PROJECTIONS SUR CARTE DE DEPLACEMENTS MODELISES

Jean-Noël Zeh

Ipsos, France, jean-noel.zeh@ipsos.com

Résumé. Ce document décrit une partie de la modélisation des déplacements mise en place pour la mesure de l'audience extérieure en France. L'objectif est de générer par simulation les déplacements en France pour chaque individu de la population virtuelle et d'obtenir ainsi des flux de mobilité qualifiés pour l'ensemble du territoire.
Le modèle est basé sur plus de 600 000 déplacements captés avec un boîtier GPS auprès de 17 000 panélistes. Il est ensuite appliqué à la population virtuelle (extraction de 4 millions d'individus/lignes pour couvrir l'ensemble du territoire français sur une année complète).
Le document décrit également la constitution de la population virtuelle à partir du recensement INSEE, l'attribution des IRIS (zones géographiques) aux individus, la projection du modèle de déplacement sur la population virtuelle, et l'ajout de variables utiles au projet Mobimétrie.
Enfin, il explique comment les déplacements sont tracés sur une carte, comment les points d'intérêt sont classés en fonction de leur attractivité, et comment les flux de trafic sont exploités.

Mots-clés. Enquêtes de mobilité, Mesure d'audience, Données massives, Intégration de données de différentes sources, Dijkstra

Abstract. This document discusses the modeling of trips for outdoor audience measurement in France. The model is based on over 600,000 movements captured with a GPS+ device from 17,000 panelists. The model is then applied to a virtual population of 4 million individuals to cover the entire French territory over a year.
The virtual population is created from the INSEE census and is representative of the population studied. The model uses a "nearest neighbors" method to transfer mobility from the panel to the virtual population.
The document also describes the creation of a virtual population, the projection of the movement model on the virtual population, and the projection of the origin and destination of each virtual movement on a map.
The document further explains the use of the PageRank algorithm to rank Points of Interest (POIs) based on their attractiveness. The final step involves tracing each origin-destination route on a map.
The routing of all individuals must meet traffic constraints for each transport mode. The traffic values come directly from the framing data: PTV for vehicle flows, MyTraffic for pedestrian flows, and GTFS base for public transport.

Keywords. Mobility surveys, Audience measurement, Big data, Integration of data from different sources, Dijkstra

1 Introduction

Ce document décrit une partie de la modélisation des déplacements mise en place pour la mesure d'audience extérieure en France : le tracé sur carte.

L'objectif est de générer par simulation les déplacements en France pour chaque individu de la population virtuelle et d'obtenir ainsi des flux de mobilité qualifiés pour l'ensemble du territoire.

La base d'apprentissage du modèle est l'ensemble des 600 000+ déplacements captés avec un boitier GPS+ auprès de 17 000 panélistes.
Le modèle est ensuite appliqué à la population virtuelle (extraction de 4 millions d'individus/lignes pour couvrir l'ensemble du territoire français sur une année complète).
Il est calé par rapport à diverses statistiques de mobilité issues de CERAMA (distance moyenne parcourue par mode de transport, distribution modale des trajets domicile-travail, taux de non-déplacements…)

L'enregistrement de référence est le déplacement horodaté d'un individu (hors sol, sans coordonnées GPS). Le déplacement est caractérisé par :

- Une origine
- Un motif
- Un mode de transport
- Une durée
- Une distance
- Un horodatage, lui-même est caractérisé par :
    - Une heure (de départ)
    - Un type de période : semaine, samedi, dimanche, vacances zonées

L'individu est caractérisé par :

- Sexe
- Age
- CSP
- Revenus
- Constitution du foyer
- Taille d'agglomération du domicile

La mobilité est transférée du panel à la population virtuelle avec une méthode des « plus proches voisins ». L'algorithme recherche pour chaque individu virtuel, son/ses jumeau(x) panéliste(s) afin de lui associer ses déplacements.

2 Constitution de la population virtuelle

La population virtuelle est constituée à partir du recensement INSEE :

- Fichier des individus (FD_INDCVI_2019.txt, 20 millions de lignes)
- Fichier des ménages (FD_LOGEMT_2019.txt, 24 millions)

Il s'agit d'en extraire un échantillon de 4 millions d'individus/lignes (avec le poids INSEE moyen de 3 cela correspond à plus de 12 millions d'individus pondérés) qui soit représentatif de la population étudiée soit : les 11 ans et plus dans leur intégralité pour les agglomérations de plus de 10 000 habitants et uniquement ceux qui se déplacent régulièrement dans les agglomérations de plus de 10 000 habitants lorsqu'ils habitent dans des agglomérations de moins de 10 000 habitants (apports extérieurs). Le territoire est la France Métropolitaine hors Corse.

2.1 Première étape

Elle consiste à construire une base d'individus augmentée de l'information de commune/IRIS manquante pour les petites communes pour des raisons d'anonymisation. Cette information étant disponible au niveau des ménages, elle sera modélisée avant projection dans le fichier individus.

Le modèle d'attribution des IRIS au sein de la base individus est construit au niveau de chaque canton ville. La base des ménages pour laquelle tous les IRIS sont renseignés sert de base d'apprentissage (Machine Learning). Nous appliquons le modèle sur les individus représentant du ménage pour lesquels l'IRIS est manquant. L'IRIS de chaque représentant est ensuite appliqué aux autres membres de son ménage.

21 variables descriptives du représentant du Ménage (âge, diplôme, nombre de personne de son ménage, …) et du logement (la période d'achèvement, nombre de pièces) ont été utilisées.

2.2 Deuxième étape

Elle consiste en un tirage sur la base des individus enrichie par l'étape 1 avec la règle suivante :

- Agglomérations de plus de 100 000 habitants : elles regroupent 11 995 IRIS et 25 millions d'habitants de plus de 11 ans. 250 individus/lignes sont tirés au sort par IRIS (cahier des charges Mobimétrie).
- Agglomérations de moins 100 000 habitants : elles regroupent 36 436 IRIS et 29 millions d'habitants de plus de 11 ans. Le nombre d'individus/lignes tirés au sort par IRIS est logiquement contraint étant donné le nombre total d'individus à tirer.

Pour chaque IRIS appartenant à une agglomération de plus de 10 000 habitants, le tirage est fait en respectant le profil INSEE de l'IRIS pour la distribution (sexe et âge). Pour les autres, il doit être représentatif de la population extérieure (au moins un déplacement effectué par semaine dans une agglomération de plus de 10 000 habitants). Comme cette définition ne se traduit pas directement par une variable INSEE disponible, le tirage au sein de cette deuxième population se fait en respectant les distributions sociodémographiques (sexe et âge) de cette population issue du panel de l'enquête.

Le tirage se faisant au niveau IRIS et communes non irisées, il peut faire apparaitre des différences par rapport à la population totale à cause de la petite taille de la population.
Un redressement national est donc effectué (avec un RIM weighting très faible).

Après tirage, un coefficient de pondération final est appliqué pour être calé par rapport au dernier chiffre connu de la population des 11 ans et + France métropolitaine hors Corse, soit 57 756 000 individus.

2.3 Troisième étape

Adjonction à la population virtuelle des variables utiles au projet Mobimétrie.
Il y a 3 types de variables :
Variables directes sans retravail
    La majorité des variables (âge, sexe…)
    CESP chef de famille
Variables avec regroupement de modalités
    Regroupement d'IRIS pour stratification en taille d'agglomération / unité urbaine
    Structure familiale du ménage
Variables transformées
    Revenus

La variable « revenus » n'existant pas dans le fichier des individus, une modélisation a été effectuée à partir du fichier FiLoSoFi.

1. Un ranking des individus sur la base de la CSP et du niveau d'étude a été créé, ce dernier permet par la suite de ne pas attribuer des tranches de revenus élevées à des individus ayant un ranking bas ou l'inverse.
2. Les individus ayant le même code géographique ont été équirépartis en 10 groupes ordonnés en se basant sur le ranking établi précédemment.
3. Pour chaque groupe parmi les 10, un revenu sup/inf lui a été attribué en se basant sur le fichier FiLoSoFi (déciles).
4. Pour chaque individu virtuel, un revenu compris entre le revenu inf et le revenu sup de son groupe d'appartenance lui a été attribué.

3 Projection du modèle de déplacement sur la population virtuelle

3.1 Première étape

Elle consiste à appliquer le modèle de déplacements aux individus virtuels. Le résultat est une collection de vecteurs déplacements associés à chaque individu virtuel avec l'origine et la destination (POI), la durée, la distance, le mode de transport et l'horodatage. A ce stade, nous n'avons ni coordonnées (latitude, longitude) ni projection sur carte. Les déplacements modélisés sont associés par série temporelle à chaque individu afin de mesurer les répétitions de parcours. Pour chaque individu, l'ensemble des séquences de parcours sur une année est donc généré.

3.2 Seconde étape

Elle concerne la projection sur carte de l'origine de chaque déplacement virtuel.

L'origine du premier déplacement est en général le domicile de l'individu. Ce dernier est construit par tirage aléatoire (en latitude et en longitude) dans chaque IRIS. Il prend en compte la densité fine de population apportée le fichier INSEE « car_m » carroyée à 200 mètres. Les points sont ensuite placés sur les segments HERE 2019Q4.

3.3 Troisième étape

Elle consiste à associer à chaque POI un coefficient d'attractivité.

Description de l'algorithme
Il s'agit d'un ranking des POIs basé sur l'algorithme **PageRank**.
Cet algorithme a été développé par Google afin de mesurer la popularité d'une page web et indirectement personnaliser sa position au moment de l'affichage des résultats d'une recherche.
Dans notre cas les **pages web** sont assimilées à des **nœuds Here**, **les liens entre les pages** web à des **links Here**.
Pour chaque link here, un poids est attribué qui n'est rien d'autre que son flux.
Ci-dessous la formule utilisée pour calculer le score pour chaque nœud (POI dans notre cas) :

$$Coeff\ (POI\ cible) = \sum_{k=1}^{n} \frac{Coeff(POI\ k)}{Nombre\ de\ flux\ sortant\ (POI\ k)} \qquad (1)$$

n = *nombre de* POI *ayant un flux direct avec l*a POI *cible*

A partir des coefficients d'attractivité des POIs, nous en déduisons une probabilité qui sera utilisée lors du choix du POI à visiter.

$$\Pr(\text{POI } cible) = (1-d) + d * coeff(\text{POI } cible) \qquad (2)$$

$$d = facteur\ d'amortissement\ (\text{à } 0.85\ par\ défaut)$$

Scoring des POIs sur carte
Les scores pour tous les nœuds du graphe HERE France ont été calculés en se basant sur les flux synthétiques véhicules.
Ci-dessous un screenshot du résultat de l'algorithme pour Lille.

Figure 1

3.4 Quatrième étape

Elle concerne la projection sur carte de la destination de chaque déplacement virtuel pour les 3 classes de flux : véhicules, piétons & transport en commun.
A la fin de cette étape nous aurons donc une suite de déplacements avec origines et destinations placées sur carte sans le parcours.

Préparation des données
Cette étape consiste à réduire la taille moyenne des séquences de déplacements. Les séquences avec une étape intermédiaire par le domicile sont coupées en sous-séquences. Ainsi, si une séquence de 5 étapes [1,2,3,4,5] où l'étape 3 est le domicile, alors la séquence [1,2,3,4,5] est découpée en deux séquences [1,2,3] et [4,5]. De même une séquence de 9 étapes [1,2,3,4,5,6,7,8,9] où les étapes 2 et 5 sont des étapes au domicile, alors la séquence est découpée en trois séquences [1,2,3], [4,5], [6,7,8,9].
Le graphe ci-dessous présente la distribution des séquences par taille avant et après le coupage.

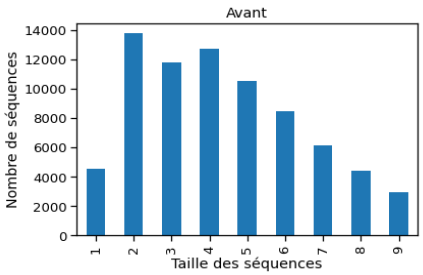

Figure 2

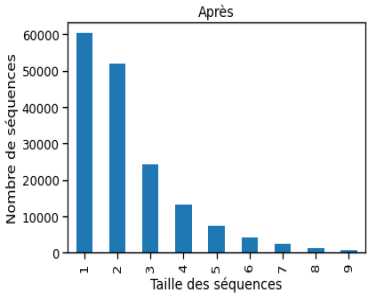

Figure 3

Algorithme

Pour permettre la recherche selon la distance géodésique, les points d'intérêt sont indexés dans des arbres. L'algorithme se déroule itérativement sur l'ensemble des séquences préparées. Pour chaque séquence :

1. Commencer par le nœud de départ et le considérer comme "nœud actuel"
2. Pour le nœud actuel et le déplacement suivant, rechercher les candidats avec un but correspondant et une distance similaire en utilisant les arbres construits précédemment
3. Filtrer la liste des candidats en enlevant ceux trop loin du domicile : somme des distances restantes inférieures à la distance directe entre le candidat et le domicile
4. Choisir n nœuds parmi les candidats pour continuer l'exploration. Le choix est fait en tirage au sort en respectant la probabilité de passage par ces nœuds selon le fichier des points d'intérêt.
5. Pour chaque point choisi P, P est considéré comme "nœud actuel" et les étapes 2-5 se répètent.
6. Lorsque le candidat est le dernier point de départ d'une séquence, vérifier si la distance de la dernière étape correspond à celle directe au domicile. Si oui, marquer la chaîne actuelle comme une solution.

La règle d'inégalité triangulaire (d(A,C) < d(A,B) + d(B,C)) est appliquée lors de la détermination des déplacements virtuels d'une boucle (hors promenade).

Illustration d'un déplacement D0, D1, D2, D0

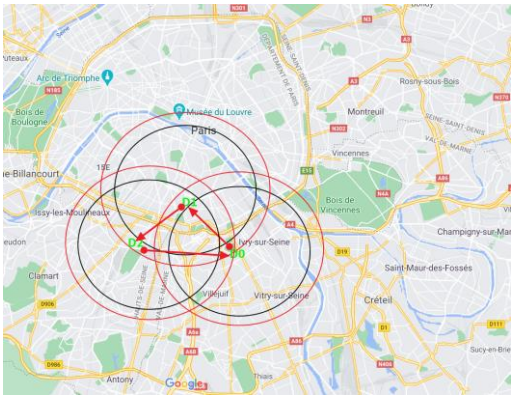

Figure 4

3.5 Cinquième étape

C'est le tracé final sur carte de chaque parcours origine-destination.

Pour chaque individu virtuel et pour chaque jour type décrit par une succession de O&D sur carte, un algorithme de recherche du **chemin optimal** sur **graphe orienté pondéré** (les links Here) est appliqué.
Le graphe est parcouru pour chaque couple (O,D) et chaque mode de transport.
La pondération (qui est la valeur à minimiser / c'est donc l'inverse d'un attracteur) de chaque link dépend du mode de transport.

- Pour les véhicules, la pondération dépend des attributs Here (vitesse et classe fonctionnelle) pour optimiser les temps de parcours.
- Pour les transports en public, les stations d'arrêt sont affectées d'un très faible poids pour jouer le rôle d'attracteur (sur la partie « piéton » du parcours).
- Pour les piétons, la pondération est simplement la distance.

Le graphe lui-même dépend du mode de transport :

- Pour les véhicules et les piétons, il s'agit directement des link Here filtrés sur l'attribut qui précise si les voies sont autorisées pour le mode de transport correspondant.
- Pour le transport en commun (hors métro), il s'agit des links associés aux réseaux de transport (provenant de l'Open data au format GTFS) et des links HERE «piétons» .

Chaque déplacement en transport en commun est complété par une partie « piéton » en début et fin de parcours pour se rendre à la station d'arrêt depuis le point de départ ou atteindre le POI depuis la station d'arrivée.

Le résultat pour chaque séquence journalière d'un individu virtuel est une série de **links Here** reliant les O&D entre eux.

C'est l'algorithme du chemin optimal **Dijkstra** (bibliothèque Python **networkx**) qui a été mis en place. Il permet de calculer rapidement le plus court chemin (minimisation des pondérations cumulées) entre un point de départ et un point d'arrivée.
Sa complexité est polynomiale. Pour n nodes et l links elle est de O(l+nlogn).
Il procède par élimination. Il choisit d'abord le sommet non visité avec la distance (poids) la plus faible. Il calcule ensuite la distance à travers lui à chaque voisin non visité et met à jour la distance du voisin si elle est plus petite.
Il marque le sommet visité lorsqu'il a effectué cette opération avec l'ensemble des voisins.

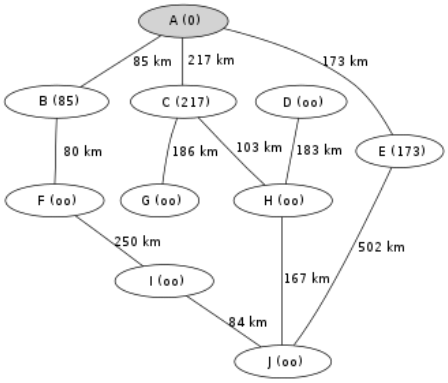

Figure 5

L'algorithme a été déployé sur le Google Cloud Computing.

<u>Exploitation des flux/trafic</u>

Le routing de l'ensemble des individus doit répondre à des contraintes de flux à atteindre par tronçon (link HERE) pour chaque mode de transport.
Les valeurs de trafic à atteindre proviennent directement des données de cadrage :

- PTV pour les flux véhicules
- MyTraffic pour les flux piétons
- Base GTFS pour les transports en commun

Ces valeurs de trafic sont disponibles pour un échantillon de links HERE pré-sectionnés pour être représentatifs de la mobilité.

Ces contraintes sont injectées par ajustement de la pondération des links dans l'algorithme **Dijkstra** de la manière suivante **:**

- Tirage aléatoire de blocs de 10 000 trajets à affecter sur voie.
- Révision de la pondération des links après chaque bloc affecté sur voie de la manière suivante :
    - Attractivité très forte des links ayant des objectifs de flux et dont le seuil n'est pas encore atteint
    - Après atteinte des seuils, l'attractivité est mise à une valeur très faible

Après routing, une vérification des flux/trafic issu du routing est effectué.
Il est a noté qu'à l'issu de ce processus, les flux TIM (Trafic Intensity Model) seront disponibles sur l'ensemble des links et l'ensemble des modes de transport (pour tracé des serpents de charge et exploitation au sein des outils de restitution).

Bibliographie

Partenaires académiques
Laboratoire LVMT (Laboratoire Ville Mobilité Transport), Olivier Bonin
Jean-Loup Madre, Professeur Emérite IFSTTAR/AME/DEST
Gilbert Saporta, professeur Emérite, titulaire de chaire de statistiques appliquées au Conservatoire national des arts et métiers
Laurent Vuillon, Professeur Université de Savoie, expert R&D en transfert de technologie académie vers les entreprises

Références bibliographiques
« Transport Survey Methods » dont l'un des auteurs est Jean-Loup Madre, correspondant scientifique d'Ipsos pour le projet.

« Introduction to Transportation Engineering » Tom V. Mathew and K V Krishna Rao
Chapter 31: « Fundamental relations of traffic flow »

« TRAFFIC STREAM CHARACTERISTICS » BY FRED L. HALL, Professor, McMaster University, Department of Civil Engineering and Department of Geography, 4 1280 Main Street West, Hamilton, Ontario, Canada L8S 4L7.

« RESEARCH ON TRAFFIC FLOW SPEED OF ARTERIAL STREETS IN URBAN AREAS »
Ziedonis Lazda1, Juris Smirnovs2
Faculty of Civil Engineering, Riga Technical University, 16, Azenes Street, LV 1048, Riga, Latvia.