Rapport de 1ère année
de Master Informatique
Année scolaire 2021-2022

DM de Bases de données non traditionnelles

Université de Caen, Normandie
UFR des Sciences

Rédigé par
Guillaume LETELLIER (21804030)
Corentin PIERRE (21803752)

1 But de l'application

L'objectif de cette application est de créer une interface de visualisation de données choisies par nous-mêmes. Il faut être capable de traiter les données, ajouter les entrées dans des collections au sein de MongoDB et les afficher sur une interface web. Pour cela, nous avons utilisé MongoDB comme base de données, Apache pour faire le serveur, PHP et langages front end afin d'afficher les données et enfin, GraphQL pour faire le lien entre la base de données et l'interface web. Le but d'utiliser ces différentes technologies est de nous initier aux méthodes de développement d'applications modernes.

2 Les données

2.1 D'où viennent-elles ?

Nous avons décidé de prendre des données concernant le retard des trains de la SNCF. En effet, nous nous sommes intéressés à ce jeu de données car nous étions curieux de connaître les raisons des retards des TGV, avancés par la SNCF. Cette idée nous est venue en cherchant un jeu de données pour réaliser l'application. En nous baladant sur différents sites comme Kaggle, nous sommes tombés sur celle-ci. Nous avons donc ensuite décidé de chercher des données plus complètes et nous avons donc très rapidement trouver le site "Open Data" de la SNCF [1]. Toutes les données que nous avons eu proviennent de ce site. À noter que les données concernant les départements et une partie du code servant à leur affichage proviennent de ce tutoriel.

2.2 Comment sont-elles obtenues ?

Comme dit précédemment, nous les avons obtenus principalement sur le site de la SNCF. Sur ce site, plus de 200 jeux de données sont disponibles. Cela peut être des données sur la régularité des trains, sur la fréquentation des gares, les infrastructures ou encore sur les mouvements sociaux. Pour notre part, nous nous sommes plutôt attardés sur la régularité des TGV et sur l'affichage de trajets liés aux données que nous avions. Voici la liste des jeux de données récupérés :

— régularité mensuelle des TGV ;
— référentiel des gares de voyageurs géré par la SNCF ;
— liste des gares actuellement en service ;
— liste des circuits de voie.

2.3 Comment sont-elles traitées ?

Pour les traiter, nous avons réalisé des scripts en Python. En effet, certains jeux de données sont très complets, d'autres moins, certains ont beaucoup de données par rapport

1. `https://data.sncf.com/pages/accueil/`

à ce qu'on avait besoin, etc. Pour pallier ces problèmes, nous avons croisés des jeux de données afin d'obtenir un jeu de données d'une bonne qualité.

De plus, concernant les jeux avec trop de données, on peut citer en exemple celui montrant le réseau ferroviaire français. Au départ, il y avait environ 74 000 entrées, une correspondant à un point à afficher sur la carte de la France (que nous verrons dans la partie suivante). Grâce au script, nous avons filtrer les voies qui ne sont pas liées aux gares que nous possédions pour les données de retard et ainsi, nous avons réussi à réduire à environ 32 000 entrées. Cela a permis de ne pas stocker des données totalement inutiles pour notre application dans notre base de données mais aussi et surtout rendre l'affichage des trajets plus rapide et ainsi rendre la vue un peu plus interactive.

On peut noter que nous avons tous les trajets eu pour les TGV. Cependant, dans les gares mentionnées dans les trajets, des approximations sont réalisées et donc pour certaines gares, nous n'avons pas de points à afficher sur la première vue. Pour pouvoir les afficher, il faudrait réaliser des regroupements de gares locales mais au risque par exemple de regrouper les gares de Paris (Montparnasse, Lyon, Nord, etc) en une seule ce qui n'est pas très judicieux pour des visualisations.

2.4 Comment sont-elles visualisées ?

Pour visualiser les données, nous avons utilisé D3.js. C'est une bibliothèque JavaScript permettant de manipuler des documents basé sur des données. Il permet notamment de réaliser des graphiques interactifs, réaliser des systèmes de sélection, et plein d'autres choses qui sont de très bonne qualité.

Pour la visualisation, nous avons décidé de partir sur deux vues qui sont complètes au lieu d'un complète et deux très simples.

Première vue La première vue permet de sélectionner une gare d'arrivée et une gare de départ ainsi qu'une année et mois. Une fois cela sélectionné, plusieurs parties s'affichent à l'écran comme le montre la figure 1.

On y voit trois parties :

— un donut : permettant de représenter graphiquement les raisons des retards des TGV ;
— la carte de France : permettant de situer les gares sélectionnées ;
— les données contenues dans notre base : permet de connaître les chiffres en terme de retards, mais aussi connaître le nombre de TGV devant circuler sur ce trajet pour un mois ou année donnée.

Mais dans cette vue, il est possible de sélectionné qu'un champ et laisser les autres vides. Comme le montre la figure 2, on voit que seule l'année 2018 a été sélectionnée. Cela implique donc que nous connaîtrons les retards des TGV pour l'ensemble des trajets et mois durant l'année 2018, mais aussi le nombre de trajets réalisés (ici un total de plus de 400 000 trains).

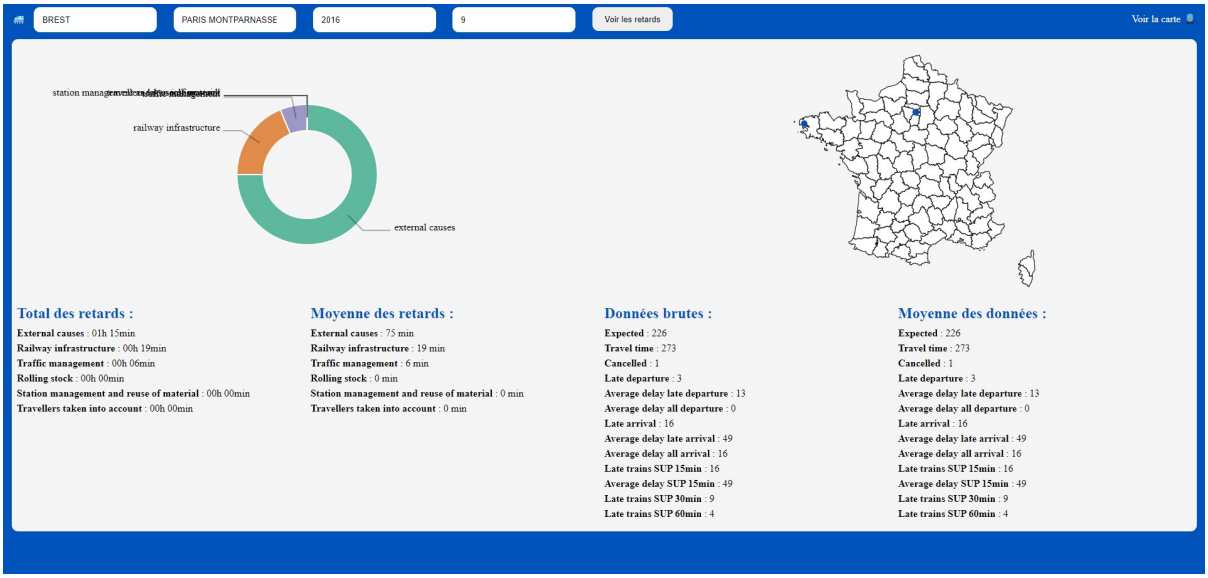

FIGURE 1 – Capture d'écran de la première vue

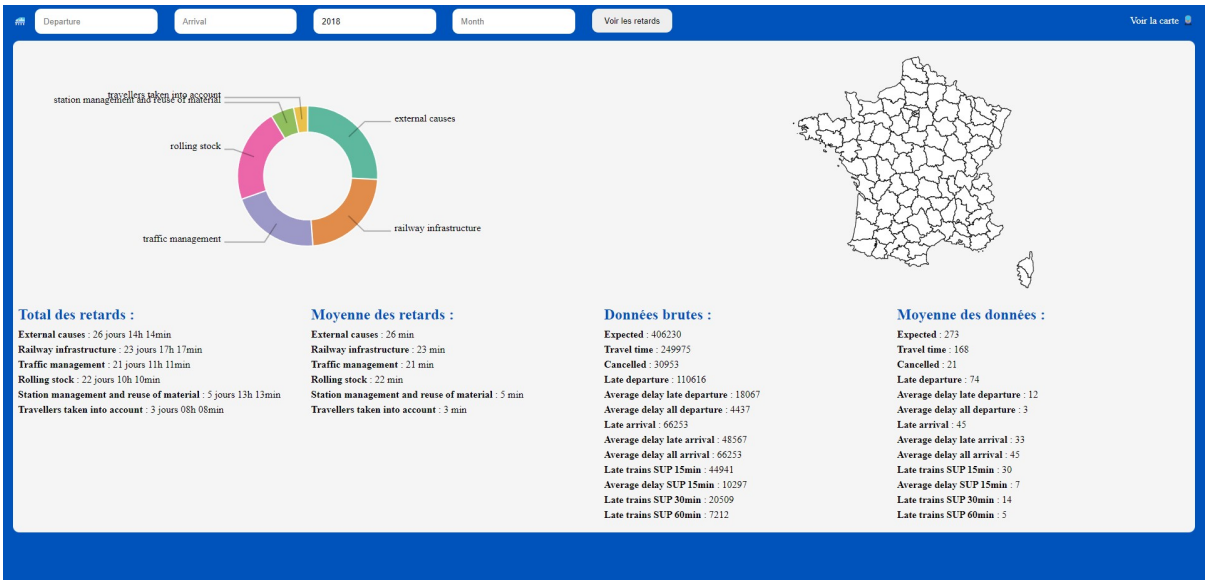

FIGURE 2 – Capture d'écran de la première vue avec qu'un seul champ rempli

Seconde vue Le seconde vue, comme dit plus tôt, présente la carte de France métropolitaine avec les gares et lignes desservies par TGV. La figure 3 montre la vue sans aucune sélection réalisée. Un petit guide d'utilisation est présent en nous incitant à survoler et à cliquer sur les éléments. Lorsque l'on survole une gare, le nom de la gare, la ville où elle se trouve et le département sont affichés dans un tooltip. Quant aux lignes, quand on en survole une, nous affichons le code de la ligne et désaffichons les autres lignes afin de mieux étudier le trajet de la ligne survolée.

Lors d'un clic sur une ligne, des informations sont affichés dans le cadre se trouvant

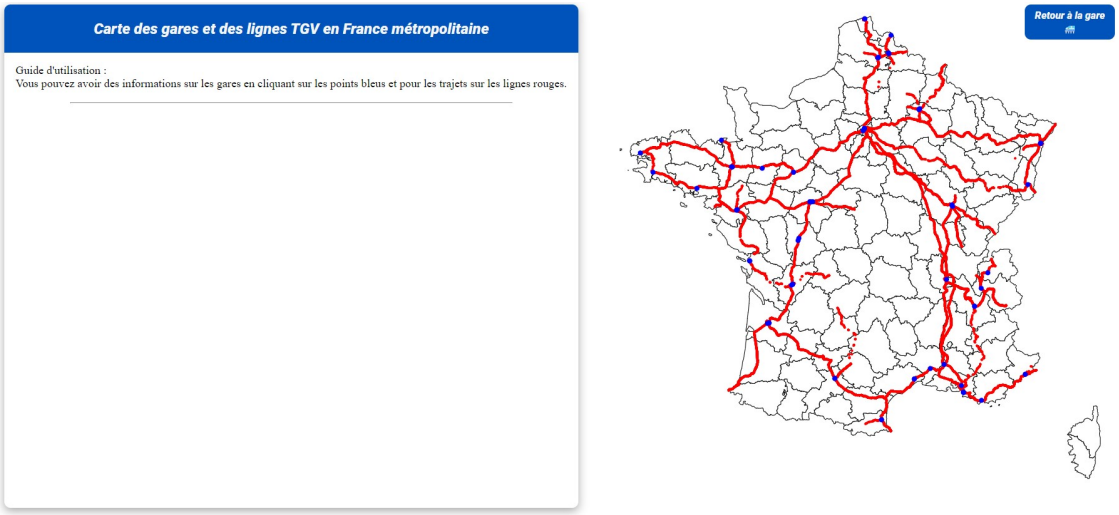

FIGURE 3 – Capture d'écran de la seconde vue

à gauche, comme le montre la figure 4. La ligne sélectionnée ici est celle reliant Brest à

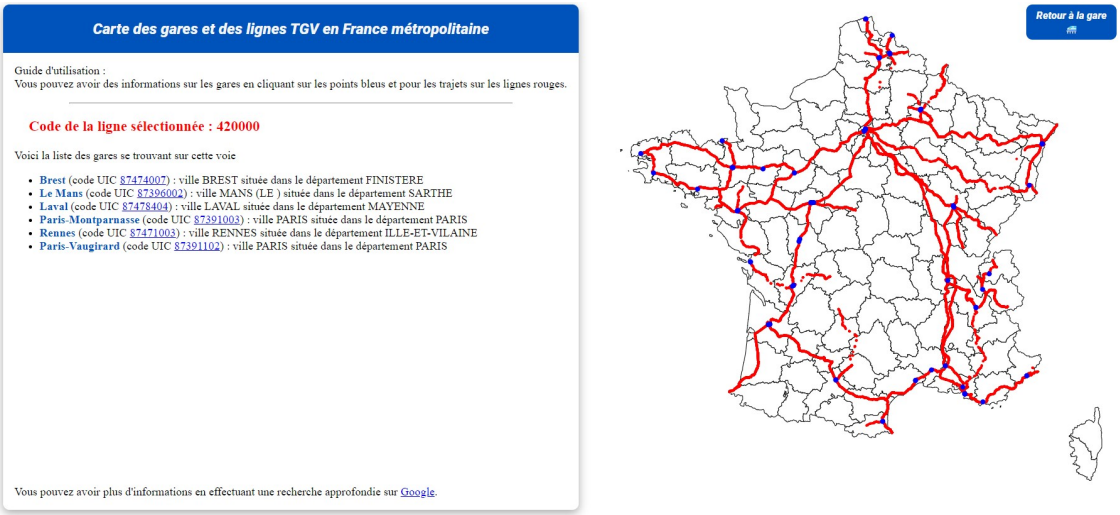

FIGURE 4 – Capture d'écran de la seconde vue après un clic sur la ligne Paris-Brest

Paris. On voit dans le cadre l'ensemble des gares principales traversées par cette ligne. Pour chaque gare, un lien externe permet de réaliser une recherche Google nous ramenant sur le site encyclopédique Wikipédia afin de pouvoir avoir de plus amples informations sur les gares ou sur la ligne (disponible en bas du cadre).

Concernant les clics sur les gares, nous affichons les mêmes informations mais inversé. C'est-à-dire que nous listons toutes les lignes passant par la gare sélectionnée, comme le

montre la figure 5 avec la sélection de la gare de Bordeaux. Comme pour les lignes, des

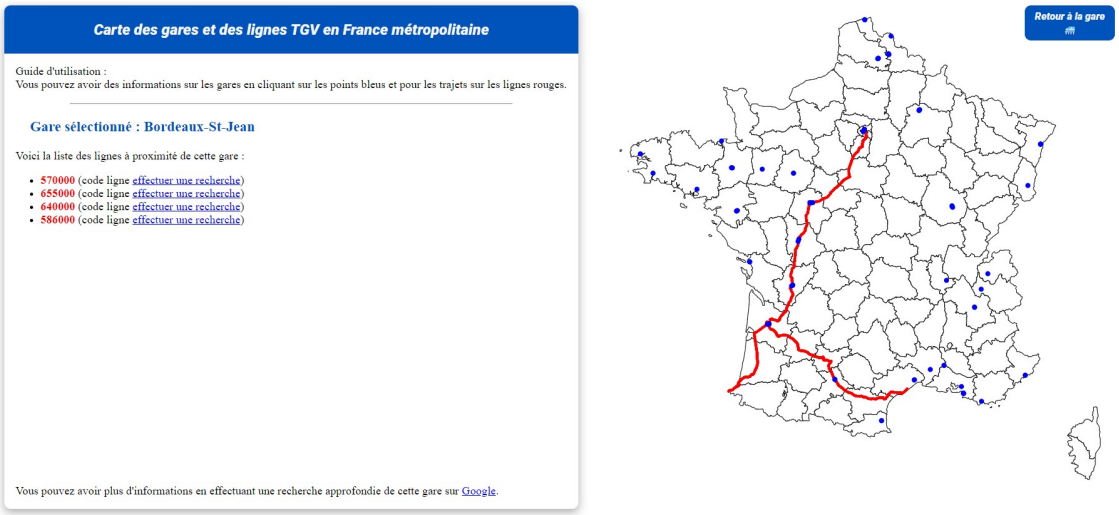

FIGURE 5 – Capture d'écran de la seconde vue après un clic sur la gare de Bordeaux

liens externes sont présents afin de réaliser des recherches. Cette partie est néanmoins différente. En effet, lorsque nous cliquons sur une gare, seules les lignes la traversant restent affichées et les autres disparaissent. Pour les retrouver, on peut recliquer sur la gare sélectionnée ou si nous voulons avoir les détails d'une autre gare, il est possible de cliquer sur une autre afin d'obtenir ses informations.

Intérêts de chaque vue Notre application a pour but de visualiser les retards sur les TGV gérés par la SNCF. Il est donc nécessaire d'afficher une visualisation simple en forme de donut afin de voir rapidement l'importance de chacune des causes, mais aussi voir les statistiques très simples afin de comprendre l'ampleur des retards subis par les usagers. D'autres statistiques permettent de montrer aussi la quantité de trains circulant sur le réseau TGV en France. La carte nous permet quant à elle de situer les gares en France car, personnellement, nous ne connaissons pas du tout le nom de l'ensemble des gares françaises !

La seconde vue a donc principalement été créée afin de mieux observer l'ensemble des lignes et gares TGV en France métropolitaine. En effet, de nombreuses voies existent, qu'elles soient destinées aux TGV, aux TER, au transport de fret ou encore des voies de dépôt et il faut pouvoir les visualiser simplement. Cette vue permet d'observer très rapidement ces données géographiques. Grâce à l'interactivité de cette vue, on peut en apprendre plus très facilement !