Machine Learning and Data-Driven Approaches in Spatial Statistics: A Case Study of Housing Price Estimation

Sarah Soleiman, Julien Randon-Furling, Marie Cottrell

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], Marie Cottrell[1]

[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

Abstract. 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 this cluster information into the spatial diffusion process of the GWR.

Keywords: SOM, Spatial Statistics, Real Estate Dynamics

1 Indroduction

Spatial modeling of the real estate prices is justified by the idea that prices are highly correlated within geographical areas. These issues, related to the valuation of real estate properties and the identification of real estate sub-markets, have been approached from different angles, from so-called hedonic models or more complex geo-statistical models [10], [8], [4], [9].

One of the most widely used models is the Geographically Weighted Regression (GWR). In particular, Bitter et al., in 2006, study the phenomenon of spatial heterogeneity in Arizona using the spatial expansion method and GWR [1]. The application of geographically weighted regression gives much better results than the spatial expansion method.

However, modern urban areas are very heterogeneous in nature, comprised of residential blocks, metro lines, parks and other geographical landmarks. Each

of these structures interfere with real estate prices in its neighborhood. These urban areas are also characterized by strong socio-spatial dissimilarities.

In this study, we describe socio-spatial patterns present in residential areas using public socio-economic data and incorporate this piece of information into real estate prices model.

2 Data

We used databases from Meilleurs Agents. Raw data are apartment's transactions described by the date (yyyy-mm-dd) of the transaction, it's exact location (land plot level), it's price and characteristics describing the apartment : the number of rooms, floor, area, price, presence/absence of an elevator. We use 5 years of those past transactions, between january 2014 and september 2019 in the city of Les Lilas only, which is a total of $N = 386$ past transactions.

Pre-processing of the data is two-fold : we first apply filters to avoid atypical apartments and get rid of the outliers (number of rooms must not exceed 15, area between 8 and 500 m^2), and then update the price of transactions adjusted for the real estate price index. The latter step enables to compare transactions at a fix time, we choose to set the model date at 2019-09-01.

Table 1: Description of the training set

Label	Variable	min	1st quantile	Median	3th quantile	max	Total
room_count	number of rooms	1	2	2	3	6	386
floor	floor	0	1	2	4	18	386
area	area (in m^2)	13	36	47	68	170	386
price	price (in €)	24000	200000	281000	400000	945000	386

47% of the transactions have a presence of an elevator in their description.

3 Spatial diffusion process

We apply GWR (Geographically Weighted Regression) [6], 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 location of the point considered.

To explain the GWR model, we follow the presentation done by M. Charlton in [6]. For each transaction i, $i = 1, ..., N$, one knows its price P_i (dependent

variable) and p independent variables $x_i^1, ..., x_i^p$, as well as its position in a geographical system. We use i to denote the transaction and its location itself.

The main equation related to a transaction at location u can be written :

$$P_i(u) = \beta_{0i}(u) + \beta_{1i}(u)x_i^1 + ... + \beta_{pi}(u)x_i^p + \varepsilon_i \tag{1}$$

Unlike a classical linear regression model, parameters $\beta_{0i}(u), \beta_{1i}(u), ..., \beta_{pi}(u)$ depend on location u.

Equation (1) represents a weighted regression model where the function to be minimized is :

$$\mathcal{E}(u) = \sum_{i=1}^{N} w_i(u)\varepsilon_i^2 \tag{2}$$

where

$$w_i(u) = \exp \frac{||i-u||^2}{2\sigma^2} \tag{3}$$

is the geographical weight, which takes into account the distance (in meter) between location u and locations of other transactions.

Parameter σ is chosen by cross-validation. In the following, $w_i(u)$ is denoted by $w_i^{GEO}(u)$.

The estimation of vector $\beta(u)$ is

$$\hat{\beta}(u) = (X^T W(u) X)^{-1} X^T W(u) P(u) \tag{4}$$

where $W(u) = diag(w_1(u), ..., w_N(u))$, X is the independent variables ($N \times p+1$) matrix (whose first column is only composed of 1) and P is the N-vector of prices.

However, two apartments located on both sides of the same street can present very different housing quality. Thus, considering the geographical distance only would bias the estimation. To account for this, we decide to use other variables that would describe the neighborhood atmosphere. We use public data from the French national office of statistics – INSEE (Institut national de la statistique et des études économiques). A clustering of location based on these extra variables will allow us to better qualify the apartments in the studied area.

4 GWR x SOM : Merging socio-spatial information into a spatial diffusion process

4.1 SOM

We choose to work with SOM as a clustering algorithm because it preserves topology : neighboring observations in the input space are located on the same or neighboring clusters on the SOM map.

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

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 city of Les Lilas is composed of 67 blocks in total. The set of $q = 13$ variables we use include socio-economic data such as age and income distribution, percentage of household owners, percentage of apartments in a block, etc.

Let us denote the number of blocks by n and $y_j \in R^q$ the feature vector of block j for $j = 1, ..., n$.

We run SOM on all y and get clusters $C_1, ..., C_K$ represented by prototypes $m_1, ..., m_K$.

Vector y is assigned to a cluster by :

$$y \in C_{k_0} \Leftrightarrow ||y - m_{k_0}|| = \min_{k=1,...,K} ||y - m_k|| \tag{5}$$

We denote by $C_{k_0(y)}$ the cluster y belongs to and $m_{k_0(y)}$ its prototype vector. The SOM distance between vectors y and y' is defined by :

$$d_{SOM}(y, y') = ||m_{k_0(y)} - m_{k_0(y')}|| \tag{6}$$

For any transaction i, we are looking for the block j that contains it, denoted by $j(i)$ and the extra vector y associated to i will be the vector $y_{j(i)}$.

We define SOM weights as :

$$w_i^{SOM}(u) = \exp\left(-\frac{d_{SOM}(y_{j(i)}, y_{j(u)})^2}{2\gamma^2}\right) \tag{7}$$

with γ a non-negative parameter chosen by cross validation. In the present work, we only take into account i and u such as they belong to neighboring clusters on the SOM map with radius 1.

4.2 Final model

Hence, final weights are defined by :

$$w_i^F(u) = w_i^{GEO}(u) \times w_i^{SOM}(u) \tag{8}$$

The function to be minimized is now :

$$\mathcal{E}(u) = \sum_{i=1}^{N} w_i^F(u)\varepsilon_i^2 \tag{9}$$

If we consider a neighboring point i that is on the exact same location of the current point u, then $w_i^{GEO}(u) = 1$. And if i and u belong to the same SOM

cluster then $w_i^{SOM}(u) = 1$. In this case $w_i^F(u) = 1$. These weights decrease to 0 when i is far from u in a geographical sense and/or in a socio-economic sense. An example of the impact of the new model on weights is shown on Figure 3 for a given location.

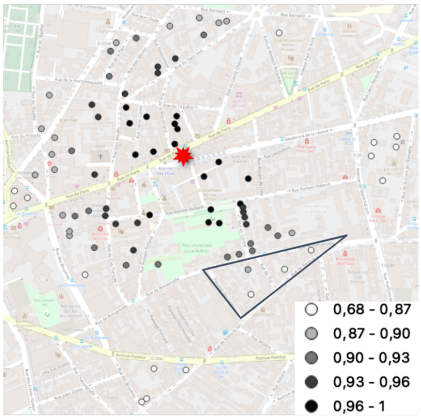

Fig. 1: Geographical weights for the considered point (red star).

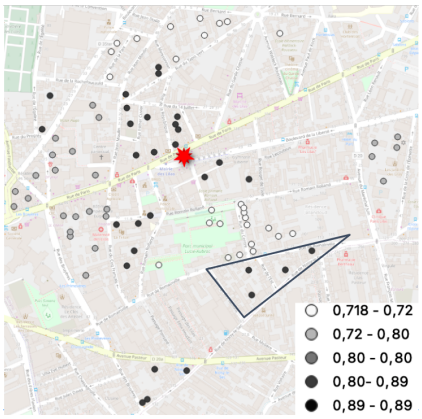

Fig. 2: SOM weights for the considered point (red star).

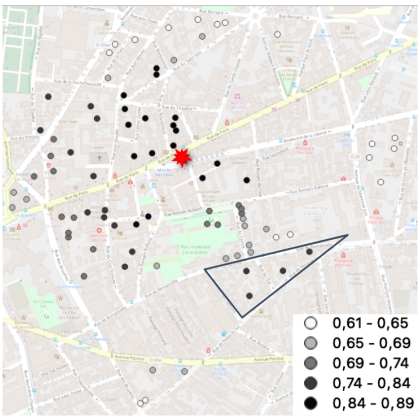

Fig. 3: Final weights for the considered point (red star). If we have a closer look at weights inside the black triangle, transactions located in this area will have more importance with our new model (Figure 3) than if we would have only take into account geographical distances (Figure 1). Indeed, neighborhood of this area and the one of the considered point (red star) appears to be similar (Figure 2).

5 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 of SOM algorithm and choose the one that minimizes the intra-classes inertia. We use a HAC algorithm to define 4 super-class that we project on the city map (see Figure 4).

We can represent the nine prototypes in Figure 5, which are 13-dimensional vector. We can see that prototypes of opposite units 1 and 9 on SOM map (Figure 5) are really different. Indeed, variable 2 "proportion of houses" has a low level for the prototype of unit 1, whereas it is has a high one for prototype of unit 9. Similarly, level of variable 3 "proportion of one-person households" is high for unit 1, and low for unit 9. And so on for the following variables. If we go from unit 1 to unit 9 passing through unit 5, we can notice that the prototype of unit 5 (which is at the center of SOM map) is pretty flat and thus, is similar to all prototypes of the map.

We set $\sigma = 600$ meters and γ equal to the median of the distances between prototypes, using cross-validation.

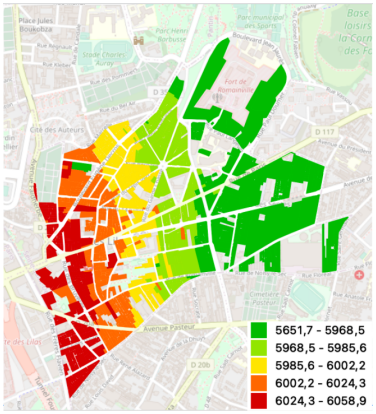

Fig. 6: Price map of Les Lilas obtained with a simple GWR.

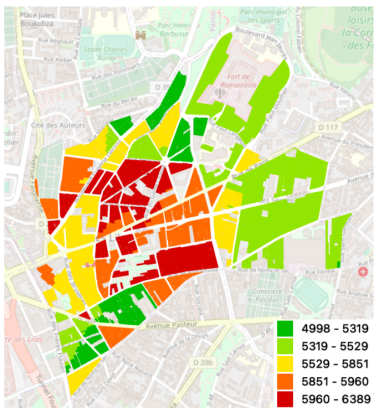

Fig. 7: Price map of Les Lilas obtained with the new model.

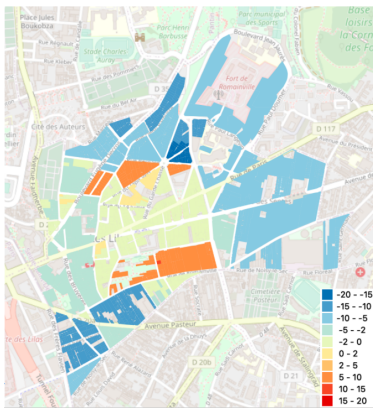

Fig. 8: Difference in percentage between Fig.6 and Fig.7

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 7** is the combination of information unveiled by the Kohonen algorithm and a single GWR.

The differences (**Figure 8**) between price indices produced by pure GWR (**Figure 6**) versus our combined method (**Figure 7**) reflect information that is

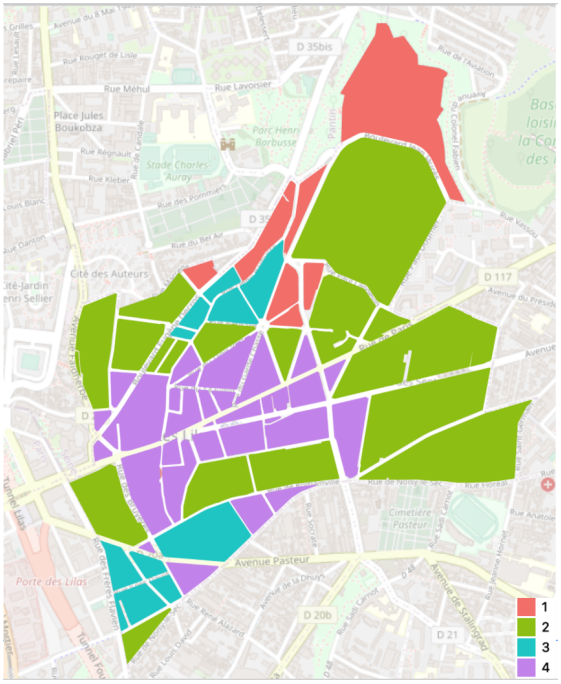

Fig. 4: City of Les Lilas after applying SOM algorithm at block level (3x3 SOM map and 4 Super-Clusters). We distinguish the suburban neighborhoods (Super-Cluster 1), the city center (Super-Cluster 4), and blocks composed of 60's building (Super-Cluster 2)

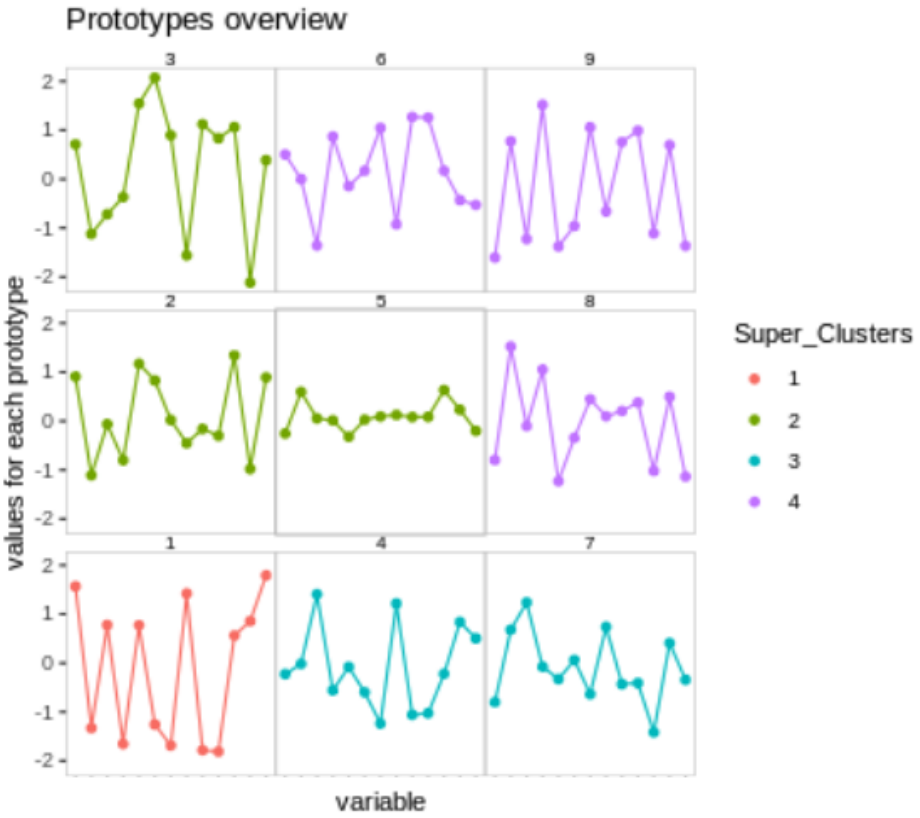

Fig. 5: Prototypes of the SOM map. In each unit, we represent a q-dimensional ($q = 13$) vector where variables are : the proportion of individuals between 25 and 65 years old, proportion of houses, proportion of one-person households, proportion of 5 persons or more households, proportion of households below the low income threshold, proportion of households in collective housing, proportion of owner households, total number of individuals, total number of households, average winsorized tax income of individuals, average area of the primary residence (in m^2).

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.

6 Performance

Performance of the new method is measured by computing errors on predicted prices. We use a simple GWR as reference model. At time t, $P_i(t)$ is the price of transaction i. We define the relative error such as :

$$E = \frac{\hat{P_i}(t) - P_i(t)}{P_i(t)} \tag{10}$$

where $\hat{P_i}(t)$ is the predicted price. We make the assumption that the real estate market doesn't have a significant fluctuation within 6 months and thus take into account transactions between 2019-09-01 and 2020-02-01. The test set is composed of 70 observations.

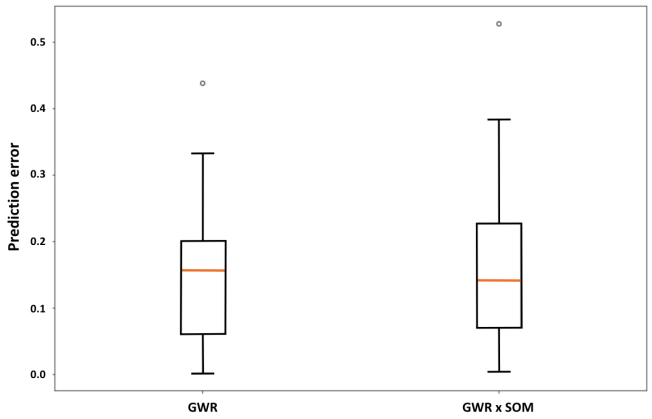

Fig. 9: Boxplots of prediction error of a GWR and of the new model mesured on the test set (n=70)

7 Conclusion

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. Christopher Bitter, Gordon F Mulligan, and Sandy Dall'erba. Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, 9, 2007.
2. 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.
3. A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, Limited West Atrium, 2002.
4. Thibodeau Goodman. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12:181–201, 2003.
5. T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012.
6. Chris Brunsdon Martin Charlton, Alexander Fotheringham. Geographically weighted regression. *Journal of the Royal Statistical Society Series D (The Statistician)*, pages 5–6, 2009.
7. 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.
8. Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82:34–55, 1974.
9. Ay Se Can and Isaac Megbolugbe. Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, 14:203–222, 1997.
10. Thouvenin. *La formation des prix des logements anciens, les apports de la théorie des prix hédoniques*. Books on Demand, 2010.