SpICE: An interpretable method for spatial data

Natalia da Silva[1], Ignacio Alvarez-Castro[1], Leonardo Moreno[1], and Andrés Sosa[1]

[1]Instituto de Estadística (IESTA), Universidad de la República, Montevideo, Uruguay.

Introduction

This paper is focused on the individual conditional expectation plot (ICE-plot), a model-agnostic method for interpreting statistical learning models and combining them with spatial information. An ICE-plot extension is proposed in which spatial information is used as a restriction to define spatial ICE (SpICE) curves. Spatial ICE curves are estimated using real data in the context of an economic problem concerning property valuation in Montevideo, Uruguay. Understanding the key factors that influence property valuation is essential for decision-making, and spatial data play a relevant role in this regard.

ICE curves in spatial problems

In supervised problems, the goal of statistical learning models is to approximate $\mathbb{E}(Y|X = x) = f(x) \approx \hat{f}(x)$
The ICE-plot (Goldstein et al. 2015) is proposed to visualise the dependence of the prediction on a feature for each observation separately, then allowing heterogeneous effects. For each observation, ICE curve is computed as

$$\hat{f}_S^{(i)}(x_S) = \hat{f}(x_S, x_C^{(i)}). \tag{1}$$

Spatial ICE (SpICE) curves

In clustering ICE curves, the key idea is to group observations with similar predictor effects on the response; thus, considering both the level and variation is important. Then, the Sobolev $W^{1,2}(\mathbb{R})$ distance (Adams and Fournier 2003) is used:

$$d_{Sob}(i,j) = \sqrt{\int |\hat{f}_S^{(i)}(x) - \hat{f}_S^{(j)}(x)|^2 dx + \int |\hat{f}'^{(i)}_S(x) - \hat{f}'^{(j)}_S(x)|^2 dx}. \tag{2}$$

Additionally, each ICE curve is associated with a specific observation. In spatial problems, it is possible to link such curves with location information. Combining these two sources of information might improve the interpretability of ICE curves.

Chavent et al (2018) propose an algorithm to construct clusters of multivariate data with spatial restrictions of contiguity. A the pseudo-inertia in cluster k can be defined as

$$I_\alpha(C_k^\alpha) = (1-\alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} d_{0,ij}^2 + \alpha \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} d_{1,ij}^2,$$

where $d_{0,ij}$ and $d_{1,ij}$ represent dissimilarities between observations i and j. Then, a Ward-like method (Ward 1963) is used for the construction of the clusters.

In this paper, this clustering algorithm is adapted to cluster ICE curves (functional data). Mainly, Sobolev distance is used to compute dissimilarities among pre-smoothed ICE curves, i.e $d_{0,ij} = d_{sob}(i,j)$ (see Equation (2)). In addition to ICE information, the geographical distance between observations is used $d_{1,ij}$.

SpICE package

The `SpICE` R package implements the computation and visualisation of SpICE curves.
The main function is `cl_sobcurve` which compute

- $D0$: Sobolev distance among ICE curves
- $D1$: Geographical distance between observations
- Create clusters with `geoClusters::hclustgeo`

Other functions:

- `plot_alpha` compute optimum value for alpha and plot results for a gange of clusters (based on geoClusters)
- `plot_clcurve` clustered ICE curves
- `plot_clcoord` geographical location (map) of clustered ICE curves

Apartment prices in Montevideo

The data are obtained from an eCommerce platform called Mercado Libre (`https://www.mercadolibre.com.uy`), where apartments and houses are offered for sale and for rent.

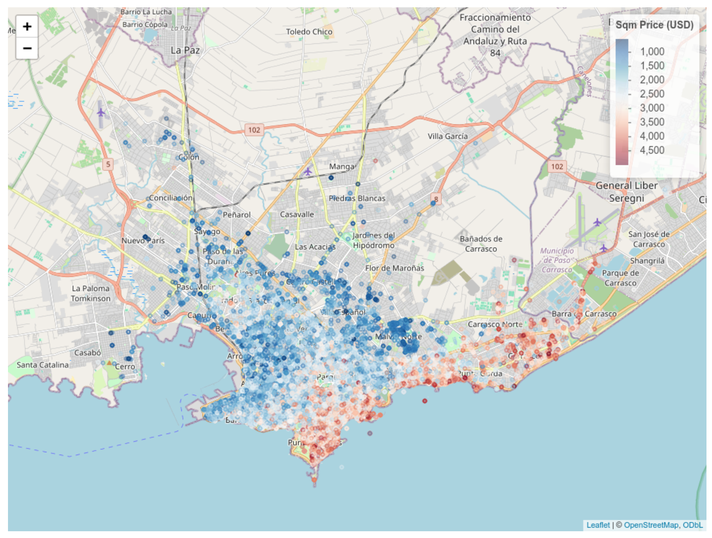

Figure 1: Montevideo apartment prices

Several models were trained using the automatic machine learning (`autoML`) procedure from the `h2o` R package (LeDell et al. 2023) to predict the apartment price as an alternative to classical methods. Table 1 shows the performance measures for the selected predictive models.

Table 1: Predictive performance measures by model

	model	rMSE	R2	MAEo	MAPEo
1	stackedensemble	0.14	0.81	242.68	9.58
2	drf	0.14	0.80	251.08	9.94
3	xgboost	0.15	0.80	254.48	10.06
4	deeplearning	0.20	0.63	395.84	15.18
5	glm	0.22	0.55	427.72	17.13

The apartment area (`larea_apt`) is the most relevant feature in every model. SpICE curves to describe effect of apartment area on apartment price are computed.

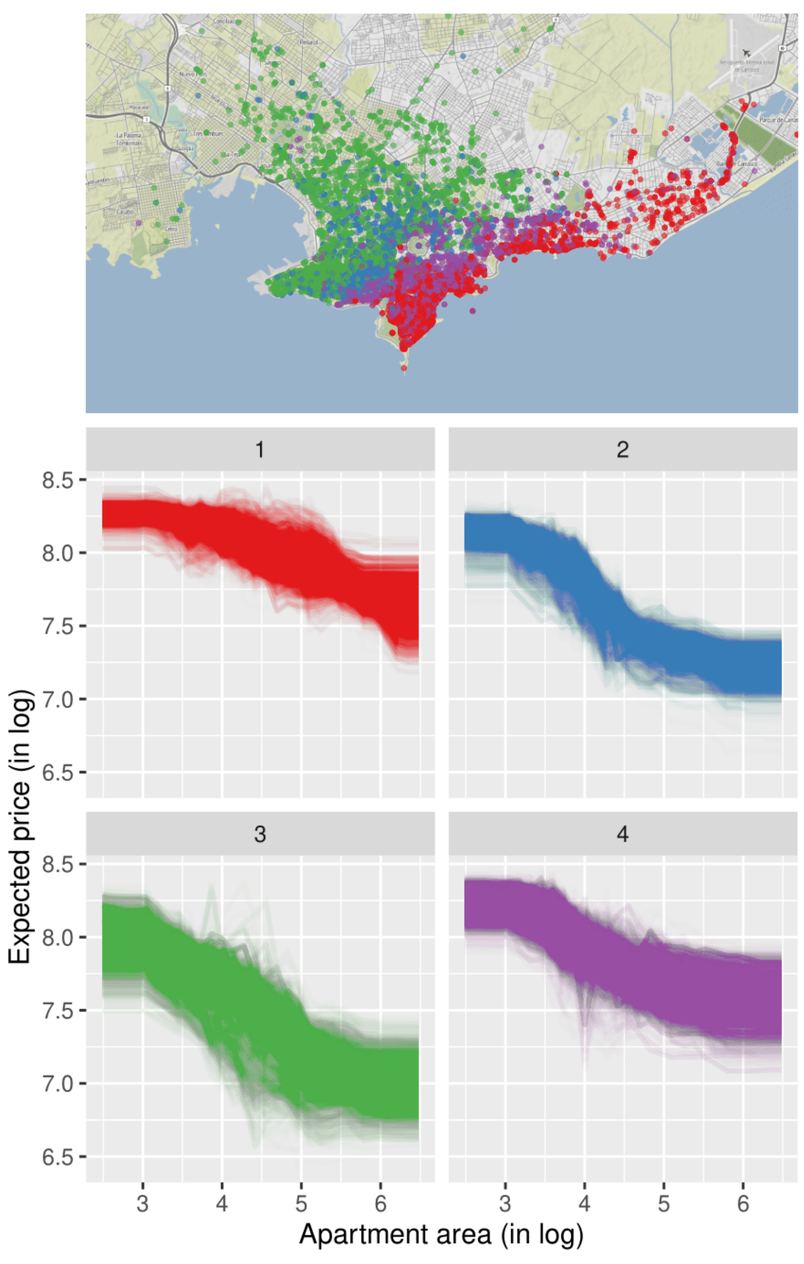

Figure 2: Geographical locations of clusters and associated SpICE curves

Paper accepted in COST, material to reproduce the paper and its results is available at: https://github.com/natydasilva/SpICE-COST

email: natalia.dasilva@fcea.edu.uy