



**HAL**  
open science

## Géocodage / calcul de temps de parcours pour les communes de la base “ mobilité professionnelles ” (INSEE)

Matthieu Viry, Timothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, Amel Feredj

### ► To cite this version:

Matthieu Viry, Timothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, et al.. Géocodage / calcul de temps de parcours pour les communes de la base “ mobilité professionnelles ” (INSEE). [Rapport de recherche] RIATE - Réseau interdisciplinaire pour l'Aménagement et la Cohésion des Territoires de l'Europe et de ses voisinages CNRS - CGET - Université Paris Diderot. 2016, 11 p. hal-03589037

**HAL Id: hal-03589037**

**<https://hal.science/hal-03589037>**

Submitted on 25 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

## 8. Géocodage / calcul de temps de parcours pour les communes de la base « mobilité professionnelles » (INSEE)

### RÉSUMÉ

Ce rapport technique décrit les méthodes utilisées pour géocoder les communes (françaises et étrangères) présentes dans le(s) fichier(s) « mobilités professionnelles » de l'INSEE et obtenir les temps de trajets automobiles correspondants aux flux qui y sont décrit.

Dans un premier temps ce rapport s'attache à décrire la méthode mise en œuvre pour géocoder les communes, notamment celles situées à l'étranger et dont le libellé dans la base INSEE peut parfois prêter à confusion (erreur d'orthographe, manque de précision, risque d'erreur en raison de la présence d'homonymes).

Les méthodes propres au calcul des temps de parcours et aux éventuels biais pouvant avoir été engendré sont détaillées dans une seconde partie.

### AUTEURS

Matthieu Viry, Timothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, Amel Feredj (UMS RIATE)

### EXPERTS

Sophie de Ruffray (UMR IDEES), Grégory Hamez (LOTERR)

**TYPOLOGIE SOCIO-ÉCONOMIQUE DES RÉGIONS FRONTALIÈRES DE L'UNION EUROPÉENNE (2000-2012)**

Étude commanditée par le Commissariat Général à l'Égalité des Territoires

## TABLE DES MATIÈRES

<b>1</b>	<b>Géocodage des communes présentes dans la base</b> .....	<b>2</b>
1.1	Objectifs et ressources disponibles .....	2
1.2	Géocodage des communes françaises .....	2
1.2.1	Les sources disponibles .....	2
1.2.1	Application à la base « mobilités professionnelles » .....	3
1.3	Géocodage des communes étrangères.....	4
1.3.1	Objectifs et difficultés à prendre en compte .....	4
1.3.2	Méthodes proposées.....	4
1.3.3	Résultats obtenus et mise en discussion .....	5
<b>2</b>	<b>Calcul des temps de parcours routiers</b> .....	<b>6</b>
2.1	Outils utilisés.....	6
2.2	Méthode proposée et résultats obtenus.....	6
<b>3</b>	<b>Discussion sur les résultats obtenus</b> .....	<b>8</b>
	<b>Annexe</b> .....	<b>9</b>

## TABLE DES ILLUSTRATIONS

Figure 1 - Flux supérieurs à 10 travailleurs (2012) .....	5
Figure 2 - Flux de travailleurs (France vers étranger ; volume > 5) et les parcours routiers correspondants .....	6

# 1 Géocodage des communes présentes dans la base

## 1.1 Objectifs et ressources disponibles

Le fichier INSEE récupéré ([http://www.insee.fr/fr/themes/detail.asp?reg\\_id=99&ref\\_id=mobilite-professionnelle-10](http://www.insee.fr/fr/themes/detail.asp?reg_id=99&ref_id=mobilite-professionnelle-10)) contient 1021610 entrées. Il s'agit d'une matrice origine/destination de villes affectées d'un flux de travailleurs. Les deux premières colonnes se rapportent à la commune de départ, les deux secondes à la commune d'arrivée et la dernière (NBFLUX\_C10\_ACTOCC15P) au nombre total d'actifs de 15 ans ou plus ayant un emploi et effectuant des parcours domicile-travail.

*« Les bases sur les flux de mobilité des « déplacements domicile-travail » fournissent, pour l'ensemble des communes (France métropolitaine et DOM), les effectifs correspondant aux croisements du lieu de résidence avec le lieu de travail. »*

Source : [http://www.insee.fr/fr/themes/detail.asp?reg\\_id=99&ref\\_id=mobilite-professionnelle-10](http://www.insee.fr/fr/themes/detail.asp?reg_id=99&ref_id=mobilite-professionnelle-10)

On peut y compter 36666 origines uniques (dont 112 situées dans les DOM) et 37440 destinations uniques (dont 134 situées dans les DOM ainsi que l'entrée ZZZZZ correspondant aux travailleurs dont la commune de destination n'est pas connue). On compte, parmi ces communes de destination, 1239 communes situées dans des pays voisins de la France (Allemagne, Belgique, Luxembourg, Monaco, Suisse et Andorre), représentant 22314 trajets à calculer.

Un second fichier relatif aux mobilités professionnelles a également été récupéré et concerne l'année 2012. Ce fichier contient quant à lui xxx communes étrangères nécessitant d'être géocodées.

Les lieux situés en France sont décrits par leur code INSEE. Les communes étrangères sont quant à elles décrites par un code (vraisemblablement interne à l'INSEE) étant lui-même associé à un libellé (parfois tronqué ou comportant des erreurs orthographiques, et dont les espaces ont été supprimé).

Différentes bases associent des noms de communes à leurs coordonnées (fond européen GEOSPECS, base GEONAMES, données OpenStreetMap, données disponibles sur les organismes statistiques des différents Etats, etc.). Différentes options sont utilisables lorsqu'il s'agit de consulter ces bases et les résultats obtenues peuvent dépendre en partie de choix effectués lors du passage de la requête et/ou d'une sélection effectuée sur les résultats du géocodage (permettant éventuellement de trancher entre des homonymes, ou de baser ce choix sur des éléments rationnels du jeu de données).

## 1.2 Géocodage des communes françaises

### 1.2.1 Les sources disponibles

Plusieurs sources fiables, notamment issues du rapprochement des données d'acteurs nationaux tels que la Poste, l'IGN ainsi que des projets OpenStreetMap. Ce projet, nommé « Base Adresse Nationale » (BAN ou parfois BANO pour *Base Adresse Nationale Ouverte* en raison de sa distribution dans le cadre d'une licence ODbL).

Cette base peut être consultée via une API publique. Ainsi une requête de la forme

<http://api-adresse.data.gouv.fr/search/?q=cityode=01100>

permettra d'obtenir les coordonnées d'une adresse située dans la commune dont le code INSEE est 01100.

Cette méthode n'est pas parfaite puisqu'elle ne permet, à moins de fournir une adresse plus précise, d'obtenir que des coordonnées correspondantes à une adresse indéterminée de la commune.

Il est également possible de récupérer des extraits de cette base de différentes façons. Soit par le téléchargement direct des données *OpenStreetMap* : les communes française (des relations OSM décrites par les tags *administrative = boundary* et *admin\_level = 8*). L'ensemble des communes françaises actuelles y est répertorié. Le code INSEE étant renseigné pour chacune d'entre elles dans le champ « *ref\_insee* ». Cette méthode peut présenter plusieurs avantages. En effet pour la majorité de ces entités OSM correspondant un membre dénommé *admin\_centre* (et correspondant généralement à la mairie) : c'est les coordonnées de ce membre qui sont à récupérer. Quelques communes ne présentaient pas ce membre lors de l'extraction réalisées, c'est alors le centroïde du polygone des limites communales qui a été utilisé. Cette méthode présente aussi l'avantage de pouvoir récupérer des informations sur d'anciennes communes : elles sont en effet parfois répertoriées dans *OpenStreetMap* (cette fois avec le tag *admin\_level = 10*) et le code INSEE est présent dans le champ « *old\_ref\_insee* ». Des informations sur les dates des évènements (création/fusion/disparition) sont également parfois fournies dans le champ « *source\_name* ».

Il est également possible de récupérer le même type d'information via une extraction disponible sur la *plateforme ouverte des données publiques françaises* : <https://www.data.gouv.fr/fr/datasets/correspondance-entre-les-codes-postaux-et-codes-insee-des-communes-francaises/> (cette extraction n'est pas mise à jour de manière rég)

Cette base ne concerne que les communes existantes actuellement. Il est possible de la compléter (pour les communes concernées par une fusion/disparition) par la base disponible sur le site de l'INSEE qui présente l'historique des modifications des communes françaises depuis 1943 (<http://www.insee.fr/fr/methodes/nomenclatures/cog/telechargement.asp>).

Quelle que soit la méthode retenue, le but est simplement d'obtenir une paire de coordonnées associée à chaque code commune afin d'effectuer une jointure entre cette table et les communes françaises de la base « *mobilité professionnelle* ».

### 1.2.1 Application à la base « *mobilités professionnelles* »

Cette base recense des départs de travailleurs depuis 36666 communes françaises (voir illustration en Annexe 1). La jointure entre ces codes communes INSEE et une base contenant des coordonnées géographiques a été effectuée sur la table des communes françaises d'une extraction *OpenStreetMap*, précédemment réalisée, prenant en compte ces critères (présence des champs « *ref\_insee* » et « *old\_ref\_insee* » notamment).

Quelques vérifications simples permettent de vérifier la cohérence du géocodage ainsi réalisé.

Il est par exemple possible de récupérer les polygones correspondants aux départements français et de vérifier via une requête spatiale que les entités se trouvent effectivement dans leur département d'appartenance. Il est également possible de vérifier qu'il n'existe pas de doublon dans les coordonnées utilisées (indiquant que deux communes seraient géocodés au même endroit).

Ces vérifications semblent confirmer la véracité du géocodage ainsi effectué ; la recherche de doublons parmi les coordonnées utilisées mettant seulement en avant les communes d'Auxon-Dessous et d'Auxon-Dessus dont les coordonnées utilisées sont les mêmes (en raison de leur fusion<sup>1</sup> sous le nom d'Auxon au 1<sup>er</sup> janvier 2015).

---

1

<http://archive.wikiwix.com/cache/?url=http%3A%2F%2Fwww.doubs.territorial.gouv.fr%2Ffactes3%2Ffiles%2Ffichiera%2Fcte56392.pdf>

## 1.3 Géocodage des communes étrangères

### 1.3.1 Objectifs et difficultés à prendre en compte

Le champ contenant le libellé de la commune de destination présente quelques particularités. En effet quelques-uns de ces libellés présentent une information entre parenthèse correspondant au canton d'appartenance de la commune : il conviendra de transmettre cette information de manière correcte à l'outil de géocodage. Ce champ est parfois tronqué (par exemple : « SAINT SULPICE (NEUCH » ou présent sous sa forme abrégée (VD pour désigner le canton de VAUD) : il conviendra ainsi d'harmoniser ces informations avant d'utiliser l'outil de géocodage.

Différentes erreurs ont également été corrigées manuellement avant de soumettre les noms de lieux à l'outil de géocodage. Plusieurs de ces erreurs, qui semblent issues d'erreur de numérisation/de saisie, sont regroupées dans une table (voir Annexe 2).

### 1.3.2 Méthodes proposées

Une fois cette étape réalisée il est possible (via un script python) de soumettre chacune de ces communes à un outil interrogeant la base *OpenStreetMap*. C'est ici l'outil *Nominatim* qui a été retenue. Ainsi chaque nom de commune étrangère de la base est transformé en une requête vers cet outil (déployé en local pour l'occasion).

L'entrée SU42Z | ECUBLENS (FRIBOURG) est ainsi transformée en une requête de la forme :

<http://nominatim.openstreetmap.org/search?q=Ecublens,%20Fribourg,%20Suisse&format=json>

ou de la forme

<http://nominatim.openstreetmap.org/search?city=Ecublens&county=Fribourg&country=Suisse&format=json>

(La seconde requête précise explicitement à l'outil quel mot correspond au nom de la ville, de la province et du pays ; on note que le résultat est le même dans les nombreux cas ne présentant pas d'ambiguïté).

L'outil de géocodage *Nominatim* peut renvoyer plusieurs résultats. Ces résultats sont alors classés par ordre de pertinence (par rapport au contenu de la requête) et selon l'importance des localités retournées. Il s'agit de ce premier résultat qui a été retenu.

Si les résultats obtenus par cette méthode de géocodage semblent correctes (hormis quelques exceptions, des coordonnées (sans doublon) peuvent être obtenues pour l'ensemble de ces communes). L'affichage sur une carte des communes ainsi géocodées ainsi que la mise en relation de la localisation de ces communes avec les communes émettrices des flux met toutefois en évidence de très probables erreurs : de nombreuses communes étrangères (notamment plusieurs recevant des flux très importants en provenance de communes françaises frontalières) sont situées à plusieurs centaines de kilomètres de la frontière française. Une vérification manuelle de quelques communes permet de mettre en évidence la présence d'homonymes, des communes généralement moins importantes, situées plus proche de la frontière française, et n'apparaissant pas en tête des résultats de l'outil de géocodage. Ce biais nécessite d'adapter la méthode de géocodage pour prendre en compte la distance à la frontière française et/ou la distance aux communes émettrices.

Ainsi un script python se charge de parcourir le fichier contenant les couples de lieux et relève, pour chaque destination, les coordonnées des communes émettrices d'un flux ainsi que l'importance de ce flux. Ces données sont utilisées, pour chaque commune destinataire, pour calculer la localisation du barycentre pondéré par le volume de travailleur des communes émettrices. Cette coordonnées pourra ainsi être utilisée pour calculer une distance avec chacun des résultats de l'outil de géocodage, permettant par exemple de sélectionner le plus proche (ou un compromis entre la proximité géographique et la pertinence du résultat) en cas d'homonymie.

→ Script `geocodage_depuis_mobpro.py`

### 1.3.3 Résultats obtenus et mise en discussion

Les résultats obtenus après plusieurs essais semblent satisfaisants : la majorité des communes géocodées s'avère à proximité (ou au moins d'une manière proportionnelle aux flux reçus) de la frontière française et les exceptions semblent être justifiées (car correspondant à des flux très faibles ou à des communes d'importance, telles que Berlin ou Genève).

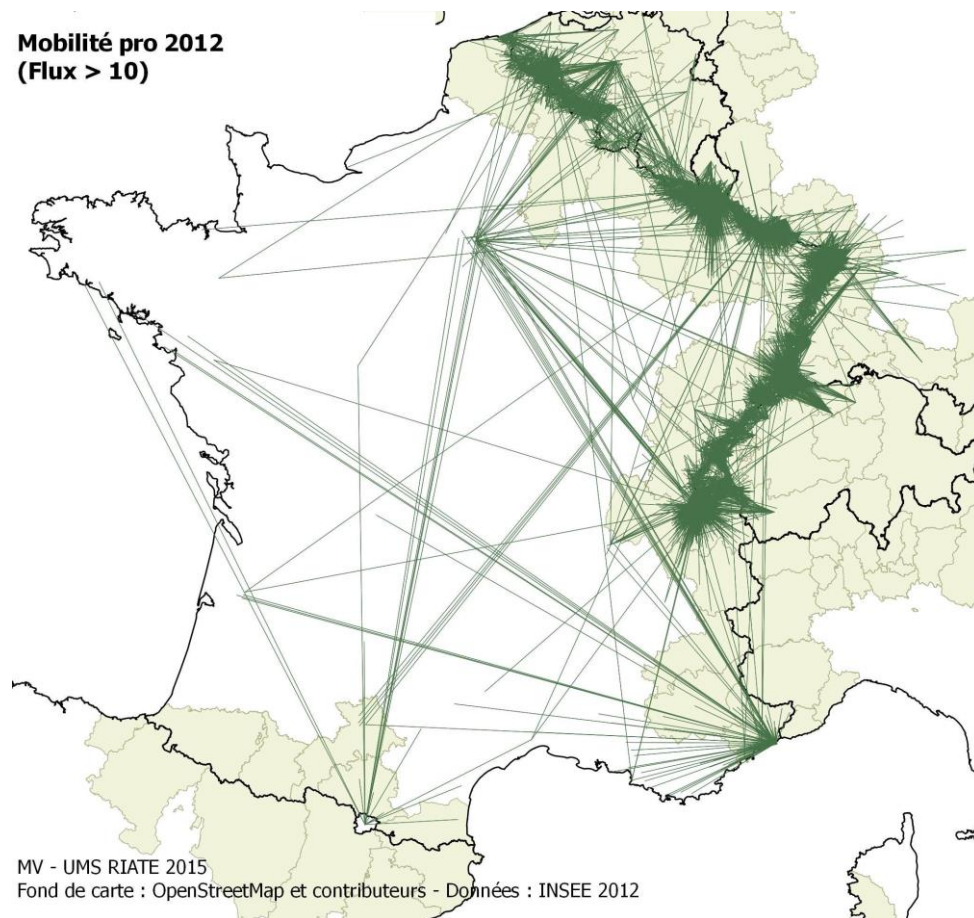


Figure 1 - Flux supérieurs à 10 travailleurs (2012)

Cette méthode présente plusieurs défauts mais également des avantages :

- De nombreux essais ont été nécessaires, notamment pour étalonner la méthode
- La méthode repose sur une ressource déployée localement (= temps d'installation + espace disques nécessaires) ou disponible en ligne (nécessitant ainsi une connexion internet et le respect des règles d'utilisation de ces API).
- Le temps nécessaire au géocodage est théoriquement assez court.
- La méthode est facilement adaptable / reproductible (il est par exemple facile de la remettre en œuvre si de nouvelles communes étrangères sont ajoutées à ce fichier et/ou si les informations présentes dans la base *OpenStreetMap* sont susceptibles de s'être améliorées).
- Les résultats fournis dépendent à la fois de la qualité de la requête (orthographe, champ utilisés, etc.) et des données présentes dans la base de données utilisée. Des erreurs/omissions peuvent ainsi exister à plusieurs étapes du processus et la vérification de la localisation de certaines communes par un œil avisé peut être nécessaire.
- La localisation de certaines communes (disparues depuis leur inscription dans la base « mobilités professionnelles ») est possible, notamment grâce à la prise en compte des noms de quartier/d'arrondissement dans l'outil de géocodage

## 2 Calcul des temps de parcours routiers

### 2.1 Outils utilisés

Les temps de parcours sont calculés par l'outil **OSRM**, sur la base d'une extraction OpenStreetMap de juillet 2015. Ces temps de parcours peuvent être récupérés via la fonction *table* qu'il propose. Toutefois, en raison du souhait de récupérer également la distance routière correspondante et la géométrie de ces routes, c'est chacun des parcours « France vers étranger » des fichiers « mobilités professionnelles » qui a été calculé (utilisant la fonction *viaroute* d'OSRM).

### 2.2 Méthode proposée et résultats obtenus

Afin d'automatiser ce processus et de le rendre facilement reproductible, un *plug-in* pour QGIS a été réalisé et permet d'obtenir, à partir d'un fichier décrivant les couples d'origines/destination ou partir de deux couches de points, l'ensemble des parcours dans un fichier shapefile. Ce plug-in permet également de récupérer une matrice de temps de parcours à partir d'une couche de points décrivant les différentes localisations à utiliser.

→ *Plugin* QGIS **OSRM**

La table présente en sortie reprend les champs de la table utilisée en entrée, auxquels sont ajoutés le temps de parcours en minutes, la distance routière en kilomètre et la distance euclidienne en kilomètres (Tableau 1).

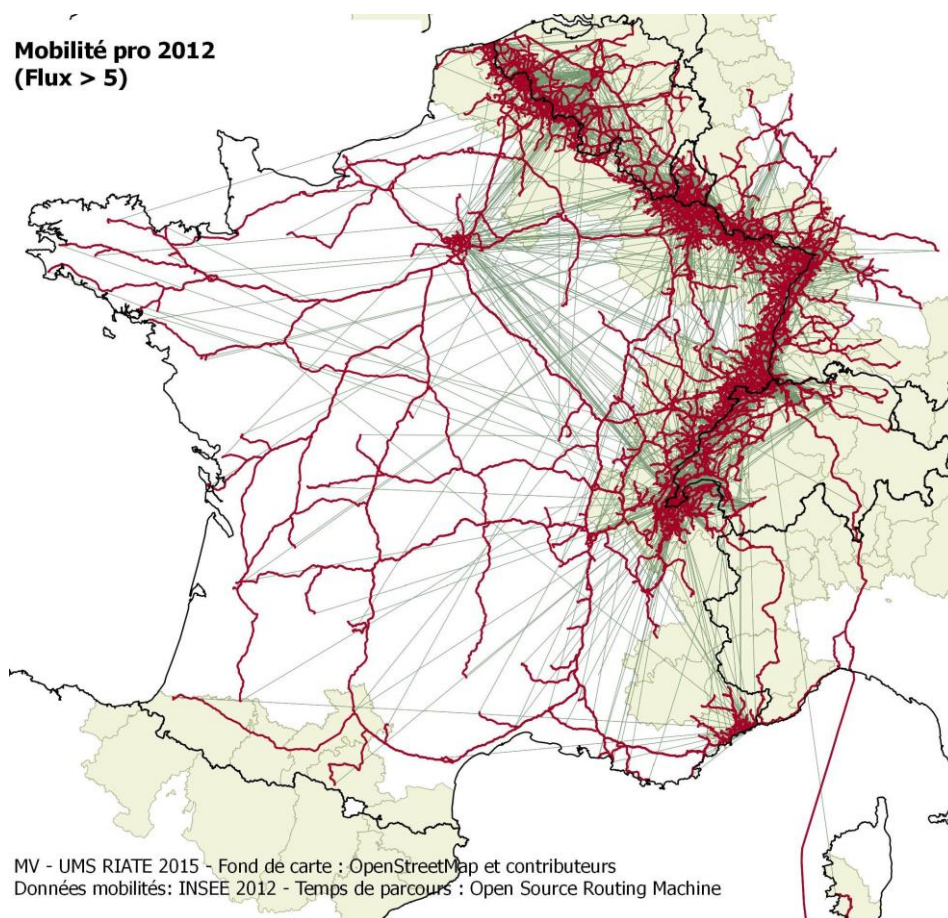


Figure 2 - Flux de travailleurs (France vers étranger ; volume > 5) et les parcours routiers correspondants



Une seconde table est générée, notamment pour vérifier qu'aucun parcours ne débute ou ne finisse loin des origines/destinations demandées. Elle comporte pour chaque parcours, la distance entre les points soumis à OSRM (origine d'une part et destination d'autre part) et les points utilisés comme points de départ et d'arrivée pour le parcours calculé (voir Tableau 1, les deux dernières colonnes). Si ces distances restent faibles, cette vérification peut-être utile pour mettre en avant des erreurs/imprécisions dans le géocodage (en effet, ce dernier étant effectué sur des communes, la distance pour s'accrocher au réseau ne devrait jamais dépasser quelques mètres ou centaines de mètres).

Origine	x_origine	y_origine	Destination	x_dest	y_dest	flux	Temps (minutes)	Distance euclidienne (km)	Distance routière (km)	Distance de déplacement du point d'origine	Distance de déplacement du point d'arrivée
1004	5,37799 g	45,96133	MO001	7,453772	43,72642	5,375	316.47	297.66	548.44	0.37	2.09
1004	5,37799 g	45,96133	SU96F	6,642894	46,5263	1	109.93	116.02	169.40	0.37	0.02
1004	5,37799 g	45,96133	SUC1L	6,142967	46,20947	2,049	70.27	65.28	105.55	0.37	0.0
1004	5,37799 g	45,96133	SUC1X	6,120111	46,16489	1,173	65.69	61.72	99.78	0.37	0.03

Tableau 1 - Extrait de la table de résultats obtenue.

### 3 Discussion sur les résultats obtenus

Les résultats obtenus, aussi bien en termes de géocodage qu'en termes de temps de parcours sont sujets à différents biais.

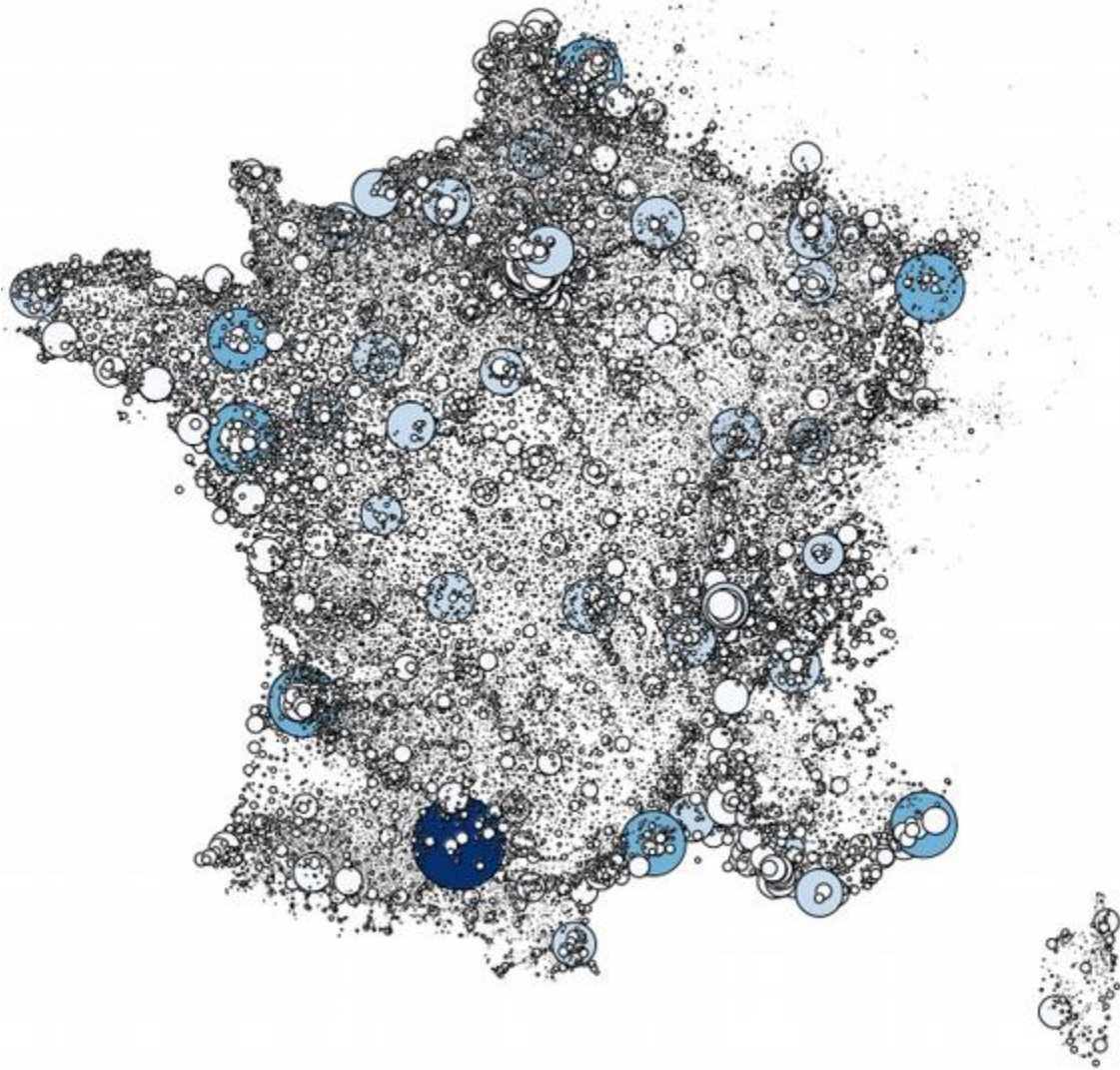
D'une part la localisation d'une partie des communes ne peut être confirmée, faute d'information supplémentaire. Cet élément pourrait notamment être résolu par l'utilisation dans la base « mobilité professionnelle » d'un identifiant unique européen (tel que le code LAU2), permettant de joindre les informations de la base avec des informations spatiales (découpage LAU2 ou centroïde des LAU2).

D'autre part une seule mesure de temps de parcours est calculée, il s'agit d'un temps de parcours routier théorique : il ne prend pas en compte les conditions de circulation (congestion, déviation temporaire, conditions météo, etc.) ni l'existence d'autres modes de transport, jouant potentiellement un rôle important dans le cadre de ces mobilités. L'interprétation des flux présents dans la base « mobilités professionnelles » ne pourrait être complète qu'avec des données permettant des calculs de temps de parcours multi-modaux et prenant en compte les heures de la journée auxquelles s'effectuent ces déplacements.

Dans le cadre d'échange sur les mobilités professionnelles avec le CGET, les résultats du géocodage, réalisé avec la méthode décrite ci-dessus, ont été comparés avec le travail de géocodage (basé sur le fond européen *GEOSPECS*) réalisé par l'Agence de Développement et d'Urbanisme de l'Aire Urbaine Nancéienne (ADUAN) dont l'un des objectifs était de représenter le volume de travailleurs français travaillant à l'étranger dans la région Grand-Est. Cette comparaison a permis de mettre en évidence des erreurs dans les deux méthodes et confirme la nécessité, quel que soit la méthode employée, de vérifier manuellement une partie des résultats du géocodage, d'autant plus s'il est possible de bénéficier de l'expérience de quelqu'un connaissant le territoire.

Enfin on constate que la possibilité d'obtenir l'ensemble des parcours correspondants aux origines-distances ouvre les portes à d'autres analyses, notamment en termes de **modélisation de la congestion** ainsi que pour calculer des **bassin de recrutement de travailleurs** autour des communes recevant des travailleurs, basée sur un paramètre fonctionnel comme la distance temps.

## Annexe



Annexe 1 - Volume de travailleurs reçu par commune  
(Source : Mobilités professionnelles INSEE 2011)

Nom d'origine	Nom corrigé
<b>Honelles</b>	Honnelles
<b>Flobeco</b>	Flobecq
<b>Muenstertauschwarz</b>	Munstertal Schwarzwald
<b>Iwaregem</b>	Waregem
<b>Ilandsthul</b>	Landsthul
<b>Iffezheim</b>	Iffezheim
<b>Lasagne</b>	La Sagne
<b>Ztefen</b>	Ziefen
<b>Peneylejorat</b>	Peney-Le-Jorat

Annexe 2 - Exemples d'erreurs orthographiques concernant les noms de communes étrangères





# CE QU'IL FAUT RETENIR...

- **La méthode de géocodage proposée n'est pas infaillible.** En effet, en cas d'homonymie et faute d'indication supplémentaire, et d'autant plus lorsque le volume des flux est faible, il est difficile de savoir quelle est réellement la commune citée par l'enquêté.
- **La validation de l'étape de géocodage peut nécessiter une intervention humaine,** notamment pour trancher des cas d'homonymie ou lors de reconstitution communale dans des pays étrangers qui pourraient être plus facilement repérables par un opérateur connaissant le terrain d'étude.
- L'étape de géocodage peut avoir un impact sur les temps de parcours calculés : une localisation trop imprécise de certaines communes pouvant par exemple engendrer des parcours utilisant des points de départ/arrivée situés sur des branches peu importantes du réseau.
- Contrairement à d'autres outils (tels que la base *Metric*® pour la France, ou l'utilisation de l'API Google®) qui proposent des temps de parcours pouvant prendre en compte la congestion (par exemple avec une mesure « heure pleine » et une mesure « heure creuse »), seul **un temps de parcours théorique** est ici fourni (ce dernier pouvant être rapproché d'une mesure dite « en heure creuse »). De plus ces calculs ne prennent en compte qu'un seul mode de transport, occultant ainsi l'offre de transports en commun disponible et les possibilités qu'elle offre en termes de mobilités professionnelles.