Master 1 in Computer Sciences

Internship Report

Author:
Fabrice Lécuyer
Supervisor:
Albert-László Barabási

Advisors:
Marc Santolini
Amar Dhand
Sean Cornelius

Model of the Hospital Network and Health Trajectories

May-August 2017, Boston

Acknowledgments

This eleven weeks internship has been a meaningful insight of the work of researchers. On the one hand, I have been integrated to the daily working process, which consists in individual work, debates and questioning within the team, and open discussions with other people of the lab. On the other hand, I have been offered to see the long-term strategy of a successful research lab, involving frequent group meetings where feedback is given and talks about related topics by outside guests.

Firstly I would like to thanks Eric[1] for giving me many contacts of complex systems researchers all around the world, and László[2] for offering me the opportunity to spend three months in his great lab whose breathtaking view of Boston is on half of the pictures of my stay, and for supervising the research. I also thank Marc[3] who was my main contact beforehand given the ease of understanding each other in French. The "Hospital Network" team was composed by Marc, Amar[4] and Sean[5], who where always here to answer my questions on a daily basis, to give me precious advice, and to discuss the next steps during group meetings; for that I express my heartfelt thanks.

This internship would not have been possible (really not) without the support of all the people involved in the dreadful paperwork required to set up the stay: N. Trotignon, D. Hirschkoff, F. Vivien, N. Saunier and the ENS Internships Office for the French side, S. Aleva, J. Stanfill and J. Robertson for the American side, and many other people from Lyon 1 and Northeastern who worked in the shadow although I did not interact directly.

Thanks to all the people of the Center for Complex Network Research who create a good atmosphere of scientific emulation, and to the numerous speakers who came to present their interesting work in the eleventh floor. The final word goes to my friends in Boston who made this stay in Boston busy and unforgettable, and to my family for its unconditional support.

This report has two objectives: first, it aims to present the work of the team and my contribution to it, so that it can be read by someone who has no previous idea of the research. Second, it must allow the future usage of my work, hence some specific details that are mostly addressed to the "Hospital Network" team.

[1] Eric Fleury: Professor at ENS Lyon, Deputy Scientific Delegate for INRIA Grenoble
[2] Albert-László Barabási: Director of the CCNR, Distinguished Northeastern Professor
[3] Marc Santolini: Postdoctoral Research Associate
[4] Amar Dhand: Neurologist and Assistant Professor at Harvard Medical School
[5] Sean Cornelius: Postdoctoral Research Associate

Contents

Introduction

Human mobility is an active domain of research that benefits from the huge amounts of data we are able to gather. Physical mobility is often tracked through people's use of technological devices[15] such as GPS location, phone calls or credit card operations. The data is hereby affected by the users, and someone who would use paper maps and cash only would be virtually nonexistent.

In the case of health trajectories however, the records come from billing statements issued by health facilities. In its Healthcare Cost and Utilization Project (HCUP), the Agency for Healthcare Research and Quality provides very precise databases containing in particular a table of hospitals and a table of hospitalizations. The first one presents many attributes for each year and each hospital, such as its name and exact geographic location, its number of beds, its insurance and teaching status[7], its specialization and reputation etc.

The second one draws up an inventory of hospital visits both in emergency and impatient departments. The data is anonymous but each person has a unique identifier called "visitlink" allowing to track her or him throughout the years. Every entry contains hundreds of fields including visitlink, hospital identifier, demographics (age, sex, race) of the patient, year of admission and length of stay[5], a very precise diagnosis of all observed afflictions, the amount of charges and whether the patient died at hospital. An additional information called "daystoevent" allows to compute the number of days between to visits of the same person without being able to know the exact day, for privacy reasons.

Some bias must nonetheless be mentioned: "visitlink" is not available for everyone for various reasons like having no social security number, to the extent that only 76% of the entries can be used. Besides, some people have several reported deaths, leading to think that there could be some mistakes in registrations or in visitlinks.

In this work, the aim is to describe the properties of hospital networks and to understand why patients move from one hospital to one other. For this purpose, we focused on California: "visitlink" is consistent at the state level only, and inter-state movement involving California is less likely than with other states because it is quite isolated of other densely populated areas.

Creating a model of health mobility can be useful in many different ways to improve the health system: while some causes of mobility may lead to better outcomes, some other could be less valuable. Thus, finding the root causes of mobility would be a way to give advice to hospital or insurance companies to prevent superfluous fluxes, to optimize individual and global outcomes, or to enhance the network by building specific hospitals in specific places.

In the first part, I will explain in details the interest and use of the tools I developed to handle database access or map visualization, so that they can be used in the future by the team. In a second part, the initial formulation of the directed network will be presented alongside the results and their comparison with the radiation model. The third part will focus on the second trial, the solutions we propose, and the further questions to be answered.

1 Tools and utility functions

1.1 Database handlers

The database has many tables and is only accessible from the lab with Ethernet wire. For convenience, some of the data had to be transferred on my computer in local. I used therefore two databases simultaneously, which can become messy without clean libraries to handle them.

These tools are all coded in `Python3` using `Pandas`, `SQLalchemy` (for distant connection in `postgreSQL`) and `SQLite3` (for local connection).

1.1.1 Connection and comments

The file `DBconnect.py` contains a class named Connection that defines several methods to handle a database. The idea is to have only one variable for each database (I had one local and one for the lab HCUP db) and to call the methods of this variable.

Once the connection is defined, the method `q` takes a reading SQL query in input and outputs a pandas data frame, while the method `do` executes a writing query (such as updates, insertions and alterations). On the other hand, a given pandas data frame can be saved as a new SQL table with the method `saveToSQL`: the structure, headers, indexes and contents are all stored.

To have a quick insight of the structure of a database, the method `getTables` can be used. It lists the tables and their owners, which is useful to retrieve a specific table among the dozens that already exist. Similarly, `getColumns(t)` will list the columns of a given table `t`. While it could seem pointless given that pandas data frames already contains the headers, it is nonetheless useful in the case of tables that contain hundreds of columns (which is the case of the hospitalization tables).

To guarantee that everyone is able to use the work of everyone else, commenting the tables and fields is necessary. For that matter, I wrote the methods `addComment`, `deleteComment` and `getComment`. The comments can be written for the whole database (for instance to explain where the data comes from and which tables are the most useful), for a given table (to describe its content or explain what the entries are), or for a given column (to mention its type or its precise signification).

For example, the following instructions could be used, resulting in Figure 1:

```
hcup.addComment("California Healthcare Cost and Utilization Project")
hcup.addComment("California Emergency Department Data", "sedd_ca")
hcup.addComment("Hospital unique identifier", "sedd_ca", "dshospid")
hcup.getComment(0, 0)
```

9	15	None	None	California Healthcare Cost and Utilization Pro...	2017-08-01	hcup_user
10	16	sedd_ca	None	California Emergency Department Data	2017-08-01	hcup_user
11	17	sedd_ca	dshospid	Hospital unique identifier	2017-08-01	hcup_user

Figure 1: List of the comments of the hcup connection

1.1.2 Disease specific queries

It is sometimes necessary to see all the patients that were diagnosed with a certain disease. The positive aspect is that the data is very precisely registered: every hospitalization can have up to twenty five diagnoses (`DX1` to `DX25`) indicated by their ICD9 code (International Classification of Diseases). However, this classification is so precise that a same disease can have many different diagnoses: if we are interested in Tuberculosis, we have to check over four hundred ICD9 codes.

To simplify the disease specific queries, I parsed the HCUP documentation[6] using `JavaScript`'s regular expressions to create two `csv` files. This files were then inserted into two new tables: `diagnoses` that relates each of the fifteen thousand ICD9 diagnoses with the corresponding disease number, and `diseases` that gives the name of near three hundred diseases.

Yet, I discovered later that the disease number was also indicated in the data (`DXCCS1` to `DXCCS25`), so the table of diagnoses is useless except for compressing the database. The table of diseases leads to three methods: `diseaseLike` outputs all the diseases whose name contains a substring (Figure 2, `diseaseIn` discriminates the hospitalizations with respect to their diagnoses (Figure 3) and `diseaseSelect` allow to select all the diagnoses and names of associated diseases (Figure 4).

```
local.diseaseLike("fracture")
```

	disease	name
0	207	Pathological fracture
1	226	Fracture of neck of femur (hip)
2	228	Skull and face fractures
3	229	Fracture of upper limb
4	230	Fracture of lower limb
5	231	Other fractures

Figure 2: Different types of fracture returned by the `diseaseLike` method

1.2 Geography

Because human mobility is always related to geography, any model requires the knowledge of some variables: distances, demographics, routes... When dozens of places are

[6] `https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt`

```
local.q("select * from sid where "+ local.diseaseIn("fracture") +" limit 3")
```

	age	ageday	agemonth	amonth	asched	asource	asource_x	aweekend	ayear	bmonth	...
0	94.0	None	None	11.0	0.0	1.0	431	0	2008	3	...
1	52.0	None	None	12.0	0.0	1.0	131	0	2004	3	...
2	83.0	None	None	12.0	0.0	1.0	131	0	2006	1	...

Figure 3: Discrimination of entries using `diseaseIn`

```
local.q("select visitlink, "+ diseaseSelect() +" from sid
```

dx1	dxccs1	dxccs1_name	dx2	dxccs2	dxccs2_name	dx3
78039	83.0	Epilepsy; convulsions	51881	131.0	Respiratory failure; insufficiency; arrest (ad...	5711

Figure 4: Selection of diseases names with `diseaseSelect`

considered, manual records are impossible. While the data is public and available, its format is often difficult to use hence the need for parsing.

1.2.1 Counties information

California is divided into 58 counties of various areas, populations and locations. Since one of our mobility model was computed on a countywide scale, I needed to extract the census information about counties. For this purpose, I used `Javascript` once again to parse a Wikipedia page[7]. The relevant information contains name, administrative center, FIPS county code, population and area. This is stored in the relation `counties` of the database.

In order to compute distances between counties, we need to have geographic coordinates. While a barycenter would have been more mathematically honest, I chose to take the location of the administrative seat which is generally the densest zone. Openstreetmaps provide a tool called Nominatim that is included in the `Python` library `geopy`: it finds a geographic location given its name (see Figure 5). Note that more than one request per second is forbidden[8] (I got temporarily banned before discovering these rules).

Finally, a benchmark of mobility is given by census data that counts the moves inside and between counties. Our healthcare data ranges from 2005 to 2011 so we chose the `2007-2001 County-to-County Migration Flows`[9]. This file is precise and complete, which means we did not need all of it for our work but wanted a clearer and less redundant

[7] `https://en.wikipedia.org/wiki/List_of_counties_in_California`
[8] `https://operations.osmfoundation.org/policies/nominatim`
[9] `https://www.census.gov/data/.../county-to-county-migration-2007-2011.html`

```
from geopy.geocoders import Nominatim
Nominatim().geocode("boston")
```

```
Location(Boston, Suffolk, Massachusetts, United States of America, (42.3604823, -71.0595678, 0.0))
```

Figure 5: Example of use for `Nominatim`s implementation in `geopy`

file. With `LibreOffice Calc` programming language, I could eliminate all the superfluous information and obtain a clean table which is stored in `counties_census` (contains the number of movers and a margin of error).

1.2.2 Distance and routing APIs

With the coordinates it was easy to compute geographic distances thanks to `vincenty` function of `geopy` (which has no limitation since it is just about the maths). However, this distance is not really relevant for human mobility since people cannot usually fly. Willing to establish the road distance and driving time between every pair of counties, I discovered the Here API[10] which is a rooting software with many options. The basic free plan allows 15,000 requests per month: enough for pairs of counties ($58^2 = 3,364$) but not for pairs of hospitals ($434^2 = 188,356$).

1.3 Visualization

1.3.1 Scatter plots

While the `pyplot` library is quite powerful and well-documented, its functions are not always convenient for intensive usage. Even though people usually prefer to write their own functions (and that's what I ended up doing), I will present to you the function `rapidScatter` whose code can be found in `rapidScatter.py`.

This function takes two data sets x and y which can be lists, arrays, data frames, etc. The next arguments set labels for x-axis, y-axis, and dots, and the title of the graph; they are all optional. It is also possible to set the style with `marker` (style of dots), `alpha` (opacity), `c` (z-axis values), `cmap` (z-axis colors).

The plot can easily be set to a logarithmic scale with the parameter `logscale`$= [1, 0]$ (x-axis in log scale, y-axis in linear scale). The window ie the min and max values of the axis can be set with a list of integers: `window`$= [xmin, xmax, ymin, ymax]$ ($None$ or incomplete list set those values automatically considering the data sets).

Various regressions can be plotted with the parameter `reg` which is a list of string values among *equal*, *power*, *proportional*, *linear*. The boolean `correlation` indicates if correlation coefficients have to be printed in addition to being returned. For a better comparison between data and regression, `boxplot` integer parameter can be used (between 10 and 20 to have a good visibility).

Lastly, parameters `show`, `save` (booleans) and `saveFolder` allow to save automatically

[10] `https://developer.here.com`

the plot in local and to choose whether it should be showed (in the Jupyter environment for instance). Most of the graphics of this report have been obtained with this function, which uses `matplotlib`, `numpy` and `scipy.stats` libraries.

1.3.2 Dynamic maps of the network

One of the first concerns of this internships was to be able to see the map of Californian hospitals. For this purpose, I discovered `OpenLayers` which is a powerful open source `JavaScript` library. The functions allow to use a map provided by OpenStreetMap (a collaborative open source cartography software) and to modify it with various types of markers.

The main realizations are showed on Figure 6 (theses `html/js` maps are not available online because of the non-disclosure agreement we signed: no specific data about hospitals or patients may be shown). Just like for Python, two versions of OpenLayers exist. I started with OL2 but it was unconvenient to draw the edges of the network, so I switched to OL3. The differences are big enough to be confusing so I recommend using the third version from the beginning. Various beautiful cartography layers are provided by OpenStreetMap, Stamen and Mapbox.

These maps have been very useful to build our intuition, but further improvements are possible. Among other ideas, we thought about using pop-ups to display information about hospitals or edges dynamically, selecting one hospital to show its neighborhood, or create a continuous heat-map of the network.

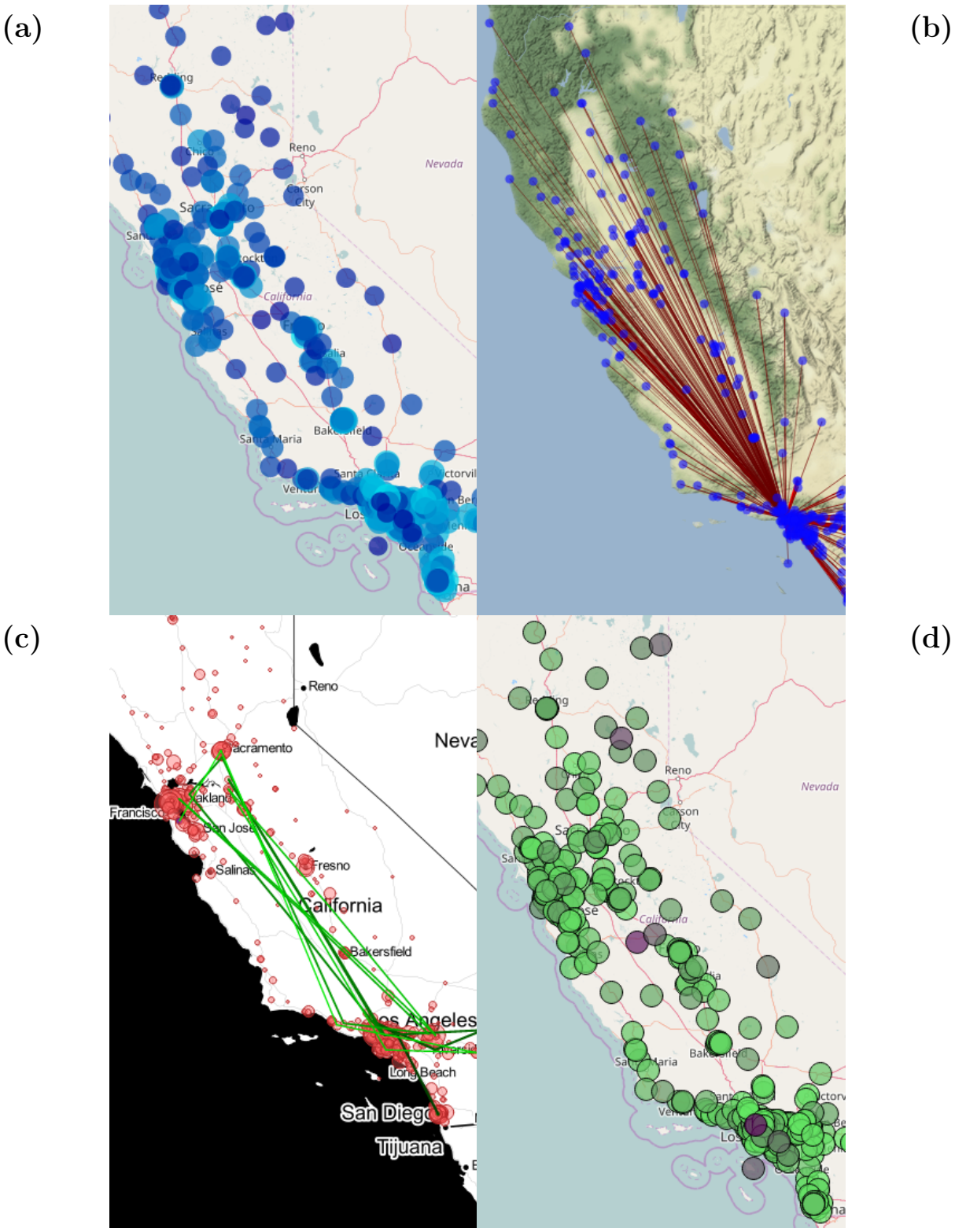

Figure 6: Dynamic maps showing different properties of the hospital network. **(a)** Californian hospitals by number of beds (lighter=bigger). **(b)** Outgoing flow from a given LA hospital. **(c)** Health trajectory of an imaginary patient (for privacy reasons). **(d)** Californian hospitals by strength in the network (weighted degree).

2 A first directed graph

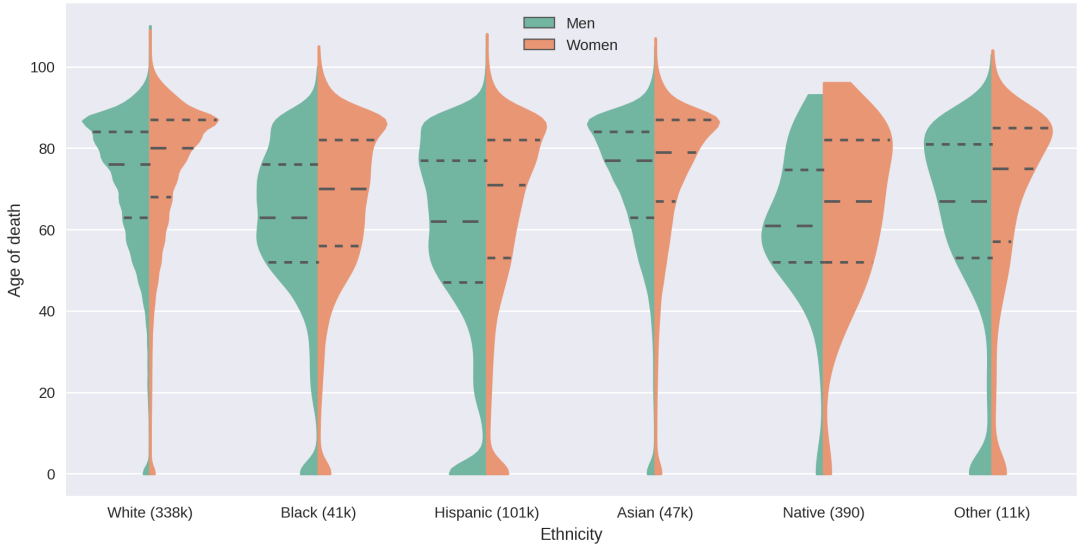

Figure 7: Life expectancy with respect to sex and ethnicity (sampled on 540k deaths; Native population may be to small for statistical relevance)

2.1 Exploration of the database

With such a big data set available, many interesting statistics can be obtained. Note that the data is confidential: we had to signed a non-disclosure agreement and to send it to HCUP's authority. For this reason, it is forbidden to display results that involve ten people or less so that no one can be identify. That is why the map 6(c) shows a random trajectory and not an individual one.

As mentioned previously, the database contains a table for hospitals, a table for emergency and one for inpatient departments. In California only, there are 480 hospitals that got 49 millions emergencies and 21 millions inpatients. Among them, 76% contain enough information to be used in our models, which makes a total of 10.5 millions of distinct patients with at least two visits between 2005 and 2011.

Using `SQL` requests, we can get lot of information about demographics. Thus the average age of people going to hospital is 55 years old, and 60% are women. About one fortieth of hospitalizations result in the death of the patient. For the overall population, the age of death is showed on figure 8: infant mortality is big, then the risk of death is small until 40 years old; after that, several Gaussian curves seem to be overlaid, with a median at 76.

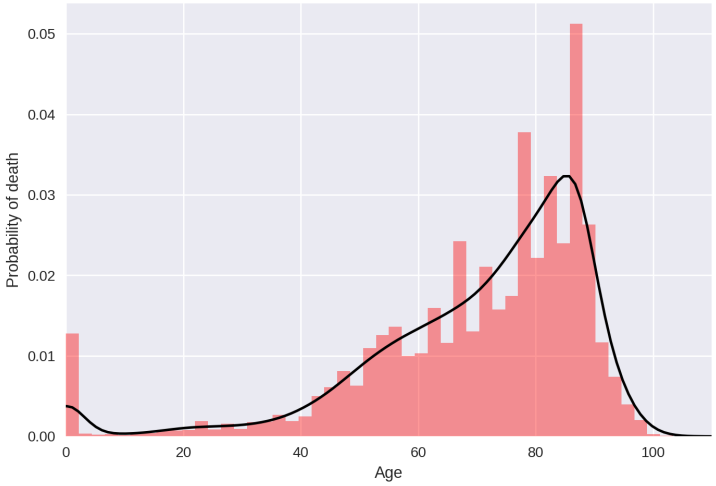

Figure 8: Age of death distribution (sampled on 540k deaths)

In addition to age and sex, ethnicity is mentioned. Even though such record can be controversial, it allows to plot an insight of racial inequalities when it comes to life expectancy. The causes of such differences won't be discussed here, but the Figure 7 clearly shows the gap between Hispanic men and White women, whose median age of death is respectively 62 and 88 years old. Another interesting fact is that the shapes are very different, especially an important amount of deaths among 60 years-old Black and Hispanic male populations that does not appear among other categories. Yet, we have to keep in mind that causality cannot be deduced so simply because of the data: there are only 500k deaths in few years and the registered deaths are only those which happen in hospitals; it means that if all 80 years-old people of one given ethnicity decided to die at home (because of inclinations such as culture or cost), the curve would tell us that this ethnicity dies very early.

Regarding hospitals, we see on Figure 9 that most of them have around 100 beds while the biggest have up to thousand of them. This is a big discrepancy of the network so we can think that this parameter will be important to predict the fluxes between two hospitals. An other important parameter should be the physical distance: Figure 10 shows that most of the pairs of hospitals are at 50km from one another, but there is a big pike at 600km which corresponds to the distance between Los Angeles and San Fransisco. Most of hospitals are indeed located in these two areas (see maps on Figure 6).

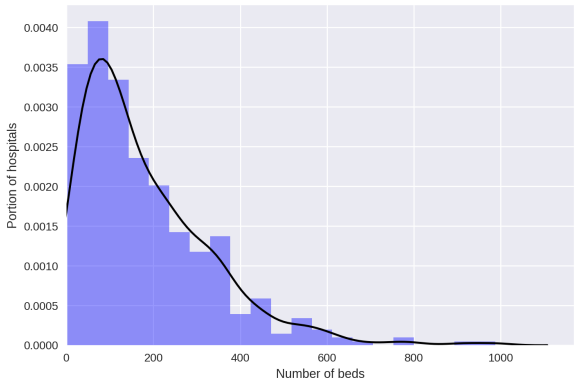

Figure 9: Number of beds

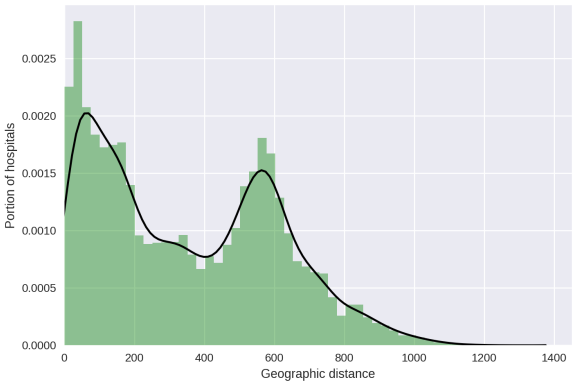

Figure 10: Distance between hospitals

To obtain these results, some hospitals have been removed because important data was missing: coordinates, size, zipcode... They are probably duplication of other entries. We also have to be cautious because every hospital has one entry per year. It was the cause for a big mistake in the first network we built: instead of taking into account the number of beds, I took the sum on all entries of the same hospital, which is not just a proportionality error since some hospitals have more entries than others.

Later, I got into trouble because many hospitals have several identifiers or location, and few are even located in different counties according to the year. Once every redundancy and inconsistency has been erased, we get 434 clean and distinct hospitals. Some of them are different units in the same location but it does not matter.

2.2 Transitive directed network

Previously, the team had been working on this network with an undirected graph: the strength of edges was related to the number of patients shared by two hospitals. Their work includes three broad topics: on the hospital scale, they link properties such as rank (which is evaluated by a dedicated commission which measures many elements related to medical efficiency and patient's satisfaction) with outcomes like cost or death rate. On the network scale, they compute new properties of the hospitals like centrality[6] or strength; these new variables can then be used to relate the outcomes to network properties. On the patient scale, they describe the mobility of individuals to classify them between returners and explorers[14] for instance.

By analogy with PageRank, where edges are oriented from the source to the destination of an hyperlink, we got the idea of creating a directed graph. For this purpose, most intuitive orientation to use is time: we know the chronology of hospital admissions for each patient with `daystoevent` variable. Therefore, the idea is to create an edge of weight 1 from A to B for every person that visited A before B.

As shown on Figure 11, this network has a property of transitivity: if one patient has been in hospitals A then B then C, we count three edges (AB, BC, AC). This decision was initially a way to improve the speed of SQL requests because they take several hours to process the millions of entries. Transitivity allows to reduce the algorithmic complexity of the request.

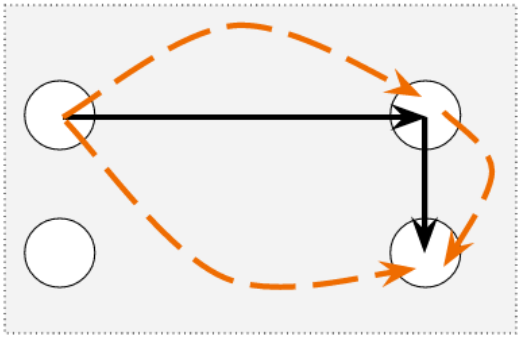

Figure 11: Transitivity of the network

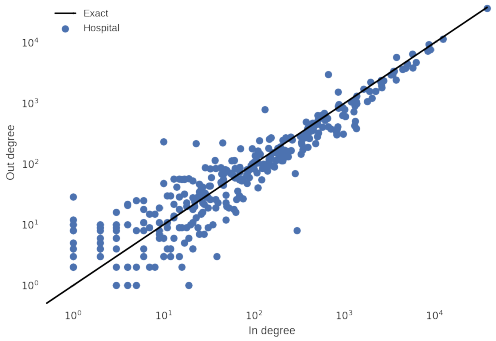

Figure 12: Proximity of in and out degrees in the directed network

Note that in this definition of the network, we used only the "sedd" (emergency department) data. Once again, this decision was a way to make the requests simpler and faster: we overcome the unions and joins that would have been necessary to include the "sid" (inpatient) table. However, we kept in mind that excluding inpatient data can lead to unforeseeable bias and problems.

To ensure some kind of backward compatibility with undirected models, we can check that in and out degrees of a given hospital are not strongly different (see Figure 12). It means that we can use the total degree the same way we used the degree in undirected models. Note that it does not mean that direction is useless, because the two opposed edges that link a pair of hospitals can may be strongly asymmetric.

2.3 Mobility as radiation

This network is based on health mobility of people and stacked over a physical layer of cities and roads. Such a structure raises issues such as: How can mobility between hospitals be described? Can the migration be explained by geography and demography? To answer these broad questions, the goal is to find or create a model that fits the flows we observe.

Most of the time, geographic mobility is explained by an instance of the gravity model[12]. Inspired by Newton's law of gravitation, it assumes that the force between two positions i and j is proportional to the product of the masses of i and j divided by their distance squared. This formula is usually modified in that way: the force becomes a flux, the masses become a parameter that depends on the problem, and the distance has an exponent γ which has no specific reason to be equal to 2. Thus we obtain:

$$F_{ij} = G\frac{m_i m_j}{r^\gamma}$$

However, this model suggests that the relation between two positions is symmetric. Since we now have a directed graph, we observe that mobility is not always similar in opposite directions, so we would like to have a model that takes into account the direction in addition to "masses" and distance. Discussing with László and other researchers of the lab, we were told about the *radiation model*[16]. Also inspired by physical phenomena, this model computes the probabilistic distribution of destinations for people leaving a given position.

This model was originally designed with the example of people who move out because they search for a better job. One person leaves a current job i valued at x and reaches the closest place j where a job with higher value $y > x$ is offered. The number of job offers is considered proportional to the population, so we obtain a model that depends heavily on the density of population. The formula is the following:

$$F_{ij} = T\frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}$$

In this equation, T is the total number of people leaving position i, and the other term corresponds to the probabilistic distribution. Masses m_i and m_j are just like in gravity

models; s_{ij} is the sum of masses whose position is in the circle of center i and radius $[ij]$ (ie the quantity of job offers that are closer to i than j). Note that this model is *not* symmetric because the denominator's first term has m_i but not m_j, and because $s_{ij} \neq s_{ji}$ (see Figure 13).

This formula has been proved very accurate in modeling several types of mobility: travellers, migrants, phone calls, freight... In particular, it was used in the paper[16] to predict the flows of migrants between US counties (counties are geographical subdivisions of states). Our idea was therefore to adapt this model to health mobility.

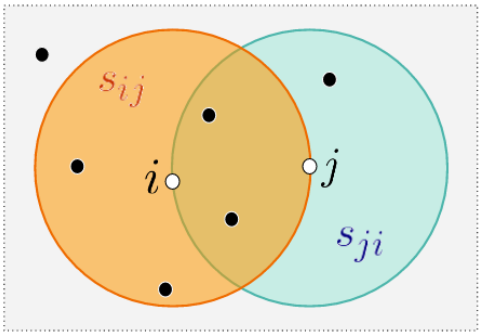

Figure 13: Illustration of s_{ij} and s_{ji}

2.4 Aggregation by counties

To compare our model with relevant data, we chose to consider health mobility between counties. Lot of inter-county data is available thanks to census (see part 1.2.1). Thus, we compute the sum of fluxes between all the hospitals of county a and all hospitals of county b and define this number as the health migration from a to b. We obtain a value for each of the 2500 pairs of counties; the aim is then to use the properties of counties and pairs (number of inhabitants and distances for instance) to build a model that approximates those numbers.

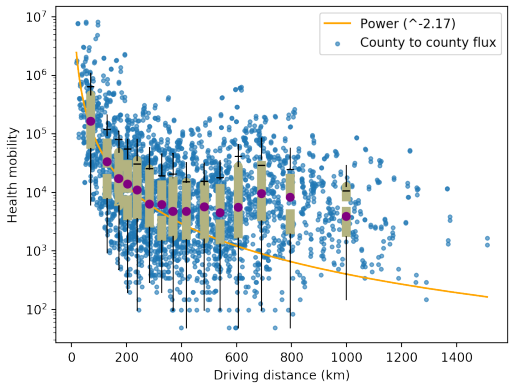

Figure 14: Health mobility with respect to driving distance

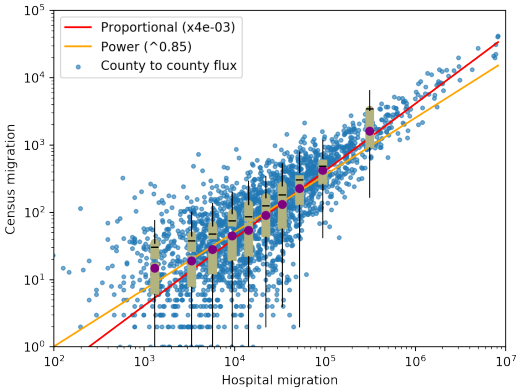

Figure 15: County to county mobility: census vs health

Before further work, we wanted to prove that geography was not enough to explain

everything. Figure 14 shows that distance explains mobility quite well for neighbors counties: mobility follows distance with an exponent -2.17 under 200km but variance becomes huge beyond that point. Moreover, the noticeable pike at 600km is visible (distance between LA and SF) which suggests that demography plays a crucial role.

Figure 15 shows that census mobility between counties is quite a good match with health mobility. It means that the order of magnitude of patients flows is strongly related to the number of people who move. Yet, there is an important noise around the proportionality (red) line because of other parameters. Moreover, we do not predict the flows between two hospitals in the same county.

2.5 Results of the radiation

The simplest definitions we can use for "masses" of hospitals is their number of beds. This information is given for every year and every hospital in HCUP database. We get the mass m_a of county a by computing the total number of beds it contains; s_{ab} is hereby the number of "beds in the circle" of center a and radius $[ab]$. Note again that county-level allows to use only 2500 pairs (while there would be 190.000 pairs of hospitals), leading to a much faster computation: we have a cubic complexity because s_{ab} is a sum of county masses computed for every pair of counties.

Remark that the number of hospitals (figure 16) is not proportional to the number of inhabitants of a county. It can be explained by the fact that denser counties can have bigger (but fewer) hospitals, which is confirmed by the regularity of the 100 inhabitants per hospital bed (figure 17).

The radiation leads to mitigate results: while it is quite satisfying for big counties like Alameda in the dense bay of San Fransisco (figure 18), it turns out to be disappointing for smaller counties like Mono near Yosemite Park (figure 19). We also note that geographic distances (blue), routing distances (orange) and driving duration (green) give very similar results while we thought it could explain some important differences between short shifts and long shifts (because of accessibility in country and speed limits in cities).

More generally, precision decreases with population (figure 20). Several hypothesis were made to explain this problem. First, our model erases the detail of inter-hospital mobility: small counties which only have few hospitals are more subject to migration due to hospitals specializations, while big counties have a more representative set of hospitals. Second, we only took emergency data: people who go to emergency department choose the closest hospital of the location they are at the crucial moment, which can be a bias. Third, the way we make the aggregation can count the same patient several times in some cases; if the rate of doubled patients is different in different counties, it would flaw the results with this unknown parameter.

In Figure 21, we see that the prediction of this radiation model does not fit the data (red line): a power regression gives an exponent of 1.61 which is far from 1, and even this orange line is not very accurate especially because the graphics is in logarithmic scale. Some predictions have several orders of magnitude which means that other parameters must be introduced. For this purpose, we re-build the whole network with inpatient data, and with emergency and inpatient mixed, but the results were not better.

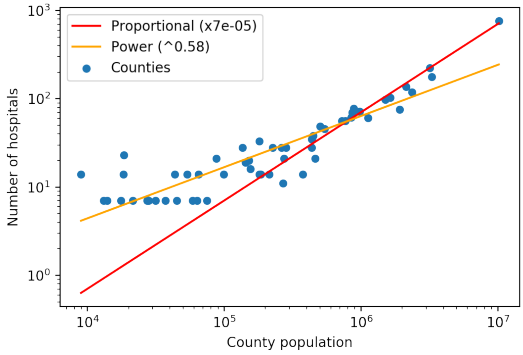

Figure 16: Hospitals per capita

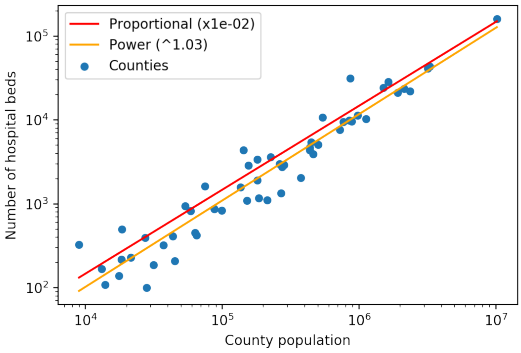

Figure 17: Hospital beds per capita

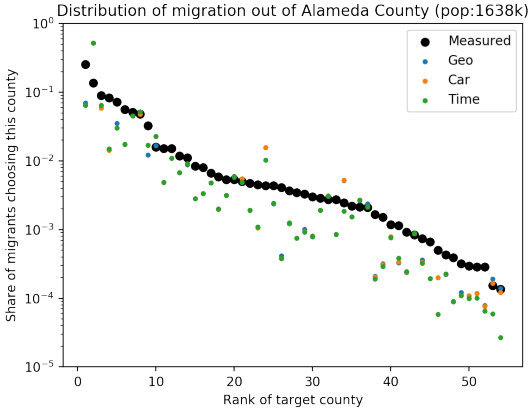

Figure 18: Destinations from Alameda

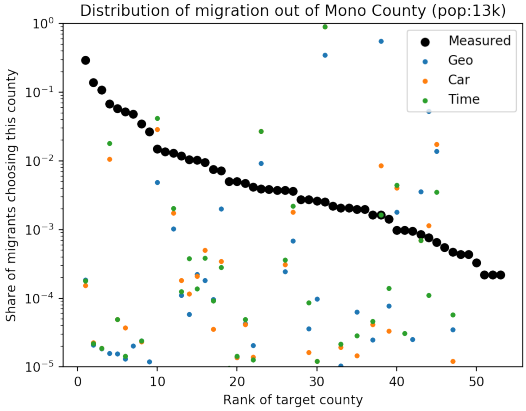

Figure 19: Destinations from Mono

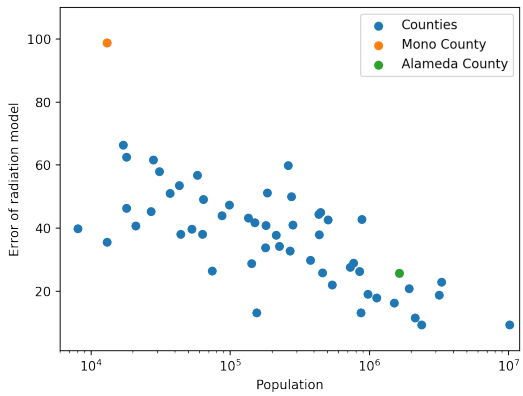

Figure 20: Error in the model with respect to county population

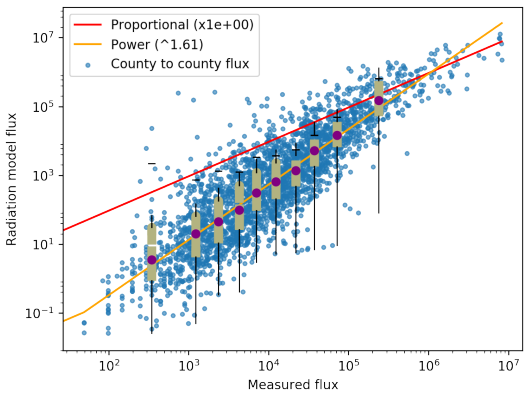

Figure 21: Model and real flows between pairs of counties

3 Daily flow of people

3.1 Redefinition of the network

The failure of this model can have two main causes: on the one hand, it is possible that a simple model taking only distance and number of beds into account is not enough to explain the complex motives of health mobility[1]; on the other hand, we have based our computation on debatable initial assumptions. Keeping in mind the "law of parsimony", we first wanted to check the second possibility before considering new refinements of the model.

The first questionable point is the county-scale: despite its computational efficiency and the relevance of census benchmarks, it rules out short range mobility in a quite arbitrary way. It was therefore decided to build a whole new network that would not be designed to be aggregated by counties: we stay at a hospital-scale, raising from 58 positions to 434. Another assumption was transitivity (see section 2.2). It was logical when we considered that an edge had to be drawn to show the temporal prevalence of an hospital upon another. However the aim is now to have a value that represents the daily flow[8] of people between hospitals. Similarly, the same person can be counted several times when several transfers have been done.

All these improvements lead to a more intuitive definition of the network but dramatically increase the computation time of every subroutine. Some optimization and many testing had to be done before running them during hours (this can be seen with some explanation in the `Jupyter` notebooks). The results now involve 190.000 pairs of hospitals and are shown in Figure 22. Boxplots show that the tendency is a 0.64 power law, but highest fluxes are a surprising good fit. This bimodality is even easier to see in Figure 23.

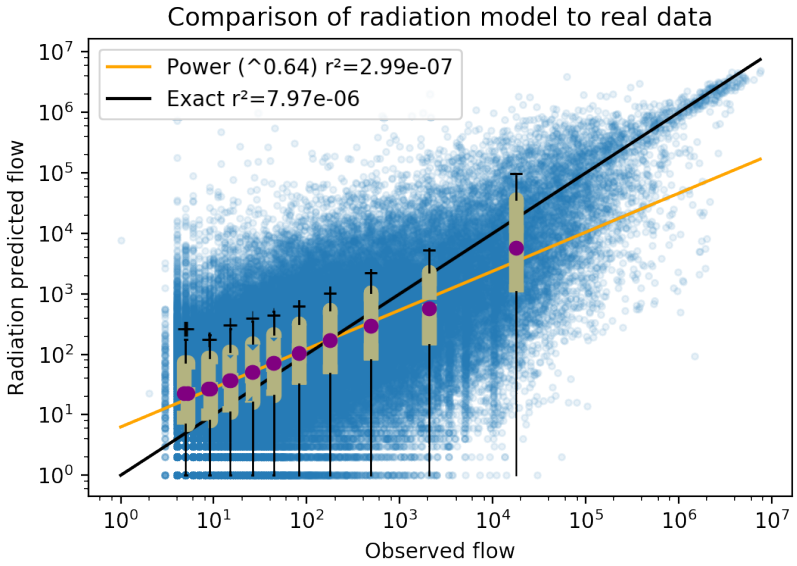

Figure 22: Model and real flows between pairs of hospitals. Boxplots and purple medians show that low fluxes follow a concave power law while high fluxes predictions match the reality.

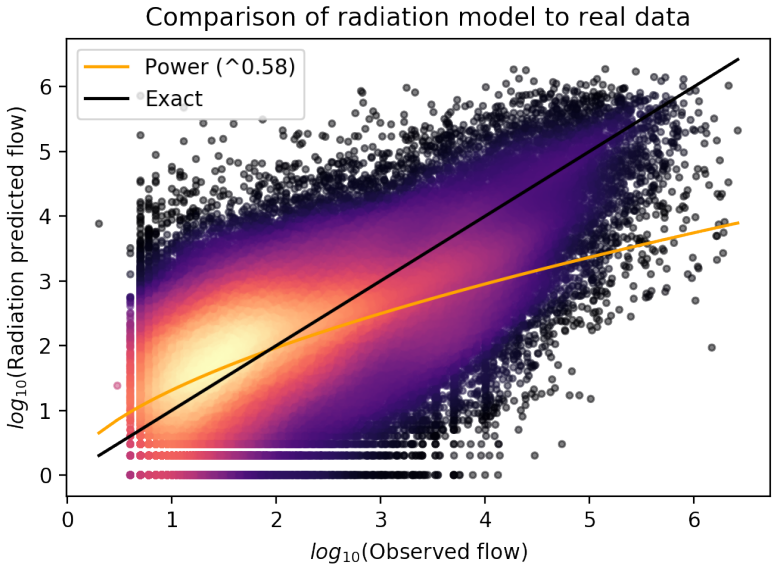

Figure 23: Model and real flows between pairs of hospitals. Self-loops have been removed but the two tendencies are still clearly visible.

3.2 Rejected hypotheses

It seems that two distinct modes are governing health mobility. To see if a simple classification of hospitals or patients could explain that, I tried to plot the same thing with different filters: size of the hospital, distance of the pair, centrality or degree of vertices... None of these attempts was successful to separate two point clouds each following one of the modes.

We also tried to see if different diseases could be a cause for different modes. With the help of Amar who is also a Doctor, we took 4 strongly different diagnoses and rebuilt a new network for each of them, taking only into account people who got these diseases. Acute myocardial infarction, Coronary atherosclerosis, Pneumonia and Fracture were selected[9]. The last one did not involve enough people to be statistically relevant, but the three other showed the same two tendencies. Therefore we think that there is no crucial link between type of disease and mobility, as Sean's work suggested.

Parsimony seems to be unable to explain health mobility, so new parameters are needed. Figure 24 shows that distance is well understood by the model, while Figure 25 shows that number of beds is not. The most intuitive idea we got to do so was to use a new definition of the mass of hospitals. However, the radiation model involves the computation of composed variables s_{ij} and hide the influence of the mass behind several additive and multiplicative terms. Finding which parameters should enter the definition of mass is hereby very difficult. For this reason, we decided to come back to gravity models where every variable appears independently, at least to find the new optimal definition of mass.

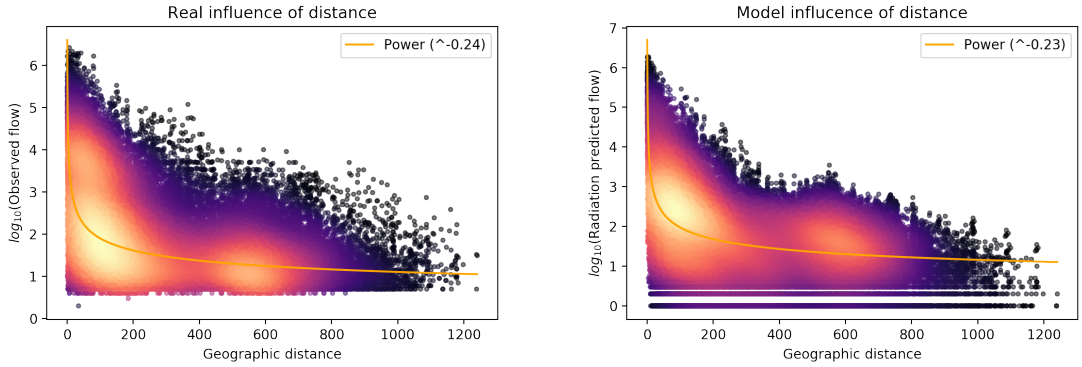

Figure 24: Flow wrt distance, in reality (left) and in model (right)

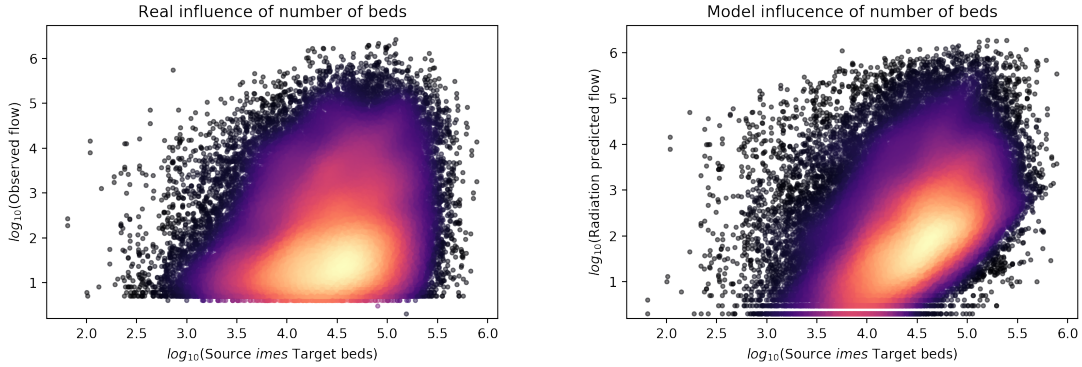

Figure 25: Flow wrt bedsize, in reality (left) and in model (right)

3.3 Further work

Mass define the importance of an hospital in the network. The variables we think to use are various and not always easily obtainable. The most important should be size and centrality, reputation and rank, which are available thanks to previous work of the team. Insurance and rehabilitation status can be a way to gather information about the average patient visiting the hospital. Clusters of hospitals and teaching status may have their importance especially in short range migrations. Finally, subjective hidden parameters such as ambulance agreements, marketing, word-of-mouth and visibility could play a role.

To find a small subset of these numerous variables that would explain health mobility with accuracy, we intend to use Lasso[17] methods. Gravity laws are indeed a linear model when we take the logarithm of each variable. In addition to give a predictive model of mobility, such a result would be very interesting in order to understand what may be the main motives for moving from one hospital to another.

Conclusion

During this eleven weeks internship, several tools have been designed to handle the database and to produce maps and graphics easily. The next focus has been to set up

the radiation model in the case of hospital networks. Noticing inconsistency in our basic assumptions, we have then tried to correct them and to define a network with intuitive specifications.

While number of beds is almost proportional to population, we found that it was not enough to define the mass of hospitals. To have a more accurate model in future works, we intend to use linear models to find the best subset of parameters to take into account in the mass.

Thanks to the dynamic team and the multidisciplinary lab, the work was very interesting. I learned a lot about medicine, biology, statistics and sociology, especially during the talks about various subjects that were taking place weekly in the lab. For this memorable opportunity, I thank all the people involved again.

References

[1] G. Abraham, G. Byrnes, and C. Bain, *Short-term forecasting of emergency inpatient flow*, Transactions of Information Technology in Biomedicine, (2009).

[2] Albert-László Barabási, *Linked*, Basic Books, 2002.

[3] ——, *Bursts*, Plume, 2010.

[4] ——, *Network Science*, Cambridge University Press, 2016.

[5] M. Barnett, D. Grabowski, and A. Mehrotra, *Home-to-home time: Measuring what matters to patients and payers*, New England Journal of Medicine, (2017).

[6] S. Borgatti, *Centrality and network flow*, Social Networks, (2005).

[7] L. Burke, A. Frakt, D. Khullar, J. Orav, and A. Jha, *Association between teaching status and mortality in us hospitals*, American Medical Association, (2017).

[8] X. Chen, L. Wang, J. Ding, and N. Thomas, *Patient flow scheduling and capacity planning in a smart hospital environment*, IEEE, (2016).

[9] K. Dharmarajan, A. Hsieh, V. Kulkarni, Z. Lin, J. Ross, L. Horwitz, N. Kim, L. Suter, H. Lin, S.-L. Normand, and H. Krumholz, *Trajectories of risk after hospitalization for heart failure, acute myocardial infarction, or pneumonia: retrospective cohort study*, British Medical Journal, (2014).

[10] J. Fernández-Gracia, J.-P. Onnela, M. Barnett, V. Eguíluz, and N. Christakis, *Influence of a patient transfer network of us inpatient facilities on the incidence of nosocomial infections*, Scientific Reports, (2017).

[11] C. Hidalgo, *Why Information Grows*, Basic Books, 2015.

[12] Y. Mansury and J. Shin, *Size, connectivity, and tipping in spatial networks: Theory and empirics*, Computers, Environment and Urban Systems, (2015).

[13] N. Nakatsuka, P. Moorjani, and 15more, *The promise of discovering population-specific disease-associated genes in south asia*, Nature Genetics, (2017).

[14] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and Albert-László Barabási, *Returners and explorers dichotomy in human mobility*, Nature Communications, (2015).

[15] C. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. González, *Unravelling daily human mobility motifs*, Royal Society Interface, (2013).

[16] F. Simini, M. González, A. Maritan, and Albert-László Barabási, *A universal model for mobility and migration patterns*, Nature, (2012).

[17] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, (1996).