
4	10691	Terreaux	S1A:R	false
5	30226	Dupeuble	C17A:A	true

	NUM_LIEU	NOM_LIEU	X	Y
1	10	Les Abattoirs	526578.1	1844788
2	11	Les Abattoirs	526581.2	1844757
3	20	Achiary	530268.1	1844710
4	21	Achiary	530272.7	1844720
5	30	Aérodrome	530070.2	1841883

	id	nom	geometry
1	46049	Ampère Victor ...	c(4.829175147...
2	46051	Bellecour	c(4.833259216...
3	46023	Brotteaux	c(4.859365997...
4	46055	Charpennes	c(4.864220838...
5	46050	Cordeliers	c(4.835737932...

	id_localisation	libelle	X	Y
1	100023	Arènes	525772.9	1843944
2	100010	Argoulets	530511.8	1847325
3	100025	Bagatelle	525204.0	1842414
4	100012	Balma-Gramont	530961.0	1847796
5	100045	Barrière de Paris	527048.7	1847556

	id_localisation	libelle	geometrie
1	46049	Ampère Victor ...	c(4.8291751478...
2	46051	Bellecour	c(4.8332592165...
3	46023	Brotteaux	c(4.8593659977...
4	46055	Charpennes	c(4.8642208388...
5	46050	Cordeliers	c(4.8357379322...

	id_localisation	libelle	geometrie
1	100023	Arènes	c(1.4186501443...
2	100010	Argoulets	c(1.4767906165...
3	100025	Bagatelle	c(1.4118382968...
4	100012	Balma-Gramont	c(1.4822848039...
5	100045	Barrière de Paris	c(1.4339124622...

	id_localisation [PK] integer	id_ville integer	type_transport type_transport	libelle character varying (50)	geometrie geometry	en_service boolean
36	10006521	2	metro	Trois Cocus	0101000020E61...	true
37	10006621	2	metro	Université Paul Saba...	0101000020E61...	true
38	4604911	1	metro	Ampère Victor Hugo	0101000020E61...	true
39	4605111	1	metro	Bellecour	0101000020E61...	true
40	4602311	1	metro	Brotteaux	0101000020E61...	true

Figure 6 : Procédé du traitement effectué sur les emplacements des stations de métro

Ce schéma simplifié montre bien le cheminement depuis les données brutes jusqu'à la table *localisation* dans la base de données.

Les données brutes sont issues de deux sources différentes, mais doivent "rentrer" dans la même table. En effet, sur cet exemple les sources brutes sont totalement différentes, dans les données de l'open data de Lyon, pour déterminer si une station est une station de métro et non de bus, il faut vérifier la desserte (301 = ligne A, 302 = ligne B…) via une expression régulière.

```
filter(str_detect(desserte, "^[a-zA-Z0-9:,]*(301|302|303|304)[a-zA-Z0-9,:]*$"))
```

Extrait de code 1 : Filtre station métro Lyon

Tandis que pour les données issues de *Tisséo* pour Toulouse, les stations de métro ont un identifiant supérieur ou égal à 100010.

```
%>% filter(NUM_LIEU >= 100010) %>%
```

Extrait de code 2 : Filtre station métro Toulouse

De plus, sur les données de Toulouse, le format des coordonnées GPS n'est pas le même, en effet, les coordonnées des stations de Lyon sont au format géométrique `Point ( x , y )` en *WGS 84* tandis que sur le fichier de *Tisséo*, les données sont représentées sur deux colonne X et Y au format *Lambert II Etendu*, un format très peu utilisé et seulement en France. Un traitement supplémentaire a été nécessaire.

Une fois ce traitement terminé les deux jeux de données se retrouvent exactement au même format. Avant de l'insérer dans la base, on rajoute les colonnes nécessaires qui ont été déterminées en amont pour pouvoir manipuler plus tard les données avec plus d'aisance, notamment la colonne *id_ville* qui permettra de localiser la station lors d'une requête par exemple, ou encore la colonne *id_localisation*, qui permet d'identifier une station, car dans les jeux de données, certaines stations de villes différentes peuvent avoir le même identifiant. Pour cela, le traitement suivant est effectué pour différentier ces stations :

$$id\_localisation = id\_station * 100 + code\_ville$$

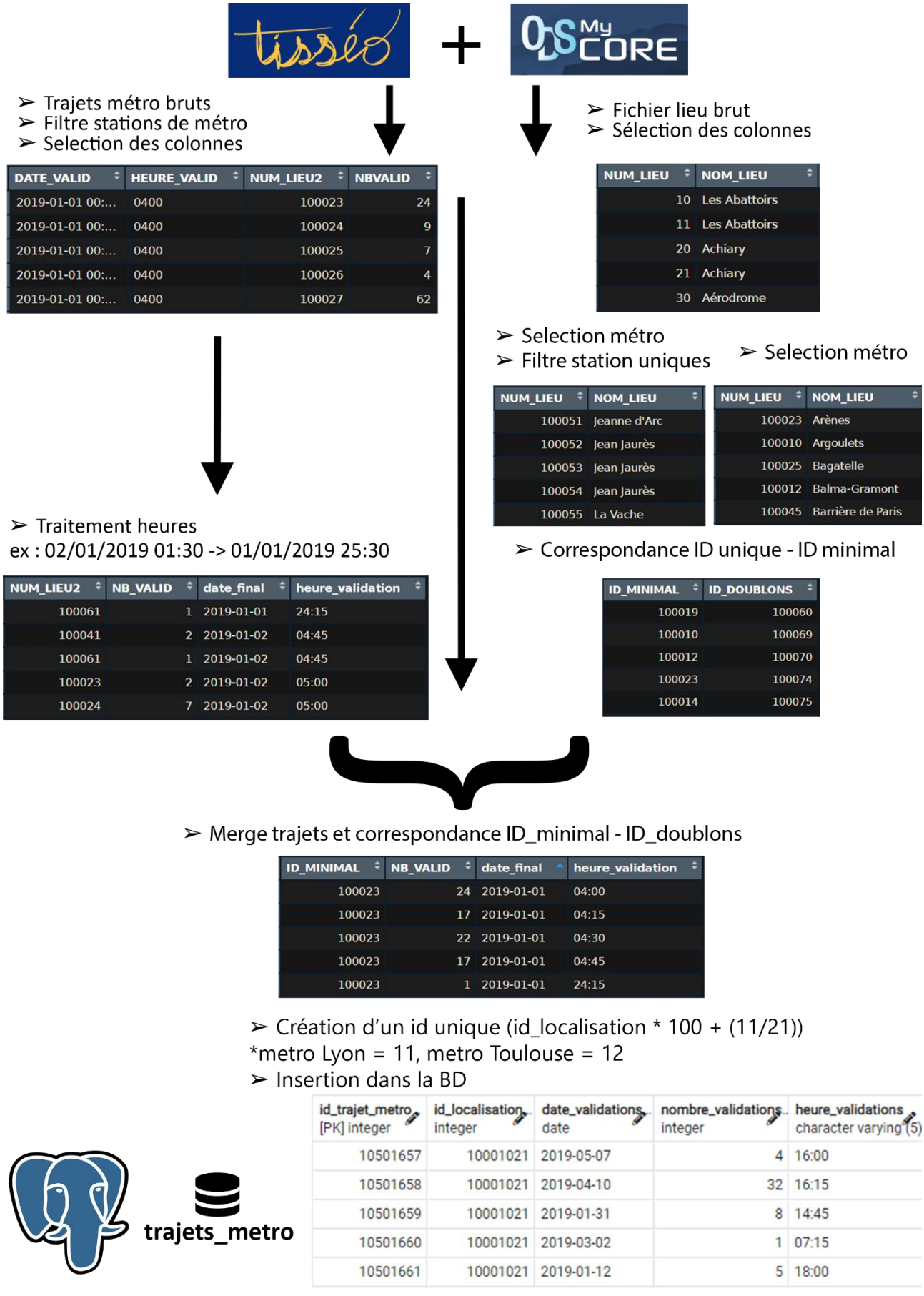
code_ville correspond à une constante attribuée arbitrairement pour chaque ville et chaque type de transport, voici la liste des constantes.

	ville	id_ville	id_unique_metro	id_unique_velo
1	lyon	1	11	12
2	toulouse	2	21	22

Figure 7 : Dataframe des identifiants pour chaque ville et type de transport

Ainsi, grâce à cette formule. chaque station est assurément unique et aucun conflit de clé unique ne peut se produire lors de l'insertion.

Format trajet métro Lyon



DATE_VALID	HEURE_VALID	NUM_LIEU2	NBVALID
2019-01-01 00:...	0400	100023	24
2019-01-01 00:...	0400	100024	9
2019-01-01 00:...	0400	100025	7
2019-01-01 00:...	0400	100026	4
2019-01-01 00:...	0400	100027	62

NUM_LIEU	NOM_LIEU
10	Les Abattoirs
11	Les Abattoirs
20	Achiary
21	Achiary
30	Aérodrome

NUM_LIEU	NOM_LIEU
100051	Jeanne d'Arc
100052	Jean Jaurès
100053	Jean Jaurès
100054	Jean Jaurès
100055	La Vache

NUM_LIEU	NOM_LIEU
100023	Arènes
100010	Argoulets
100025	Bagatelle
100012	Balma-Gramont
100045	Barrière de Paris

NUM_LIEU2	NB_VALID	date_final	heure_validation
100061	1	2019-01-01	24:15
100041	2	2019-01-02	04:45
100061	1	2019-01-02	04:45
100023	2	2019-01-02	05:00
100024	7	2019-01-02	05:00

ID_MINIMAL	ID_DOUBLONS
100019	100060
100010	100069
100012	100070
100023	100074
100014	100075

ID_MINIMAL	NB_VALID	date_final	heure_validation
100023	24	2019-01-01	04:00
100023	17	2019-01-01	04:15
100023	22	2019-01-01	04:30
100023	17	2019-01-01	04:45
100023	1	2019-01-01	24:15

id_trajet_metro [PK] integer	id_localisation integer	date_validations date	nombre_validations integer	heure_validations character varying (5)
10501657	10001021	2019-05-07	4	16:00
10501658	10001021	2019-04-10	32	16:15
10501659	10001021	2019-01-31	8	14:45
10501660	10001021	2019-03-02	1	07:15
10501661	10001021	2019-01-12	5	18:00

Figure 8 : Procédé simplifié du formatage des trajets métro Toulouse

Voici un deuxième exemple de formatage de données. Il s'agit de l'insertion de trajets de métro de la ville de Toulouse dans la table *trajets_metro*. Dans ce cas-ci, les sources des données sont issues du même organisme, mais plusieurs fichiers sont nécessaires afin de pouvoir formater les données.

En effet, le système de *Tisséo* est conçu de telle sorte que chaque groupe de borne de validation de ticket corresponde à un id et à un nom de station associée, il y a donc des stations avec plusieurs identifiants, ce qui serait gênant lors de l'affichage de ces données par la suite.

NUM_LIEU	NOM_LIEU
100051	Jeanne d'Arc
100052	Jean Jaurès
100053	Jean Jaurès
100054	Jean Jaurès
100055	La Vache

Tableau 2 : Extrait d'un fichier TISSEO contenant l'emplacements des stations de transports en commun

Dans cet extrait de données, on voit bien que la station de métro Toulousaine Jean Jaurès est associé à plusieurs identifiants, dans ce cas, l'importation du fichier contenant les lieux des stations est nécessaire en plus de celui des trajets, car il faut pouvoir déterminer un identifiant unique pour chaque station. Le traitement est effectué sur le fichier des lieux et ensuite on *merge* les trajets et les lieux afin d'obtenir un identifiant unique pour chaque station de métro.

3) Optimisation

Diminution du temps de traitement

Comme indiqué sur le diagramme de GANTT, les scripts de formatage du point précédent ont été créés pendant les deux premières semaines du stage. Cependant au vu de nos connaissances de *R*, les programmes n'étaient pas du tout optimisés et les temps d'exécution des premières versions de ces programmes était très longs pour certains (plusieurs semaines pour le trafic notamment). J'ai donc décidé d'optimiser ces fichiers, voici un exemple du type de manipulation que j'ai réalisé et qui ont permis un gain de temps notoire sur l'exécution des programmes.

```
boucle <- function(i, trajets, file, subpart, qte) {
  print(paste("Part", file, "/", subpart, ":", i*100/ qte,"%"))
  trajetsTemps <- trajets[trajets$date_temps_metro == trajets$date_temps_metro[i],]
  rowName <- as.numeric(row.names(trajetsTemps[1,]))
  j <- i-rowName+1
  idTrajet <- trajetsTemps$id[j]
  ids <-
      na.omit(metroUnique$NUM_LIEU[metroUnique$NOM_LIEU == metroUnique$NOM_LIEU
      [metroUnique$NUM_LIEU == idTrajet]
                                    & metroUnique$NUM_LIEU != idTrajet])
  idTrajetMin <- min(append(ids, idTrajet))
  if (length(ids) > 0) {
    trajets$id[i] <<- idTrajetMin
  }
}
lapply(1:qte,FUN=boucle,trajets=trajets,file=file,subpart=subpart, qte = qte)
```

```
trajets <- base::merge(x = trajets, y = correspondance_id_stations, by="ID_DOUBLONS")
```

Extrait de code 3 : Optimisation script d'insertion des trajets de métro Toulouse

Programmation de l'interface graphique

Cette fonctionnalité était facultative, mais en rédigeant un premier manuel d'utilisation, je me suis rendu compte que l'insertion via les scripts R était très laborieuse pour un non-informaticien et ne fonctionnait pas à la moindre erreur. J'ai donc réalisé **une interface graphique** réactive, l'utilisateur choisit le type de données qu'il souhaite insérer, et l'interface s'adapte en fonction de ce qu'il souhaite.

4) Nouvelles contraintes et ajout de fonctionnalités

Nouvelles fonctionnalités

Au fil des réunions, de nouvelles suggestions et idées ont été émises par les chercheurs, si certaines de ces idées nécessitait un ajout ou tout simplement une légère modification, d'autres nécessitaient une modification profonde de la structure que nous avons établie.

Par exemple, lors d'une réunion le 23/05, lorsque l'application était donc à un état assez avancé, le chercheur Bruno Revelli a émis le problème suivant : comment faire pour observer les données de 23h00 à 02h00 chaque jour sur une période donnée ?

Car en effet, l'application contenait un calendrier pour choisir une période ainsi qu'un curseur allant de 00:00 à 23h59, ce qui rendait ce cas d'utilisation impossible bien qu'il soit pertinent.

Nous avons donc choisi de modifier le stockage des dates sur toute les tables de la base contenant un champ *date* afin de permettre ce cas d'utilisation. Ce qui a nécessité les changements suivants sur les tables *trafic*, *trajets_metro*, *trajets_velo* et *meteo*.

date_heure (timestamp)
2020-10-23 01:15

date_... (date)	*heure (varchar(5))*
2020-10-22	25:15

Figure 9 : Exemple de modifications effectuées sur la structure de la base de données

Nouveaux jeux de données

Lors du stage, nous avons dû faire face à d'autres imprévus, en effet lors du début du stage, nous avion remarqué que pour les données du métro Lyonnais, nous ne disposions que des données allant de Janvier 2019 à Octobre 2020, il manquait les données des 2 derniers mois de l'année 2020. J'ai donc programmé les scripts ainsi que tout le traitement nécessaire au formatage de ces fichiers. Cependant lorsque notre tuteur nous a remis plus tard le 18/05 le reste des données, il s'agissait des mêmes données mais formatées différemment.

```
Date jour (CAS tr) Mois   Année Code ligne   Station      Jour   Tranche       horaire
      Heure (CAS tr)      Nb entrées total (CAS tr) Nb sorties total (CAS tr)
01/01/2019   1      2019   B      Brotteaux    2      (vide) 05:00:00     36     2
07/03/2019   3      2019   B      Jean Macé    4      (vide) 19:30:00     335    403
26/01/2019   1      2019   D      Valmy  6     (vide) 00:45:00     1      72
```

```
Ligne  StationCode ligne TITAN      Code ligne     Date jour (CAS tr)    Heure (CAS tr) Nb
entrées total (CAS tr)       Nb sorties total (CAS tr)
Charpennes - Gare d'Oullins  BRO    302    B      01/01/2019    05:00  36     2
Charpennes - Gare d'Oullins  MAC    302    B      07/03/2019    19:30  335    403
Gare de Vaise - Gare de Vénissieux  VMY    304    D      25/01/2019    24:45  1      72
```

Figure 10 : Modifications du fichier des trajets de métro de Lyon

Il a fallu refaire tout le programme, car dans le nouveau fichier avait une structure totalement différente de l'ancienne, particulièrement pour ce qui est du format de l'heure qui est passé à *00:00 - 27:59* ainsi que les noms de stations *Station* qui sont passés du nom en toutes lettres à un identifiant de 3 lettres impossible à déterminer à partir du nom de la station. Il a fallu créer un fichier de correspondance nom de station-code.

5) Visualisation des données

La partie carte et visualisation des données a été réalisée par Mattéo tandis que je m'occupais de l'optimisation et de l'interface graphique. Voici brièvement le travail qu'il a effectué.

Maquette de l'application

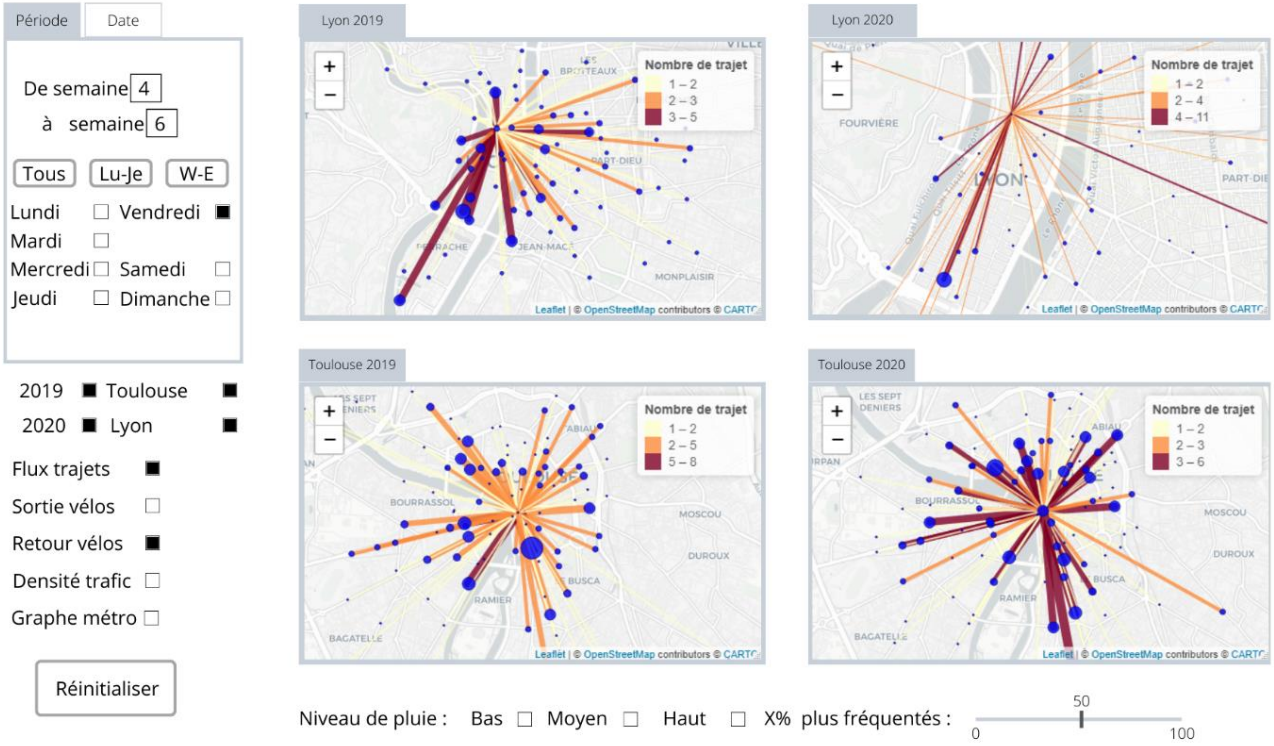


Figure 11 : Maquette de l'application

Cette maquette a été réalisée à la suite d'une réunion tenue avec le tuteur de stage, nous sommes partis de la maquette des M2 et avons supprimé/ajouté des fonctionnalités visibles sur la maquette.

Programmation de l'application

Cette partie consiste à programmer la maquette précédemment établie sur *R* avec l'aide de la bibliothèque *Shiny*. **Les cartes** sont générées grâce à *Leaflet* et les données affichées proviennent de la base de données.

IV) Résultat final et évaluation

1) Insertion des données

Base de données

Concernant la base de données, nous avons fourni une base de données remplie et formatée accessible sur un serveur fourni par la *TGIR Huma'num.* En plus du MLD correspondant ainsi que du fichier .sql permettant l'importation du schéma de la base.

Scripts R permettant l'insertion de données

Voici l'ensemble des scripts R permettant le formatage et l'insertion des données que nous avons programmées lors du début du stage et que j'ai par la suite optimisé.

Script R	Explications	Tps. Exec. (sans insertion)
FORMAT_LIEUX.R	Format et insertion des fichiers bruts dans une BD	<1 sec
FORMAT_TRAJETS_VELO.R		Toulouse : 2 min 50 / trimestre Lyon : 4 min / an
FORMAT_TRAJETS_METRO.R		1 min / an / ville
FORMAT_TRONCONS.R		<3 sec
FORMAT_TRAFIC.R		30 min / an / ville
FORMAT_METEO.R		<3 sec

Tableau 3 : Liste des scripts de formatage fournis au client

Le premier point qui était d'automatiser les tâches effectuées à la main par les M2 a été remplie avec succès. Les chercheurs ont simplement besoin en entrée des données brutes, et obtiennent en sortie les fichiers formatés et directement insérés sur la BD. De plus, le traitement est relativement rapide.

Manuel d'utilisation

J'ai commencé à rédiger un manuel d'utilisation afin d'expliquer tout ce qui est en lien avec l'insertion, en expliquant le format exact que doivent avoir les fichiers ainsi que la procédure à suivre pour insérer les données dans la base. Le document n'est pas encore terminé à l'heure du rendu de ce rapport.

Interface graphique

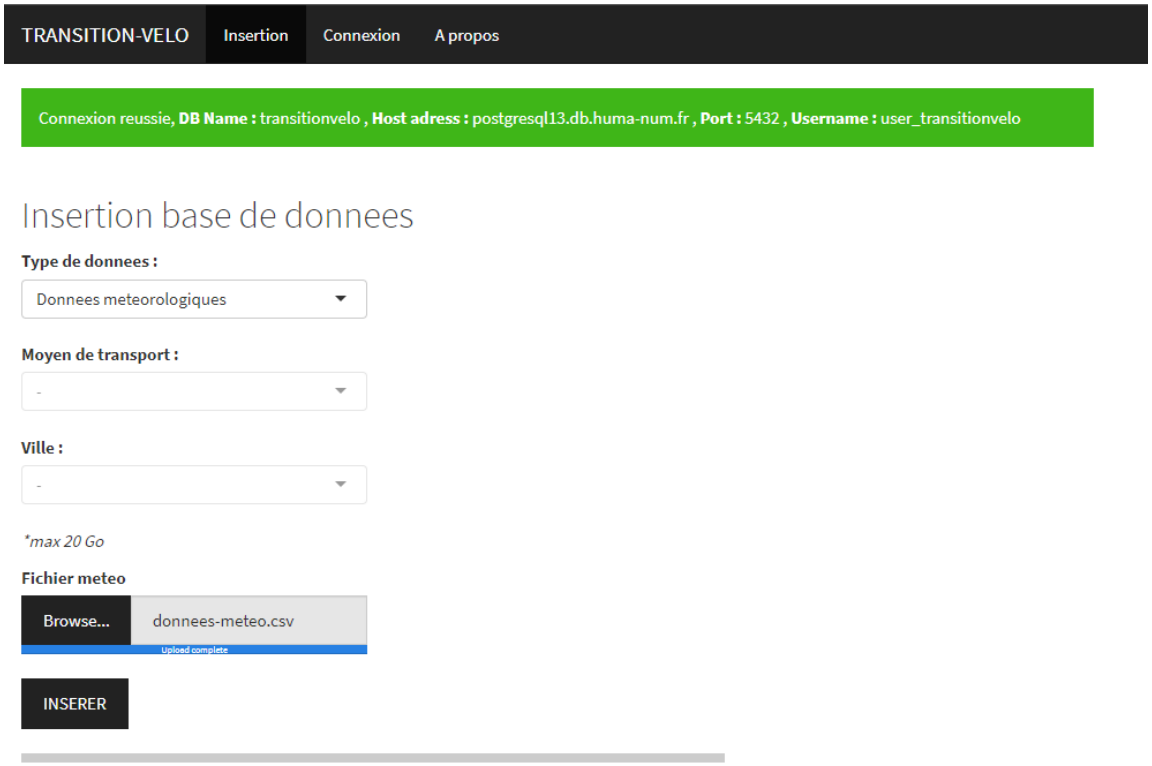


Figure 12 : Capture d'écran de l'interface graphique de l'insertion des données

Cette interface n'était pas une fonctionnalité primordiale, mais elle facilite grandement la tâche pour les futurs utilisateurs.

2) Visualisation des données

Cartes *R Shiny*

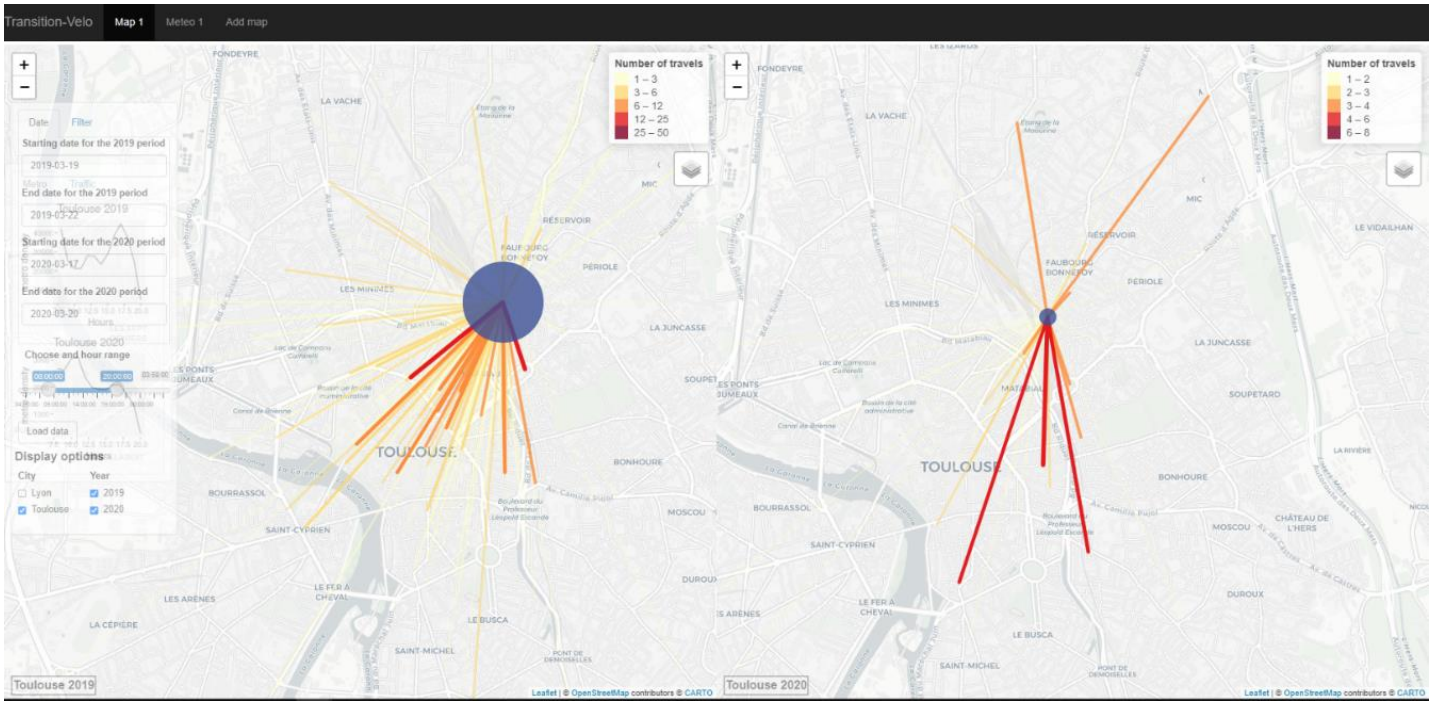


Figure 13 : Capture d'écran de l'application

3) Evaluation des résultats

Afin d'évaluer les résultats, nous nous sommes basés sur les remarques de notre tuteur, elles étaient très positives à l'égard de notre travail. Lors de la phase de tests, nous avons fait parvenir au tuteur une **fiche de recueil d'avis** afin d'avoir clairement son retour en forme écrite, une idée suggérée par notre tuteur. Il était très satisfait du résultat en plus de nous avoir fourni des remarques afin d'améliorer le webSIG.

4) Points à améliorer / fonctionnalités non implantées

A l'issue de ces 10 semaines de stage, les objectifs principaux ont été réalisés.

- automatiser le traitement et l'insertion des données
- créer une application R Shiny pour visualiser ces données
- faire en sorte qu'elle soit accessible aux chercheurs (*à l'heure où le rapport est rendu, le stage n'a pas été terminé, le déploiement est en cours*)

Cependant, certaines fonctionnalités qui pourraient améliorer l'application n'ont pas été implantées, soit dû à une limitation des outils, soit dû à un manque de temps. Voici la liste de quelques fonctionnalités notables non-implantées :

- carte affichant le trafic (manque de temps)
- export d'une carte en pleine qualité (difficulté -> manque de temps)
- possibilité d'ajouter une ville ou une année de données (pas les données)

V) Bilan

Bilan professionnel

A l'issue de ce stage, nous avons pu fournir aux chercheurs une application permettant la visualisation des données des déplacements VLS des villes de Toulouse en 2019 et 2020. En plus d'une base de données ainsi que d'une application permettant l'insertion de ces dites données. Comme vu précédemment, le tuteur est très satisfait de notre travail, tandis que l'avis des autres chercheurs sont en attente et devraient arriver après la rédaction de ce rapport.

Cependant, l'application peut être encore soumise à amélioration, notamment en ajoutant les fonctionnalités non implantées telles que la visualisation du trafic ou bien des autres transports en commun (bus) ou encore en rajoutant la possibilité de comparer les données d'autres villes en Europe ou d'autres années à Toulouse et Lyon.

Bilan personnel

Lors de ce stage, j'ai été surpris de ce dont j'ai été capable de produire, en effet, j'ai pu appliquer les enseignements que j'ai reçu à l'IUT sur le terrain, et j'ai pu trouver un lien direct entre les connaissances acquises en cours et les moyens techniques mis en œuvre pour résoudre de nombreux problèmes.

Techniquement, ces deux mois de travail à temps plein sur *R* m'ont permis d'acquérir une bonne maîtrise de ce langage, ainsi qu'une première expérience avec de telles quantités de données. J'ai également pu développer mon autonomie, notamment en apprenant à bien lire les documentations des fonctions ou bien en recherchant en ligne sur des forums des réponses aux problèmes rencontrés. De plus, j'ai acquis de nouvelles compétences diverses : utilisation de *pgAdmin*, PostgreSQL, *postGIS*, des données géométriques, *Git*, connexion avec une base extérieure et non en local, de nouvelles fonctionnalités sur *Microsoft Word* et bien d'autres…

Pour ce qui est des compétences de communication, j'ai pu découvrir la communication en entreprise, notamment l'aspect réunion hebdomadaire afin de faire un point sur l'avancée des résultats ainsi que l'aspect envoi et échanges de mails pour tenir au courant toute l'équipe concernée.

Conclusion

En conclusion, ce stage a concrétisé à merveille mes deux années d'apprentissage au sein de l'IUT. J'ai pu mettre en pratique ainsi qu'améliorer l'ensemble de me compétences qu'elles soient d'un point de vue technique ou bien communicationnelles lors de la confection du webSIG pour les chercheurs travaillant dans le cadre du projet *transition-vélo*.

Maintenant, il reste à voir si l'application servira dans le temps, et si l'ergonomie et les choix visuels de la représentation des données seront exploitables et si les chercheurs arriveront à repérer des phénomènes intéressants à travers cette application.

Actuellement, notre tuteur nous a proposé un contrat à durée déterminée de deux semaines après ce stage afin de rajouter quelques fonctionnalités, notamment la visualisation du trafic sur la carte.

J'ai personnellement très bien vécu ce stage et je me suis senti à ma place dans l'équipe, si travailler dans le milieu de l'informatique s'apparente à cela, je serai alors ravi de poursuivre mes études et de commencer ma carrière dans ce domaine. Cependant cette expérience s'est déroulée dans le milieu de la recherche et non dans le monde en entreprise en tant que tel, j'aimerais lors de mon prochain stage essayer le milieu en entreprise afin de me forger un avis d'un autre point de vue.

Glossaire

SIG = Système d'Information Géographique web

webSIG = Système d'Information Géographique consultable sur un page web

VLS = Vélo en Libre-Service

TC = Transports en Commun

M1, M2 Sigma = Master Sciences Géomatiques en Environnement et Aménagement

Géomatique = Ensemble des outils et méthodes permettant d'acquérir, de représenter, d'analyser et d'intégrer des données géographiques. Le mot « géomatique » est issu de la contraction des termes géographie et informatique.

BD = Base de données

MLD = Modèle Logique des Données

MCDi = Modèle Conceptuel des Données avec identifiants

SGBDs = Système de gestion de base de données spatiales

Dataframe = Un tableau ou une structure de type tableau à deux dimensions.

Bibliographie, sitographie

KHLIFI Ali, FAUCHER Paul, & HADDOUCHE Chihab. (2022). *Web-SIG permettant de visualiser la comparaison des mobilités en période COVID et hors période COVID à Toulouse et à Lyon*.

Dorian Martineau. (2021). *Effet de la COVID-19 sur l'utilisation des vélos en libre-service à Toulouse : Approche par la physique statistique*.

Hervé Groléas. (2015, 15 janvier). *Points d'arrêt du réseau Transports en Commun Lyonnais*. data.grandlyon.com. Consulté le 20 avril 2020, à l'adresse https://data.grandlyon.com/jeux-de-donnees/points-arret-reseau-transports-commun-lyonnais/info

Jean-Luc Moudenc. (2011, 19 mars). *Station Vélô Toulouse - Toulouse*. data.toulouse-metropole.fr. Consulté le 20 avril 2022, à l'adresse https://data.toulouse-metropole.fr/explore/dataset/station-velo-toulouse/information/

Annexes

WEBSIG TRANSITION-VELO – FICHE DE RECUEIL D'AVIS

Camin Mattéo : matteo.camin@etu.iut-tlse3.fr
Escudié Tom : tom.escudie@etu.iut-tlse3.fr

Nom & prénom : Jouve Bertrand

Points positifs :

- Bonne rapidité de l'application
- Tout est assez intuitif donc facile à utiliser

Points à améliorer :

- Dans les cartes à afficher (non, output, input, output+input), être plus explicite sur la légende : qu'est-ce que c'est. Le diamètre des disques autour des stations doit être adapté. On ne voit rien en 2019.
-

J'ai testé la comparaison des 12-15 mai 2020 avec 14-17 mai 2019.

Suggestion / idées de fonctionnalités à ajouter (recherche, interface...) :

- Date sarting and end dates : en choisissant 2020, 2019 doit se remplir automatiquement
- Les fenêtres mobiles ne peuvent pas être affichées sur un autre bureau
- Je ne sais pas si c'est intéressant de mettre « metro » et « traffic » comme case à cocher car ils sont sur une fenêtre séparée et ça n'a pas l'air de couter grand-chose de les avoir.
- On peut peut-être ajouter la possibilité de fermer une carte à partir de la carte elle-même, de façon classique avec une case à cocher de type une croix dans un carré ou un cercle.
- SI on dispose de deux écrans, ce pourrait être bien de pouvoir afficher deux cartes sur un écran et deux autre sur l'autre ?
- La grosseur des traits de flux est peut-être trop importante. Si on filtre peu, on ne voit rien.
- Peut être mettre « querying the database, please wait a moment » en affichage clignotant. D'ailleurs cette phrase importante, s'efface pendant que les cartes s'affichent ce qui est dommage.
- Comment est découpée l'échelle des valeurs de flux (couleurs) ?

Bogue découvert (méthode pour reproduire, effet attendu, effet effectif) :

- Quand on change deux fois d'horaires de façon assez rapide mais qu'on a appuyé sur « «loading data » , l'appli charge quand même les deux requetés de façon successives. Il faudrait que la deuxième requête soit prioritaire et arrête la première.
- Quand on veut changer du métro au traffic, les courbes de traffic ne s'affichent pas
- J'ai l'impression que le nombre d'output ou d'input sur une station ne correspond pas à la somme des outputs (input) des flux afférents à la station.
-

Votre avis sur l'application :

SUPER !!

Application version 25/05/2022

Annexe 1 : Fiche de recueil d'avis de notre tuteur lors de le première phase de test