Machine Learning and Data-Driven Approaches in Spatial Statistics

A case study of housing price estimation

Sarah Soleiman, Julien Randon-Furling,Thomas Lefebvre

1 SAMM, Université Paris Panthéon-Sorbonne – P2M2 (FR2036) CNRS, Paris, France
2 MeilleursAgents, 7 Boulevard Hausmann, 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 statis-tique 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 de-limitations, 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:

$$\min \sum w_{ik} \gamma ||x_i - x_k||^2$$

where $w_{ik}$ is the weight and $x_k$ the variable under consideration (eg. price).

Weights

Writing $V_{ge}(r)$ for the set of spatial neighbors of $x$ (with $r$ a distance in meters), we define geometric weights as:

$$w_{geo} = \exp\left(-\frac{|x-x_i||x-x_j||^2}{2\sigma^2}\right)$$

for $i, j \in V_{ge}(r)$

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 $d_{SOFM}$ between clusters on the Kohonen map and we define SOM weights as:

$$w_{SOFM} = \exp(\gamma d_{SOFM}(C(x_i), C(x_j)))$$

where $C(x_i)$ cluster of observation $x_i$.

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

$$N_{k} = \{C, C \text{ neighbor of } C_{k}\}$$

Final weights are then:

$$w_{i} = w_{geo} \times w_{SOFM}$$

And the loss function becomes:

$$\min \sum_{i \in N_{k}} w_{i} \gamma \left||x_i - x_k||^2\right.$$