Modal INF473G
Déserts médicaux et prévalence de maladies

Jérémie Touati et Paul Vanborre

1 Introduction

Notre projet porte sur le thème de santé et plus particulièrement sur l'accessibilité aux soins dans l'ensemble du territoire français. Le but est de comprendre si les déserts médicaux ont une influence sur la santé des gens qui y habitent et une incidence sur leur probabilité d'être malade.

2 Nos sources de données

Pour traiter ce sujet, nous avons utilisé deux jeux de données issus du site *data.gouv.fr* :

- Le premier liste les établissements de santé français et donne pour chacun un certain nombre d'informations dont sa localisation (adresse) et sa nature (est-ce un centre hospitalier ou une pharmacie... ?). Il est téléchargeable au lien suivant : *https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/#/resources*

- Le second contient les effectifs de patients en France par pathologie, par catégorie d'âge, par sexe et par département, tout en donnant les prévalences associées. Il est téléchargeable ici : *https://www.data.gouv.fr/fr/datasets/effectif-de-patients-par-pathologie-sexe-classe-dage-et-territoire-departement-region/*

Comme nous le verrons dans la suite de ce rapport, nous avons choisi de mener notre étude à l'échelle de la ville et non du département, alors que le second jeu de données portait sur les départements. Pour pouvoir réaliser ce changement d'échelle, nous avons fait appel à la base de données Wikidata afin d'obtenir des informations sur les communes de chaque département.

3 Création des nœuds du graphe

Notre graphe contient deux types de nœuds : des nœuds de type *établissement* et des nœuds de type *malade*. Ceux-ci sont reliés par une arête pondérée par la distance entre l'habitation du malade et le lieu de l'établissement de santé, calculée par l'API d'OpenStreetMap. Pour représenter nos données, nous avons opté pour la création d'un graphe de type Data Graph à l'aide de Neo4j. En effet, un tel graphe était plus adapté qu'un Label Graph car simplement des labels sur les nœuds n'étaient pas suffisants. En effet, nous souhaitions ajouter à nos nœuds représentant les malades le nombre d'habitants

de la commune, le nom et code postal de la commune, la prévalence de la maladie dans cette commune... Pour les établissements, nous voulions également ajouter leur adresse et leur code postal.

Etant donné que notre premier jeu de données contenait près de 100 000 lignes et le deuxième en contenait plus de 3 millions, cela aurait signifié de créer un graphe d'autant de nœuds et de $3 \times 10^6 \cdot 10^5 = 3 \times 10^{11}$ arêtes. Ces calculs ne prennent d'ailleurs même pas en compte notre changement d'échelle : le passage du département à la ville multiplie en fait ces nombres. Neo4j ne peut pas supporter un graphe d'une telle taille et nous avons donc dû procéder à un certain nombre de simplifications dans nos jeux de données. Ces simplifications ont constitué une étape importante de pre-processing.

3.1 Nœuds de type *établissement*

Pour créer ces nœuds, nous nous sommes basés sur le fichier listant les établissements de santé français (environ 100 000 lignes). Celui-ci mélangeait des établissements de type très variés (hôpitaux, pharmacies, centres d'accueil pour personnes âgées ou pour enfants, services de soins spécialisés...). Afin de réduire le nombre de lignes dans ce fichier et donc de nœuds et d'arêtes dans notre graphe, nous nous sommes restreints aux établissements de type "Centres Hospitaliers" et "Hôpitaux Locaux" que nous avons considérés comme étant les plus représentatifs pour estimer la couverture de santé d'un département. Cette étape a permis de réduire le fichier à un total de 1474 lignes.

Pour pouvoir effectuer les calculs de distance avec l'API d'OpenStreetMap, nous avions besoin en particulier du nom de la ville de chaque établissement. Le fichier contenait cette information, mais la ville était parfois suivie de "CEDEX", de "CEDEX 9"... Nous avons pour cela enlevé les occurrences de "CEDEX" et les chiffres qui étaient présents dans les noms. Le nom de ville ainsi "nettoyé" a été ajouté comme nouvel attribut à notre fichier et donc à nos nœuds *établissement*.

3.2 Nœuds de type *malade*

Afin de dimininuer considérablement la taille du fichier des pathologies, nous nous sommes restreints aux lignes qui contenaient des informations relatives à l'année 2020, année la plus récente du fichier. Nous n'avons considéré qu'une seule pathologie parmi celles présentes : le cancer du poumon, et n'avons considéré que les hommes, qui sont les plus touchés par cette maladie. Nous nous sommes également restreints à la classe d'âge des 60-64 ans, ce que nous justifions par le fait que cette tranche d'âge est parmi les plus touchées par les cancers du poumon. Nous obtenons ainsi, pour chaque département, un nœud indiquant la prévalence du cancer du poumon chez les hommes de 60-64 ans en 2020.

Pour passer de l'échelle du département à l'échelle de la commune, nous avons réalisé une query Wikidata donnant toutes les villes d'un département et leur nombre d'habitants. Pour éviter de considérer un nombre trop important de petites villes, nous nous sommes restreints aux villes de plus de 2500 habitants. Le site Association des Maires de France nous a permis de choisir ce chiffre judicieusement. La query Wikidata est présentée figure 1.

```
SELECT ?city ?cityLabel ?population
  WHERE {
    ?city wdt:P31 wd:Q484170 . #est une commune française
    ?city wdt:P1082 ?population. #on récupère sa population
    ?city wdt:P131 wd:Q{departement_code}. #commune du département untel
    FILTER(?population > 2500) #on se limite aux communes de population > 2500 :
                               #justifié car autour de chaque petit village on a forcément
                               #à quelques kilomètres une commune de plus de 2500 habitants
    SERVICE wikibase:label {{ bd:serviceParam wikibase:language "fr". }}
  }
```

Figure 1: Query Wikidata renvoyant les villes d'un département donné et leur nombre d'habitants

Nous avons donc remplacé chaque nœud (correspondant à un département) par n_{dep} nœuds, où n_{dep} est le nombre de villes de plus de 2500 habitants dans le département en question. Nous avons fait l'approximation que la prévalence de la maladie pour dans la commune était la même que celle dans le département (voir figure 2). Nous avons ainsi obtenu un total de 4485 nœuds de type *malade*.

3.3 Importation des nœuds sur Neo4j

En utilisant les queries Cypher suivantes, nous avons chargé tous les nœuds représentant les établissements puis ceux représentant les malades :

```
LOAD CSV WITH HEADERS FROM "file:///etablissements_structure.csv"
AS row FIELDTERMINATOR ";"
CREATE (n:Etablissement)
SET n=row

LOAD CSV WITH HEADERS FROM "file:///nouveaux_malades_poumon.csv"
AS row FIELDTERMINATOR ";"
CREATE (n:Malade)
SET n=row
```

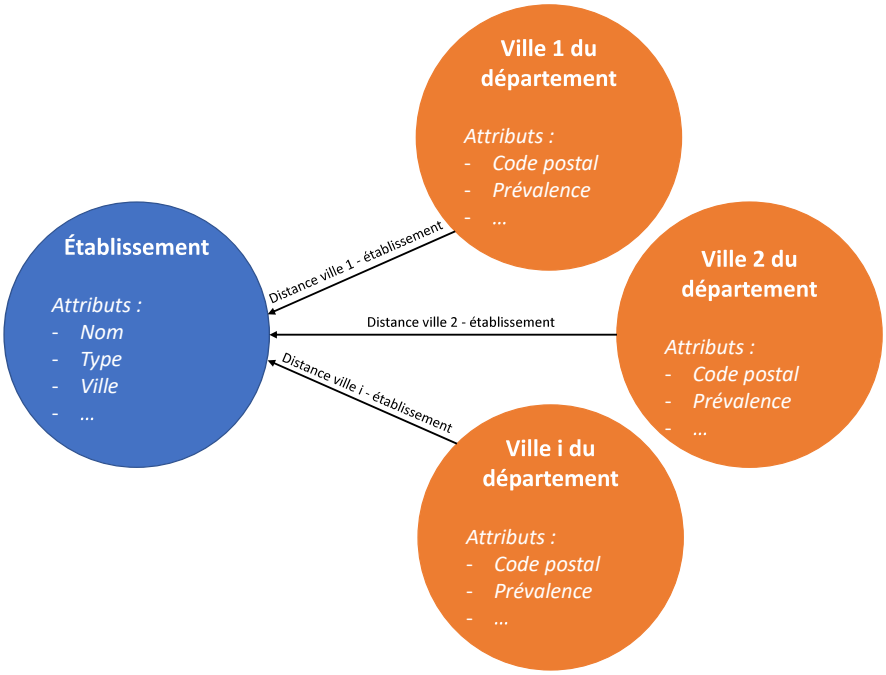
4 Création des arêtes du graphe

4.1 L'API OpenStreetMap

La première idée que nous avons eue pour calculer les distances entre établissements et malades a été d'utiliser l'API d'OpenStreetMap pour calculer l'itinéraire, en voiture, de l'adresse de l'établissement à l'adresse du malade. Malheureusement, nous sommes dans l'incapacité de connaître l'adresse précise du malade. C'est pourquoi nous avons limité la précision de notre calcul à la ville de résidence du malade et à la ville de l'établissement de santé. OpenStreetMap comprend les requêtes de ville et choisit un point de la ville voulue pour réaliser ses calculs.

Nous voulions donc réaliser un calcul d'itinéraire pour chacun des $1471 \times 4485 \approx 6.5 \times 10^6$ couples (établissement, ville d'un malade). Chaque calcul grâce à l'API prenant approximativement une seconde, la durée totale de processing aurait été beaucoup trop longue.

(a) Exemple d'un extrait de graphe à l'échelle départementale



(b) Même extrait de graphe après passage à l'échelle communale

Figure 2: Réduction d'échelle, du département à la ville

4.2 Distance à vol d'oiseau

Pour remédier à cela, nous avons envisagé une autre approche. Au lieu de calculer la distance de l'itinéraire en voiture entre deux points, nous avons calculé la distance à vol d'oiseau entre ces deux points, qui en est une bonne approximation en premier lieu. Pour cela, il suffisait simplement de calculer les coordonnées (latitute et longitude) de chaque ville grâce à l'API d'OpenStreetMap. La formule de Haversine, s'exécutant en un temps très rapide, permet alors de déterminer la distance à vol d'oiseau en fonction des coordonnées de deux points sur la Terre.

4.3 Importation des arêtes sur Neo4j

Nous avons cependant rencontré un nouveau problème de dimensionnalité en tentant d'importer sur Neo4j les 6.5×10^6 arêtes ainsi générées. Dans notre fichier d'arêtes au format *.csv*, nous avons donc conservé uniquement les arêtes de distance inférieures à x km. Pour $x = 10$ et $x = 50$, Neo4j est parvenu à construire le graphe complet. Pour $x = 100$, le nombre d'arêtes était trop grand et nous avons donc conservé pour notre étude la valeur $x = 50$.

Voici la query Cypher permettant de créer les arêtes de type *HOSPITALISED_IN* reliant les établissements de santé aux communes françaises en les pondérant par la distance à vol d'oiseau entre la commune du malade et la commune de l'hôpital :

```
LOAD CSV WITH HEADERS FROM "file:///edges_poumon_50km.csv"
AS row FIELDTERMINATOR ";"
MATCH (m:Malade), (e:Etablissement)
WHERE m.index = row.id_malade AND e.nofinesset = row.id_etablissement
CREATE (m)-[edge:HOSPITALISED_IN]->(e)
SET edge=row, edge.Weight=toFloat(row.dist)
```

À la figure 3 est présenté un sous-graphe de notre graphe, contenant tous les nœuds de type *malade* du département 05 (qui contient 7 villes de plus de 2500 habitants) et les établissements de santé auxquels ils sont reliés.

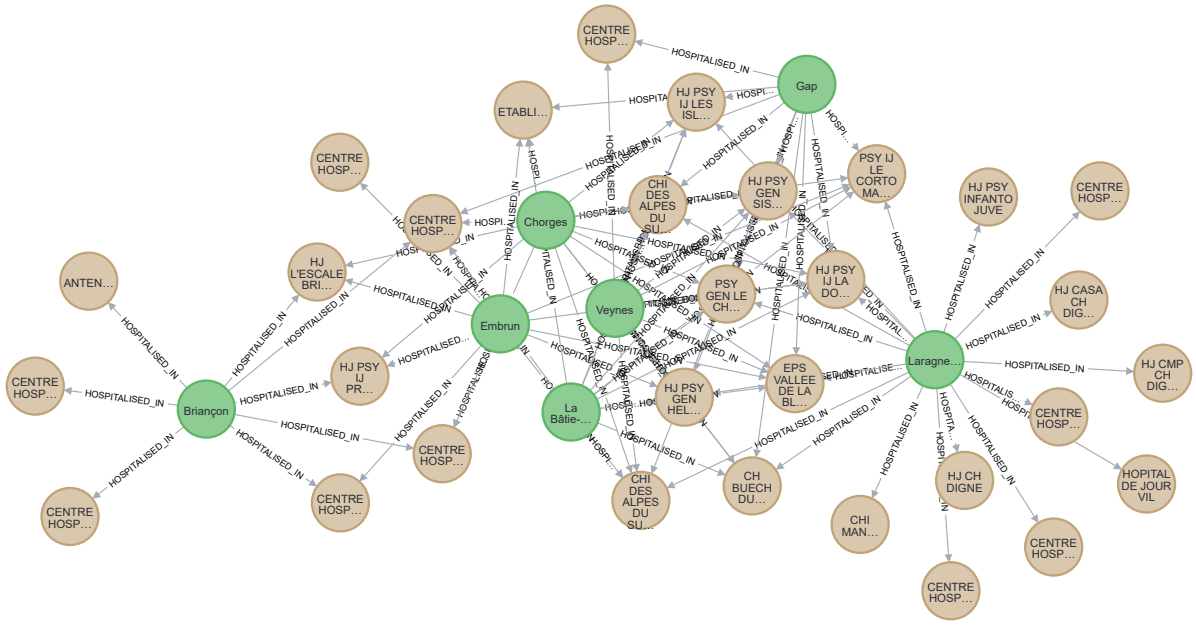


Figure 3: Sous-graphe contenant les malades du département 05

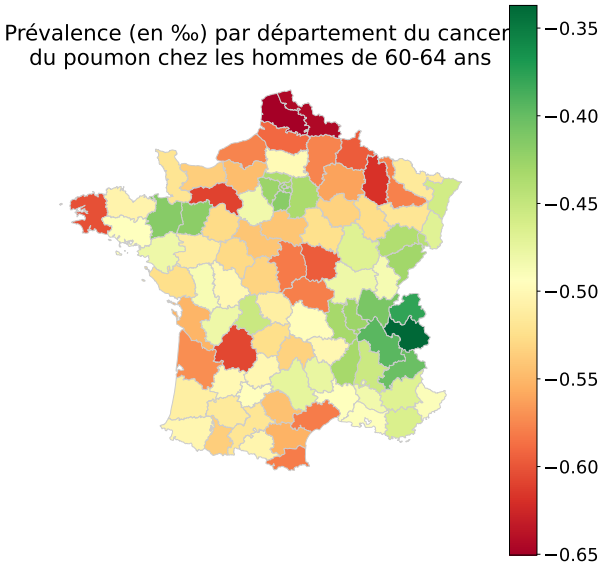
5 Analyse du graphe : requêtes Cypher et visualisation en Geopandas

Chacune des requêtes Cypher que nous avons effectuée renvoie un tableau contenant une valeur pour chacun des 95 départements français, tableau dont Neo4j permet l'export au format *.csv*. Plutôt que de simplement représenter les résultats de nos requêtes comme des tableaux de 95 lignes, nous avons opté pour une visualisation en Python à l'aide de la librairie Geopandas. À cet effet, nous avons chargé une carte des départements français et nous colorons celle-ci en fonction des valeurs obtenues dans chaque requête.

Tout d'abord, pour vérifier la pertinence de notre dataset sur les malades, nous avons affiché la prévalence par département du cancer du poumon pour les hommes de $60-64$ ans grâce à la requête suivante.

```
MATCH (m:Malade)-[edge:HOSPITALISED_IN]->(e:Etablissement)
RETURN toInteger(e.departement) as d, AVG(toFloat(m.prev))
ORDER BY d
```

Nous l'avons affichée à la figure 4 et comparée avec une carte de *Santé publique France* donnant la prévalence du cancer en France par département. Bien que cette dernière ne porte pas uniquement sur les cancers du poumons chez les hommes de $60-64$ ans, nous observons les mêmes tendances : le Nord, la Bretagne et le Centre sont des zones à forte prévalence, tandis que l'Île-de-France et le Rhône-Alpes présentent une prévalence faible. Cela confirme la cohérence de notre jeu de données concernant les malades.



Résultats de notre dataset

Prévalence du cancer en France selon *Santé publique France*

Figure 4

Nous avons par ailleurs cherché à quantifier de plusieurs manières différentes la difficulté d'accès aux soins dans chaque département. Pour cela, nous avons en premier lieu

analysé les disparités en terme de nombres d'hôpitaux à l'aide de la requête suivante. Le résultat est présenté figure 5.

```
MATCH (e:Etablissement)
RETURN toInteger(e.departement) as d, COUNT(e)
ORDER BY d
```

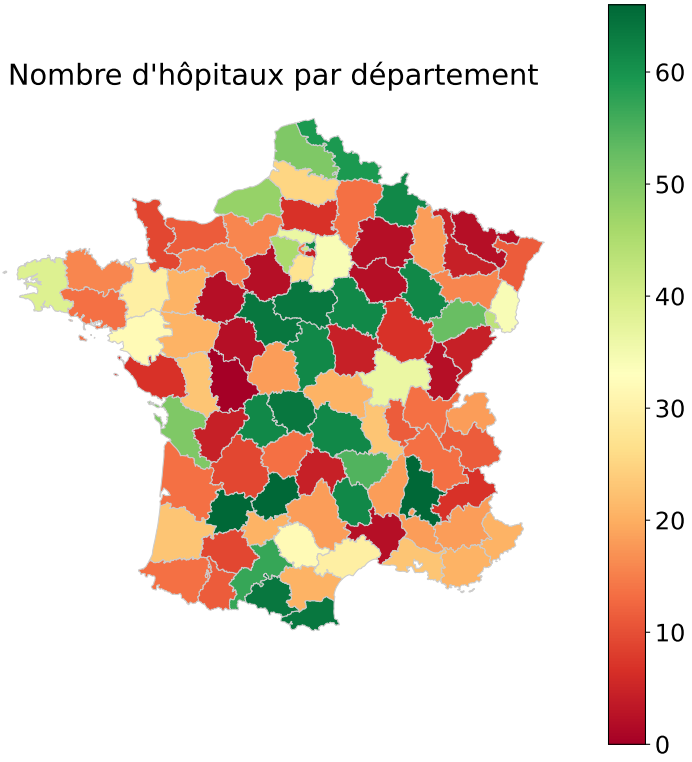


Figure 5

Une autre manière de quantifier la difficulté d'accès aux soins a été de regarder les communes n'ayant pas d'hôpitaux à 20km à la ronde. Nous avons sommé le nombre d'habitants de ces communes en rapportant la somme au nombre d'habitants dans le département (voir figure 6) grâce à la query suivante :

```
MATCH (m:Malade)
WHERE NOT EXISTS {
(m)-[edge:HOSPITALISED_IN]->()
WHERE (edge.Weight < 20)
}
RETURN toInteger(m.dept) as d, SUM(toInteger(m.nb_hab))
ORDER BY d
```

Enfin, nous nous sommes intéressés à la distance moyenne au plus proche hôpital par département en utilisant la requête Cypher ci-dessous (résultat figure 7).

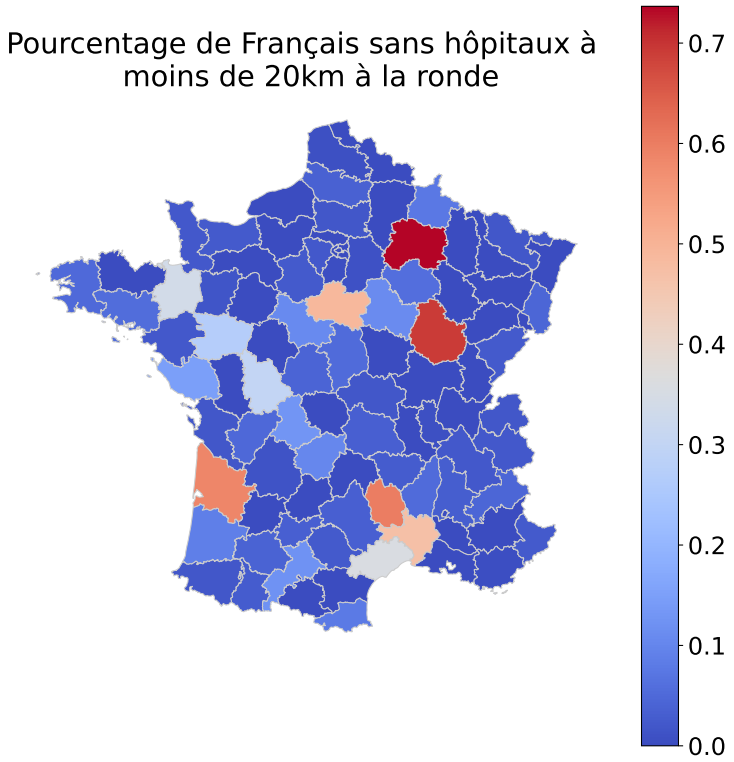


Figure 6

```
MATCH (m:Malade)
OPTIONAL MATCH (m)-[edge:HOSPITALISED_IN]->(e:Etablissement)
WITH m, MIN(edge.Weight) AS dis
RETURN toInteger(m.dept) AS d, AVG(dis) ORDER BY d
```

En comparant la difficulté d'accès aux soins (en terme de nombre d'hôpitaux, de distance moyenne au plus proche hôpital...) avec la prévalence de cancers du poumon dans les départements français, nous n'observons pas de lien particulier. En effet, nous avons réalisé que la prévalence des cancers, en particulier les cancers du poumon, n'était pas tant reliée à la proximité avec des hôpitaux qu'à d'autres facteurs plus importants : habitudes culturelles, prédisposition génétique...

6 Conclusion et ouverture

Nous avons établi un graphe permettant d'étudier la répartition des hôpitaux en France et leur lien avec toutes les communes françaises et les malades y résidant. Nous avons en effet développé un algorithme calculant les distances entre villes et nous avons ainsi pu évaluer les communes et départements n'ayant pas ou peu d'hôpitaux à proximité, dans la perspective de favoriser des propositions d'ouverture d'établissements de santé dans ces régions.

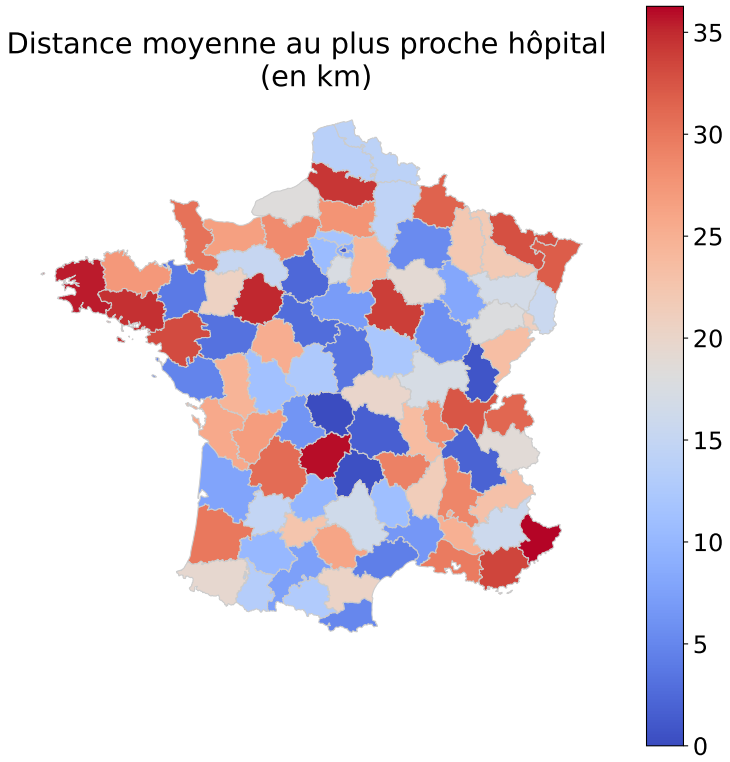


Figure 7

Cependant, au cours de notre étude, nous avons compris qu'un désert médical ne se caractérisait pas uniquement par le simple nombre d'hôpitaux par département. Au contraire, pour mener une étude complète, il faudrait aussi considérer la capacité desdits hôpitaux, le prix des consultations, la durée avant l'obtention d'un rendez-vous, la compétence des médecins y travaillant... En raison du temps qui nous était mis à disposition, nous avons restreint notre travail au facteur qui était le plus simple à obtenir et à quantifier, mais finalement peut-être pas le plus représentatif à lui seul.

Enfin, nous avons expliqué que nous nous étions en premier lieu limités à l'étude d'une maladie particulière par souci de temps et de moyens. Mais le travail que nous avons réalisé est généralisable à un ensemble plus vaste de pathologies, ensemble qui permettrait de mieux caractériser le lien avec l'accès aux soins et pour lequel nous avons d'ailleurs d'ores et déjà les données.