Machine Learning and Data-Driven Approaches in Spatial Statistics
A case study of housing price estimation

Sarah Soleiman[1,2] **Julien Randon-Furling,**[1,3] **Thomas Lefebvre**[2]

[1] **SAMM, Université Paris Panthéon-Sorbonne – FP2M (FR2036) CNRS, Paris, France**
[2] **MeilleursAgents, 7 Boulevard Haussmann, Paris, France**
[3] **Department of Mathematics, Columbia University, New York, USA**

1 Introduction

The intertwining of socio-spatial complexity with that of price formation leads to highly challenging questions when modeling real estate markets and the dynamics of property prices. The exact same apartment typically will not have the same price depending on its location in the city – due to specifics of the neighborhoods and even micro-neighborhoods that are difficult to quantify.

Traditional methods rely on the so-called hedonic approaches modified to incorporate spatial effects via geographically weighted regressions. However, the recent availability of big data pertaining to the socio-economic characteristics of cities, at a very fine-grained level, should allow one to capture in much finer detail the complex relationship between space and price in the real estate market.

Our approach is two-fold, we first apply a simple Self-Organizing Map (Kohonen) algorithm on vast sets of demographical, economical and infrastructural data in order to bring out the socio-spatial structure of a city and then use cluster information into the spatial diffusion process of the GWR.

2 Data

We use public data from the French national office of statistics – INSEE (Institut national de la statistique et des études économiques). The data bases are provided on a grid of 200 x 200 m cells covering the entire country. Since cell division does not take into account geographical, natural or urban delimitations, we map cell data to the block level, weighting by surface overlap.

The set of variables we use are comprised of age and income distribution, percentage of household owners, percentage of apartments in a block.

For prices, we refer to MeilleursAgents's data bases, with 8 years of apartment transaction data and prices updates from September 2019.

3 Method

Our method combines geographical distance with distance on the Kohonen map into a geo-statistical estimation model of real estate prices

3.1 *SOM algorithm*

One may view SOM as a non-linear projection of the probability density function $p(x)$ of the high-dimensional input data vector x onto a two-dimensional grid of neurons (the Kohonen or *self-organized* map). The projection preserves topology, in the sense that neighboring observations in the input space are located next to each other on the grid.

HAC is then applied on prototypes to produce super-clusters that are useful to visualize and interpret the socio-spatial organization of a city on a simple geographical map.

3.2 *Spatial diffusion process*

We apply GWR (Geographically Weighted Regression) as a spatial diffusion model on housing transaction prices. GWR consists essentially in a classical regression where observations are weighted according to geographical distance to the point considered. The loss function reads:

$$\min \left(\sum_{i=1}^{n} w_i \epsilon_i \right), \tag{1}$$

where w_i is the weight and ϵ_i is the variable under consideration (*eg* price).

Weights

Writing $V_x(r)$ for the set of spatial neighbors of x (with r a distance in meters), we define **geographic weights** as :

$$w_{geo_i} = \exp\left(-\frac{||x - x_i||^2}{2\sigma^2} \right) \quad \text{for } i \in V_x(r) \tag{2}$$

However, two points located on both sides of the same street but in opposite clusters on the Kohonen map will be more different to one another than two points far away in geographic space but close to each other on the Kohonen map. To account for this, we introduce a distance d_{SOM} between clusters on the Kohonen map and we define **SOM weights** as :

$$w_{SOM_i} = \exp\left(-\gamma \, d_{SOM}(C(x_i), C(x))\right) \tag{3}$$

γ : a non-negative parameter $\qquad C(x_i)$: cluster of observation x_i
$C(x)$: cluster of observation x

We take into account only neighboring clusters on the SOM map:

$$N_{C_k} = \{C, C \text{ neighbor of } C_k \text{ on SOM map and } C_k = C(x) \text{ if } x \in C_k\}$$

Final weights are then :

$$w_i = w_{geo_i} \times w_{SOM_i}$$

And the loss function becomes:

$$\min \left(\sum_{i} w_i \epsilon_i \right) \quad \text{for } i \in N_x \qquad N_x = \left\{ x_i, x_i \in V_x(r) \text{ and } C(x_i) \in N_{C(x)} \right\} \tag{4}$$

4 Results

As an example, we show here the results obtained on the city of Les Lilas, comprising 43 rectangles and 67 blocks – a small city just outside central Paris, at the heart of the 12-million Île de France metropolitan region. We run 1000 iterations and choose the best one according to Ward's method.

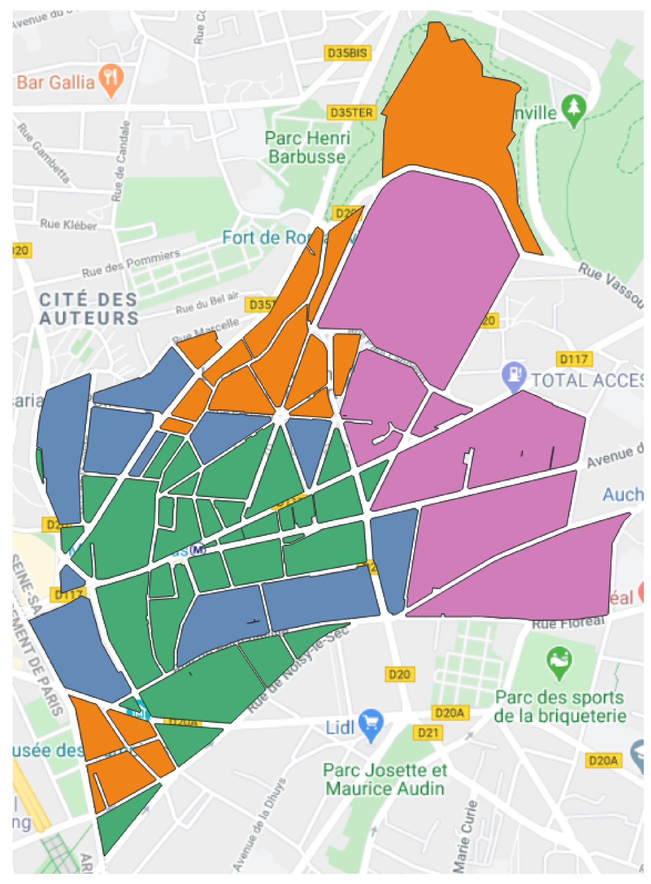

Figure 1: City of Les Lilas after applying SOM algorithm at block level (9 clusters and 4 Super-Clusters). We distinguish the suburban neighborhoods (Super-Cluster 4), the city center (Super-Cluster 1), and blocks composed of 60's building (Super-Cluster3)

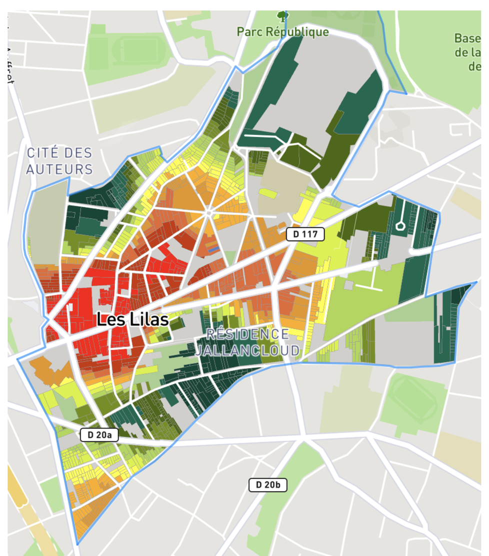

Figure 2: Price map of Les Lilas obtained with a simple GWR

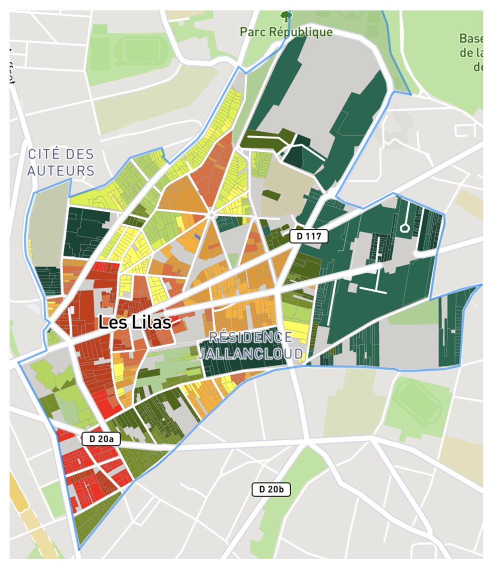

Figure 3: Price map of Les Lilas obtained with the new model

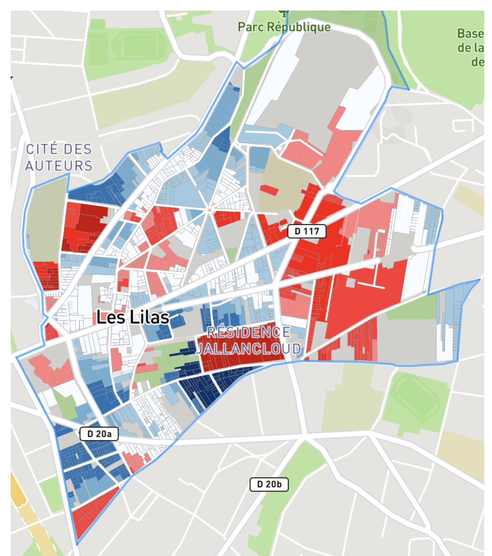

Figure 4: Difference in percentage betweend Figure 3 and Figure 4

While a simple GWR only captures spatial effects and, **with a delay**, neighbourhood quality information reflected in the housing prices, our method captures this latter information before prices of actual transactions come to reflect it — allowing one to forecast future trends at a fine-grained geographical scale. **Figure 4** is the combination of information unveiled by the Kohonen algorithm (as shown on **Figure 5**) and a single GWR.

The differences (**Figure 5**) between price indices produced by pure GWR (**Figure 3**) versus our combined method (**Figure 4**) reflect information that is hardly accessible by other means, especially if one does not or cannot have an intimate knowledge of the city under consideration: the type and quality of the buildings, the *atmosphere* of a neighbourhood, whether it will soon be a very sought-after neighbourhood or not, *etc*. Vast amounts of socio-demographical data work as a proxy for such information, provided one is able to harness it using machine learning methods.

Performance of the new method is measured by comparing errors on predicted prices. We use a simple GWR as reference model and compute the difference between the price predicted and actual transaction prices.
With a 10% improvement on median error, our method outperforms standard GWR.

5 Conclusions

- Using SOM allows one to gather information from a vast corpus of socio-economical data in order to bring out the socio-spatial structure and relate it to the dynamics of real estate prices
- Combining distances on the Kohonen map with geographical distances provides a better model of prices (at least in the real estate markets where we have tested our method)
- Our method captures a reality that is hard (or expensive) to obtain via other, human-resource based practices.

Open questions include that of the definition of mixing weights between the two types of distances used, and that of the interpretability of regression coefficients thus obtained.

References

[1] Julien Boelaert, Laura Bendhaiba, Madalina Olteanu, and Nathalie Villa-Vialaneix. *SOMbrero: An R Package for Numeric and Non-numeric Self-Organizing Maps*, pages 219–228. Springer International Publishing, 2014.

[2] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, Limited West Atrium, 2002.

[3] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012.

[4] Madalina Olteanu, Aurélien Hazan, Marie Cottrell, and Julien Randon-Furling. Multidimensional urban segregation: toward a neural network measure. *Neural Computing and Applications*, pages 1–13, 2019.

[5] Nicolas Thouvenin. *La formation des prix des logements anciens: les apports de la théorie des prix hédoniques*. PhD thesis, Université Paris Nanterre, 2005.