A space-time-categorical local linear smoother for predicting house prices
Ghislain Geniaux, Davide Martinetti

To cite this version:
Ghislain Geniaux, Davide Martinetti. A space-time-categorical local linear smoother for predicting house prices. 2. Conference on Econometrics for Environment (CE2-2018), Dec 2018, Nador, Morocco. 35p. hal-02952109

HAL Id: hal-02952109
https://hal.inrae.fr/hal-02952109
Submitted on 29 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A space-time-categorical local linear smoother for predicting house prices

Ghislain Geniaux, INRA Ecodeveloppement

Co-author, D. Martinetti, INRA Biosp

EC², Nador, 2018
UrbanSIMUL (http://urbansimul.fr) is a big geohistoric database at parcel scale that supports a decision tool for managing land supply and for designing urban policy (zoning) :

- Predict urban sprawl
- Predict building capacity of parcels
- Predict land prices
Two main methodological issues for spatial model with big data:
- Spatial Discrete Choice model (Martinetti & Geniaux RSUE 2017, ProbitSpatial R Package)
- Dealing simultaneously with spatial dependence, spatial heterogeneity and non-linearity (Geniaux & Martinetti RSUE 2017, mgwrsar R Package)
Market segmentation / submarket

A housing (resp. land) submarket can be defined, roughly, as a set of dwellings (resp. lands) that are reasonably close substitutes of one another, but there are not substitute of dwellings (resp. lands) belonging to other submarkets.
Islam and Asami (2009) → 3 approaches:

1. hedonic price models are used to cluster the properties that are similar with respect to a bundle of qualitative characteristics, such as lot size, number of rooms and bathrooms, garden, parking slot, etc. (Grigsby et al., 1986; Kauko, 2002; Leishman, 2001; Schnare and Struyk, 1976; Tu and Goldfinch, 1996; Tu, 1997)
2. On the other hand, housing market can be analyzed with respect to the spatial distribution of properties and other spatial features. In this context, spatial proximity and clustering are the prime determinants of submarket’s definition (Gallet, 2004; Goodman, 1978; Goodman and Thibodeau, 1998, 2003).
3. There exist mixed approaches that consider both topographic and quality segmentation, sometimes referred as hybrid-related submarkets, (O’Sullivan and Gibb, 2008).
OUR PROPOSAL

Since we postulate a strong dependence between house quality and its location, we cannot rely on two-stage models such as the ones proposed by (Goodman and Thibodeau, 2007; O’Sullivan and Gibb, 2008; Tu, 1997).
OUR PROPOSAL

We prefer instead a smoother approach, where the hedonic regressions coefficients can vary across space, time and submarkets.

Extended version of geographically-weighted regression with spatial dependence, namely MGWR-SAR, Geniaux and Martinetti (2017)
A space-time-categorical local linear smoother for predicting house prices

Local linear regression framework (Cleveland, 1979; Hastie and Tibshirani, 1990, 1993)

\[ Y_i = \beta(u_i, v_i; h)X_i + \epsilon_i, \]

Each Local Regression for point \( i \) is based on a local subsample
A space-time-categorical local linear smoother for predicting house prices

\[ W_{ij}Y = \beta_i W_{ij}X + \varepsilon \]

GWR with fixed spatial kernel

\[ W_{ij}Y = \beta_i W_{ij}X + \varepsilon \]

x Regression point
• Data point

A space-time-categorical local linear smoother for predicting house prices

GWR with adaptive kernel

x Regression point
● Data point

Each local subsample is defined by a kernel that produces a vector of weights based on spatial proximity between $i$ and $j$:

$$
\mathbf{w}_{ij} = K(d_{ij}, h)
$$

where $d_{ij}$ is a metric of proximity between $i$ and $j$ and $h$ a bandwidth.

Various kernels $K()$ can be used, but the main issue is to choose a suitable bandwidth $h$ using Cross Validation (leave-one-out) or Plug-in Methods.
A space-time-categorical local linear smoother for predicting house prices

OUR PROPOSAL


+ Li and Racine 2010 « Smooth varying-coefficient estimation and inference for qualitative and quantitative data ». \textit{Econometric Theory 26 (06) hereafter LR2010}
Add time differences to the kernel:

\[ w_{ij} = K(d_{ij}, T_{ij}; h_d, h_t) \]

Huang et al., 2010; Wrenn and Sam, 2014; Fotheringham et al., 2015

Wu et al. (2014) proposed a GWR technics with spatial autocorrelation,

Wei et al. (2017) proposed to extend GWR using spatial SUR models in order to explore spatio-temporal heterogeneity
Why not choosing a full non-parametric framework?

Because optimization of bandwidth is too long and precludes such option for moderate and big samples as soon as you have more than 3-5 covariates.

To provide results easier to interpret and to share with practitioners, notably using maps/time and map/housing submarkets: space + time + market segment
Why choosing categorical submarkets:

- Because by merging all submarkets in a global local linear regression, it allows to increase the amount of information used in each submarket for taking into account unobserved heterogeneity. It’s what we call « shared spatial heterogeneity ».
Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial autorcorrelation

\[ y = \beta_c X_c + \epsilon_i \quad \text{(OLS)} \]
\[ y = \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(GWR)} \]
\[ y = \beta_c X_c + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR)} \]
\[ y = \lambda W y + \beta_c X_c + \epsilon_i \quad \text{(SAR)} \]
\[ y = \lambda W y + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(0, 0, k)) \]
\[ y = \lambda W y + \beta_c X_c + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(0, k_c, k_v)) \]
\[ y = \lambda(u_i, v_i) W y + \beta_c X_c + \epsilon_i \quad \text{(MGWR-SAR}(1, k, 0)) \]
\[ y = \lambda(u_i, v_i) W y + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(1, 0, k)) \]
\[ y = \lambda(u_i, v_i) W y + \beta_c X_c + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(1, k_c, k_v)) \]
Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial autorcorrelation

\[ y = \beta_c X_c + \epsilon_i \quad \text{(OLS)} \]

\[ y = \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(GWR)} \]

\[ y = \beta_c X_c + \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(MGWR)} \]

\[ y = \lambda Wy + \beta_c X_c + \epsilon_i \quad \text{(SAR)} \]

\[ y = \lambda Wy + \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(MGWR-SAR(0, 0, k))} \]

\[ y = \lambda Wy + \beta_c X_c + \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(MGWR-SAR(0, k_c, k_v))} \]

\[ y = \lambda(u_i, v_i)Wy + \beta_c X_c + \epsilon_i \quad \text{(MGWR-SAR(1, k, 0))} \]

\[ y = \lambda(u_i, v_i)Wy + \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(MGWR-SAR(1, 0, k))} \]

\[ y = \lambda(u_i, v_i)Wy + \beta_c X_c + \beta_v(u_i, v_i)X_v + \epsilon_i \quad \text{(MGWR-SAR(1, k_c, k_v))} \]
Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial auto-correlation

\[ y = \lambda W y + \beta_c X_c + \beta_v (u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(0, k_c, k_v)) \]

+ GENERAL KERNEL PRODUCT of Li and Racine 2010
Spatial, temporal and categorical kernel are combined by means of the Generalized Kernel Product function:

\[ GPK(i, j) = K(d_{ij}, hs) \times K(T_{ij}, ht) \times K(S_i, \rho) \]
The categorical kernel (Aitchison and Aitken, 1976; Li and Racine, 2010) takes the following form:

\[ K(S_i, \rho) = \begin{cases} 
1 & \text{if } S_j = S_i = s \\
\rho_s & \text{if } S_j \neq S_i = s
\end{cases} \]
Planed Extensions of GM2017

\[ Y_i = \lambda WY + \beta_c X_c \]
\[ + \beta_v ((u_i, v_i), T, S; h_d, h_t, \rho_s) X_v + \epsilon_i, \]

\[ Y_i = \sum_s \lambda_s WY + \sum_s \beta^s_c X_c \]
\[ + \beta_v ((u_i, v_i), T, S; h_d, h_t, \rho_s) X_v + \epsilon_i, \]

\[ Y_i = \lambda ((u_i, v_i), T, S; h_d, h_t, \rho_s) WY \]
\[ + \beta_v ((u_i, v_i), T, S; h_d, h_t, \rho_s) X_v + \epsilon_i, \]
Monte Carlo design inspired by GM2017:

• (x,y) locations drawn from uniform [0,1]
• $W \rightarrow$ 4 nearest-neighbours, row normalized
• 4 covariates including intercept, some spatially correlated,
• Mixed $\beta$: some spatially varying $\beta_v(u_i,v_i)$ and some constant over the space $\beta_c$
Monte Carlo: Submarket simulation

4 simulated submarkets cases:

- Different Beta for 4 submarkets for observable covariates
- One spatially correlated covariate is not observed for all submarkets
  → introduce additional Spatial Heterogeneity + dependence between submarkets
- One submarket is designed to be independent
  → 3 dependent submarkets + one fully independent submarket
2 simulated submarkets case:

- One case with different Beta
- One case with same Beta

→ false submarket segmentation
Monte Carlo results

Results based on this model:

\[ Y_i = \lambda((u_i, v_i), T, S; h_d, h_t, k_s)WY + \beta_v((u_i, v_i), T, S; h_d, h_t, \rho_s)X_v + \epsilon_i, \]

We show that:

• Bandwidths \( \rho_s \) for the independent submarket is closed to zero and \( \rho_s >> 0 \) for other submarkets (4 submarkets case)
• Bandwidth \( \rho_s \) for “false” submarket segment is closed to one (2 submarkets case)
• \( \beta_i \) and spatial parameter \( \lambda_i \) appears unbiased
RESULTS for land Sales Data

- SAMPLE:
  - geolocalized sales in southern France (2007-2015), fiscal administration
  - 1531 sales of Developpable land
  - 8011 sales of Agricultural land
  - 1330 sales of Other types of land
  - 13057 Single Family House with Garden

→ 4 potential submarkets
• A lot of covariates from various databases form GIS UrbanSIMUL project (http://urbansimul.fr)
• Information about parcels, owners, distance to …
• Selection of covariates using piecewise linear model (MARS model) for each submarket.
## RESULTS for land Sales Data

### Agricultural lands N=8011

<table>
<thead>
<tr>
<th>Model</th>
<th>kernel type</th>
<th>bandwidth</th>
<th>LOO-CV*</th>
<th>In,sample RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS (piecewise linear)</td>
<td>none</td>
<td>none</td>
<td>1.4961</td>
<td>1.4234</td>
</tr>
<tr>
<td>SAR (W matrix)</td>
<td>nn</td>
<td>8</td>
<td>1.2710</td>
<td>1.2642</td>
</tr>
<tr>
<td>GWR (space kernel)</td>
<td>Gauss_adapt</td>
<td>1600</td>
<td>1.2810</td>
<td>1.2655</td>
</tr>
<tr>
<td>GWRSAR (W matrix)</td>
<td>bisq_adapt</td>
<td>14</td>
<td>1.1621</td>
<td>1.0896</td>
</tr>
<tr>
<td>GWRSARX (W matrix)</td>
<td>bisq_adapt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(space kernel)</td>
<td>Gauss_adapt</td>
<td>1400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWRSARX (W matrix)</td>
<td>bisq_adapt</td>
<td>14</td>
<td>1.0613</td>
<td>1.0310</td>
</tr>
<tr>
<td>(surface kernel)</td>
<td>bisq_adapt</td>
<td>3600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(space kernel)</td>
<td>Gauss_adapt</td>
<td>1400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## RESULTS for land Sales Data

### Developpable lands N=1531

<table>
<thead>
<tr>
<th>Model</th>
<th>kernel type</th>
<th>bandwidth</th>
<th>LOO-CV*</th>
<th>In,sample RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS (piecewise linear)</td>
<td>none</td>
<td>none</td>
<td>1.2941</td>
<td>1.1343</td>
</tr>
<tr>
<td>SAR (W matrix)</td>
<td>nn</td>
<td>2</td>
<td>1.0412</td>
<td>0.9776</td>
</tr>
<tr>
<td>GWR (space kernel)</td>
<td>Gauss_adapt</td>
<td>140</td>
<td>0.9790</td>
<td>0.9049</td>
</tr>
<tr>
<td>GWRSAR (W matrix)</td>
<td>bisq_adapt</td>
<td>3</td>
<td>0.9714</td>
<td>0.8929</td>
</tr>
<tr>
<td>(space kernel)</td>
<td>Gauss_adapt</td>
<td>135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWRSARX (W matrix)</td>
<td>bisq_adapt</td>
<td>3</td>
<td>0.8664</td>
<td>0.7487</td>
</tr>
<tr>
<td>(surface kernel)</td>
<td>bisq_adapt</td>
<td>950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(space kernel)</td>
<td>Gauss_adapt</td>
<td>270</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LOO-CV** GWRSARX

Ind. Developpable+Agricultural Land= 1.0432
# RESULTS for land Sales Data

Developpable lands and Agricultural Lands N= 9542

<table>
<thead>
<tr>
<th>Model</th>
<th>kernel type</th>
<th>bandwidth</th>
<th>LOO-CV*</th>
<th>In, sample RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWRSARX (W block diag matrix)</td>
<td>bisq_adapt</td>
<td>3 - 14</td>
<td>0.9965</td>
<td>0.7487</td>
</tr>
<tr>
<td>(surface kernel)</td>
<td>bisq_adapt</td>
<td>1500 - 3900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(space kernel</td>
<td>Gauss_adapt</td>
<td>2400 - 2400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Segment kernel</td>
<td>LR2010</td>
<td>0.03 - 0.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LOO-CV GWRSARX**
Developpable+Agricultural Land= 0.9965
Next Step for « shared spatial heterogeneity » idea

\[
\begin{pmatrix}
\rho_{11} & \rho_{12} & \rho_{13} \\
\rho_{21} & \rho_{22} & \rho_{23} \\
\rho_{31} & \rho_{12} & \rho_{33}
\end{pmatrix}
\]
Introducing mgwrsar R Package