French spatial disparities in Major Diseases through Multiscalar Lens

Xavier Bacon (ENS Paris-Saclay), Carine Milcent (PSE)

July 3, 2025, Paris School of Economics

- Assess the current state of **hospital visits** in France:

  $\rightsquigarrow$ A framework to investigate spatial heterogeneity in patients diagnoses

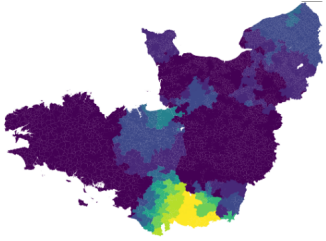


Table of Contents

Data and Pointwise Model

Data description

- **Patients' diagnosis** : information collected[a] when a patient enters a health care facility, recorded as **CMD** (**C**atégorie **M**ajeure de **D**iagnostic)
- 29 groups: 27 regular + medical consultations ('28') + Error ('90')

[a] data provided by the **A**gence **T**echnique de l'**I**nformation sur l'**H**ospitalisation (**ATIH**)

CMD	Libellé
01	Affections du système nerveux
02	Affections de l'œil
03	Affections des oreilles, du nez, de la gorge, de la bouche et des dents
04	Affections de l'appareil respiratoire
05	Affections de l'appareil circulatoire
06	Affections du tube digestif
07	Affections du système hépatobiliaire et du pancréas

	Numéro FINESS	code_PMSI (domicile patient)	CMD
0	123456789	39170	01 - Affections du système nerveux
1	987654321	14110	11 - Affections du rein et des voies urinaires
2	112233445	13320	15 - Grossesses pathologiques, accouchements et affections du post-partum
3	556677889	73460	19 - Maladies et troubles mentaux

⚠ Heterogeneous data ⚠

1. Merging data by municipality:

	Commune	Nom de la commune	CMD 1	CMD 2
0	39170	Ravilloles	17	24
1	14110	Pontecoullant	120	56
2	13320	Bouc Bel Air	21	41
3	73460	Sainte-Hélène-sur-Isère	54	79

2. Normalization to obtain frequencies:
   CMD[*area*] = frequency distribution of patients' CMD in the *area*, *e.g.*
   CMD['75014'] = (0.1, 0.25, 0.65)
3. To compare two areas, one compares their CMD's vectors by the means of a **dissimilarity**:

$$\text{dissimilarity}(\mathbf{CMD}[\text{area A}], \mathbf{CMD}[\text{area B}])$$

In the pointwise model : area A = commune, area B = France and dissimilarity = Kullback–Leibler divergence, that is:

$$\text{div}_{\text{KL}}(\mathbf{CMD}[\text{commune}], \mathbf{CMD}[\text{France}])$$

Context: Given two normalized distributions $P = (P_1, \ldots, P_N)$ and $Q = (Q_1, \ldots, Q_N)$,

$$\begin{aligned} \mathrm{div}_{\mathrm{KL}}(P, Q) &= \sum_{n=1}^{N} P_n \log\left(\frac{P_n}{Q_n}\right) \\ &= \sum_{n=1}^{N} P_n [\log(P_n) - \log(Q_n)] \end{aligned}$$

The question: *How different is Q from P?*

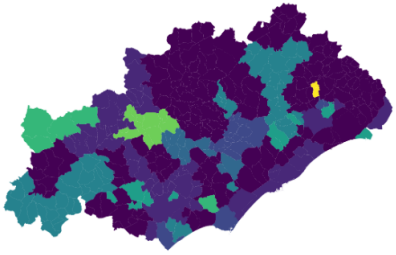
Interpretation:

- Measures how inefficient it is to assume Q when the true distribution is P
- Equal to 0 if and only if $P = Q$
- Not symmetric: $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$

Discussion regarding the dataset

- Framework's requirements : Spatial and multiscalar data.
- These disparities are measured in terms of access to healthcare.
- Both forgone and excessive use of healthcare will be interpreted as indicators of need.
- Each pathology is considered comparable here.

Pointwise model's limitations



- Spatial aspects, at both local and global scales, are completely ignored: two communes can be swapped; the dissimilarities remain unchanged.

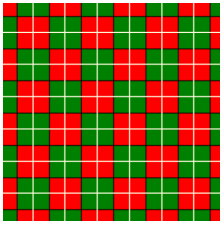


Figure: Two grid patterns

- This depends on the administrative definition of the units.

Multiscalar Lens Model

- Comes from urban segregation: Multiscalar Lens model, 2019, Olteanua M., Randon-Furling J., Clark W.

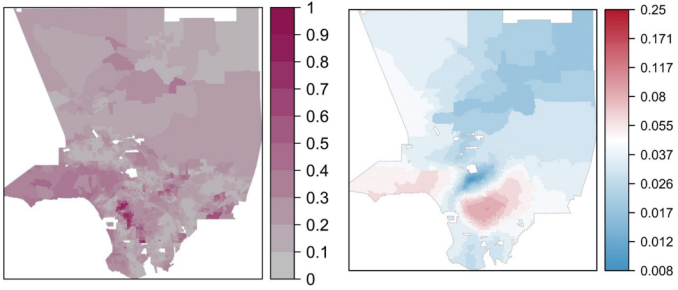


Figure: Ethnic composition in Los Angeles (Left: Pointwise Model, Right: Multiscalar Lens Model

For each commune **com** we compute the distance one needs to cover to get a relatively clear picture of the country, starting from this commune.

Nearest Neighbors

We compute the closest municipalities in order of proximity.

Aggregation process

Compute successively the divergence of **com**, then the divergence of **com** gathered with its nearest neighbor etc.

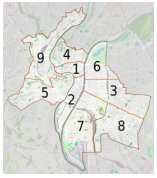


Figure: Districts in Lyon

- $\mathrm{div}_{KL}(\mathrm{Lyon}_1|\mathrm{France})$
- $\mathrm{div}_{KL}(\mathrm{Lyon}_1 \cup \mathrm{Lyon}_4|\mathrm{France})$
- $\mathrm{div}_{KL}(\mathrm{Lyon}_1 \cup \mathrm{Lyon}_4 \cup \mathrm{Lyon}_6|\mathrm{France})$

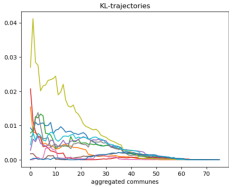


Figure: KL-trajectories

Focal distance

To quantify the speed of convergence of each trajectory, one fixes a threshold δ and determines for each **com** the step $f_{\mathbf{com}(\delta)}$ where the trajectory becomes smaller than δ and remains thereafter.

$$f_{\mathbf{com}}(\delta) = \inf_{1 \leqslant u \leqslant l-1} \left\{ n^{i,0:u} | \forall v \geqslant u, d_{KL}\left(\mathbf{CMD}^{0:v}[\mathbf{com}] | \mathbf{CMD}[\text{France}]\right) \leqslant \delta \right\}$$

Distortion coefficient

is defined as the summation of all the delta.

$$\Delta_{\mathbf{com}} = \int_0^\infty f_{\mathbf{com}(\delta)} \, \mathrm{d}\delta$$

Multiscalar Lens Model IV

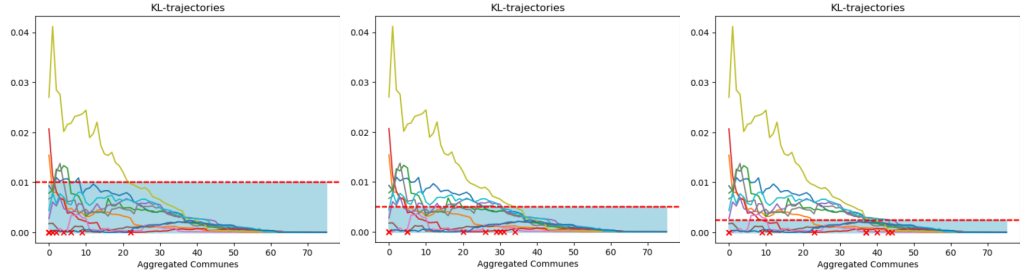


Figure: KL-trajectories

Aggregation process and KL Trajectories

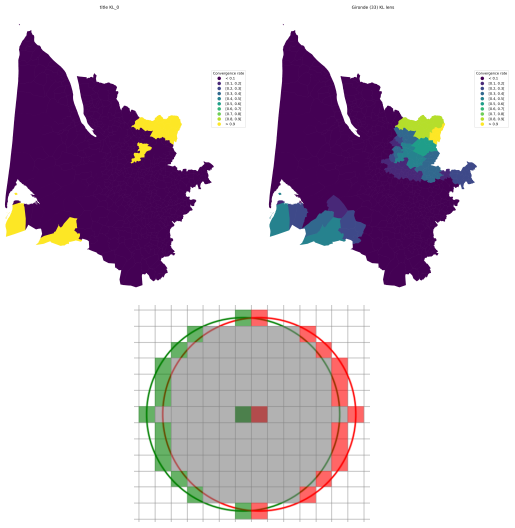
1. Calculate its nearest neighbors
2. Compute its KL-sequence

Focal distance and sommation

1. Compute the focal distance for every threshold.
2. Compute the distortion coefficient $\Delta_{\mathbf{com}}$

Comments on the Multiscalar Lens Model

- It reintegrates municipalities into their broader geographical context and no longer treated as independent entities.
- It has a smoothing effect.



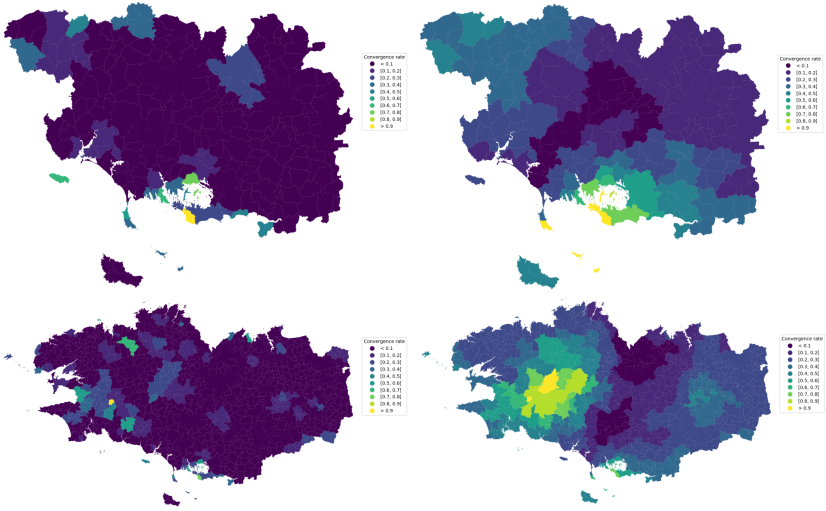


Figure: Pointwise KL (Left), Lens Model (Right), Morbihan (Top), Bretagne (Bottom)

A bit of Statistics

- Age-related causal effects...Simpson's paradox: Simpson's Paradox occurs when a trend or relationship observed within multiple separate groups reverses or disappears when those groups are combined.
- Adjust public health indicators by age/sex group (indirect standardization).

Main idea: there exists a hidden relationship between the observed data and the population structure...*A priori* knowledge: formulas that relate data to certain hidden variables. ...however, these relationships are often unknown... we therefore introduce statistical models to infer them.

$$\mathscr{M}^{\theta_k}(G) = \mathbf{CMD}_k, \text{ for all } k$$

- Consider that the population is divided into several distinct groups age, gender):

$$G = (\text{F0-14}, \ldots, \text{F75+}, \text{H0-14}, \ldots, \text{F75+})$$

  - $G^{0:U}(\mathbf{com})$ = the group distribution in the U-aggregated units around the commune **com**.
  - $\overline{G}$ = the (reference) group distribution (*e.g.* in France).

$$\mathscr{M}^{\theta_k}(G^{0:U}(\mathbf{com})) = \mathbf{CMD}^{0:U}[\mathbf{com}]$$

⇝ Estimation : $\widehat{\theta}_k$;

- Compute Corrected-CMD$^{0:U}[\mathbf{com}] = \mathscr{M}^{\widehat{\theta}_k}(G^{0:U}(\mathbf{com}))$.
- In place of considering the CMD's distribution, we consider the distribution in the area aggregated around but with a socio-demographic distribution corresponding to the national one:

$$\text{div}_{\text{KL}}(\text{Corrected-CMD}^{0:U}[\mathbf{com}] | \overline{\text{CMD}}).$$

- Linear regression : the simplest, but not suitable for categorical (count) data as it can return negative values.
- Poisson regression : Well suited to categorical data. But fits well when mean and variance are approximately equal : $\mathrm{Var}(Y) = \mathbb{E}[Y]$
- Negative binomial distribution : variance is greater than the mean (overdispersion) : $\mathrm{Var}(Y) > \mathbb{E}[Y]$, $\mathrm{Var}(Y) = \mathbb{E}[Y] + \alpha \mathbb{E}[Y]^2$

- Data: is collected and processed by the ATIH (Agence Technique de l'Information sur l'Hospitalisation).
- Multiscalar Lens model, 2019, Olteanu M., Randon-Furling J., Clark W.
- In the short term, the code will be available on the Safepaw website.
- pynsee package containing tools for searching and plotting data from INSEE and IGN. (available in R language also)