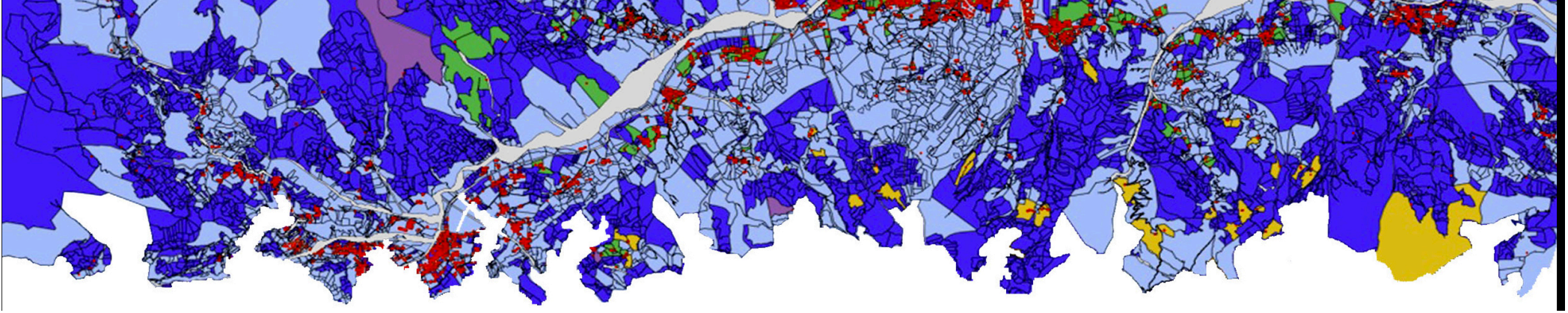
A space-time-categorical local linear smoother for predicting land/house price

Ghislain Geniaux, Davide Martinetti



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

Ghislain Geniaux and Davide Martinetti
INRA UR 767 Ecodéveloppement

17 June 2017

EcoStat 2017, Honk Kong

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

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.

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

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)

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

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).

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

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).

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

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

OUR PROPOSAL

Geniaux and Martinetti(2017) « A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models » *RSUE, forthcoming* **hereafter GM2017**

+

Li and Racine 2010 « Smooth varying-coefficient estimation and inference for qualitative and quantitative data ». *Econometric Theory 26 (06)*
hereafter LR2010

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

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

$$w_{ij} = K(d_{ij}, h)$$

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.

ADD TIME

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

ADD OTHER DIMENSIONS OF ATTRIBUTE'S SPACE ?

- Why not choosing a full non-parametric framework ?
  - Because convergence time 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

ADD OTHER DIMENSIONS OF ATTRIBUTE'S SPACE ?

- 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 heterogenity ».

Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial autorcorrelation

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

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

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

$$y = \lambda W y + \beta_c X_c + \epsilon_i \quad \text{(SAR)}$$

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

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

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

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

$$y = \lambda(u_i, v_i) W y + \beta_c X_c + \beta_v(u_i, v_i) X_v + \epsilon_i \quad \text{(MGWR-SAR}(1, k_c, k_v)\text{)}$$

Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial autorcorrelation

$$
\begin{aligned}
y = & & \beta_c X_c & & & + \epsilon_i & \text{(OLS)} \\
y = & & & & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(GWR)} \\
y = & & \beta_c X_c & + & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(MGWR)} \\
y = & \lambda W y & + \beta_c X_c & & & + \epsilon_i & \text{(SAR)} \\
y = & \lambda W y & & + & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(MGWR-SAR}(0,0,k)\text{)} \\
y = & \lambda W y & + \beta_c X_c & + & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(MGWR-SAR}(0,k_c,k_v)\text{)} \\
y = & \lambda(u_i, v_i) W y & + \beta_c X_c & & & + \epsilon_i & \text{(MGWR-SAR}(1,k,0)\text{)} \\
y = & \lambda(u_i, v_i) W y & & + & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(MGWR-SAR}(1,0,k)\text{)} \\
y = & \lambda(u_i, v_i) W y & + \beta_c X_c & + & \beta_v(u_i, v_i) X_v & + \epsilon_i & \text{(MGWR-SAR}(1,k_c,k_v)\text{)}
\end{aligned}
$$

Extending mgwrsar R package (Geniaux Martinetti 2017)
Mixed GWR + 2SLS for spatial autorcorrelation

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

+ GENERAL KERNEL PRODUCT of Li and Racine 2010

ADDING TIME AND HOUSING SUBMARKET IN THE KERNEL

Spatial, temporal and categorical kernel are combined by means of the Generalized Kernel Product function:

$$GPK(i,j) = K(d_{ij}, hs) * K(T_{ij}, ht) * K(S_i, \rho)$$

ADDING TIME AND HOUSING SUBMARKET IN THE KERNEL

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

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

$$Y_i = \sum_s \lambda_s WY + \sum_s \beta_c^s X_c + \beta_v((u_i, v_i), T, S; h_d, h_t, \rho_s) X_v + \epsilon_i \,,$$

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

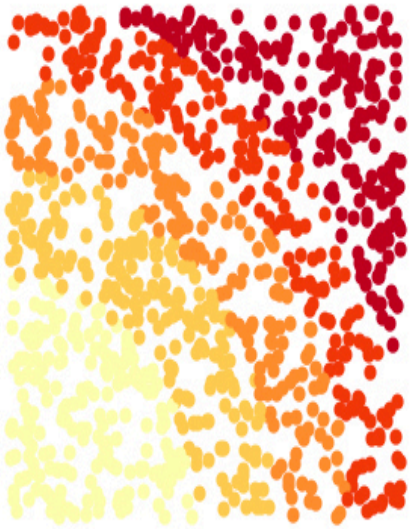
Monte Carlo

Monte Carlo design inspired by GM2017:

- (x,y) locations drawn from uniform [0,1]
- W → 4 nearest-neighbours, row normalized
- 4 covariates including intercept, some spatially correlated,
- Mixed ß : some spatially varying $\beta_v(u_i,v_i)$ and some constant over the space $ß_c$

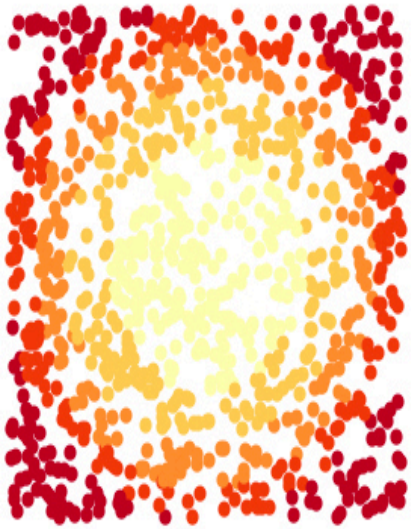
Beta0

- under –0.2
- –0.2 – –0.05
- –0.05 – 0.08
- 0.08 – 0.19
- over 0.19



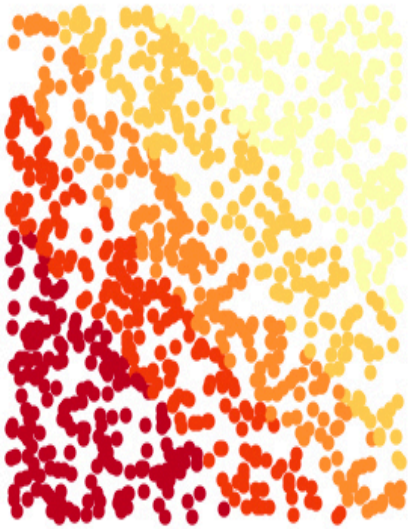
Beta1

- under 0.3
- 0.3 – 0.46
- 0.46 – 0.58
- 0.58 – 0.68
- over 0.68



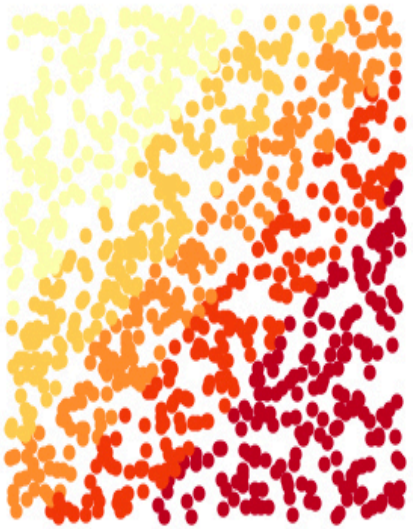
Beta2

- under 0.6
- 0.6 – 0.89
- 0.89 – 1.13
- 1.13 – 1.39
- over 1.39



Beta3

- under –1.37
- –1.37 – –1.1
- –1.1 – –0.89
- –0.89 – –0.64
- over –0.64



Monte Carlo: Submarket simulation

4 simulated submarkets cases:

- Different Beta for 4 submakets for observable covariates
- One spatially correlated covariate is not observed for all submarkets

→ introduce additionnal Spatial Heterogeneity + dependence between submarkets

- It has null Beta value for one submarket

→ 3 dependent submarkets + one fully independent submarket

Monte Carlo: Submarket simulation

2 simulated submarkets case:

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

→ **false submarket segmentation**

Monte Carlo preliminary results

Results based on this model:

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

- Bandwidths ρ_s for the independent submarket is closed to zero and $\rho_s > 0$ for other submarkets (4 submarkets case)
- Bandwidth ρ_s for ‘’false’’ submarket segment is closed to one (2 submarkets case)
- β_i and spatial parameter λ_i unbiased

RESULTS for Developable land Sales Data

- SAMPLE : 7 000 geolocalized sales of Developable land between 100 m^2 and 50000 m^2 in southern France (2007-2015)

- Simple didactic model:

log(prix)= surface + date + ndist_cbd + ndist_dense_urban_area + ndist_road | (u_i,v_i),T,S

RESULTS for Developable land Sales Data

- 4 potential submarkets based on different ways of classifying Developpable Land Sales form different data sources:
    - from fiscal administration database (1)
    - from fiscal administration database (2)
    - from digitalized Land Use Plan,
    - from SAFER (french agricultural administration that monitors undevelopped land sales).

RESULTS on Developable land Sales

MODEL	MAPE	CV (leave-one-out)	SSR
SAR	32.36 %	1412.12	1286.27
GWR	30.71 %	1376.69	1213.11
GWR + TIME	30.75 %	1377.084	1217.09
GWR + Spatial Dependence (GWRSAR)	29.54 %	1263.224	1164.58
GWRSAR + endogenous categorical segmentation	24.11 %	862.38	947.72

W = Adaptative bisquare (with 10 first spatial neigbours belonging to the same submarket)

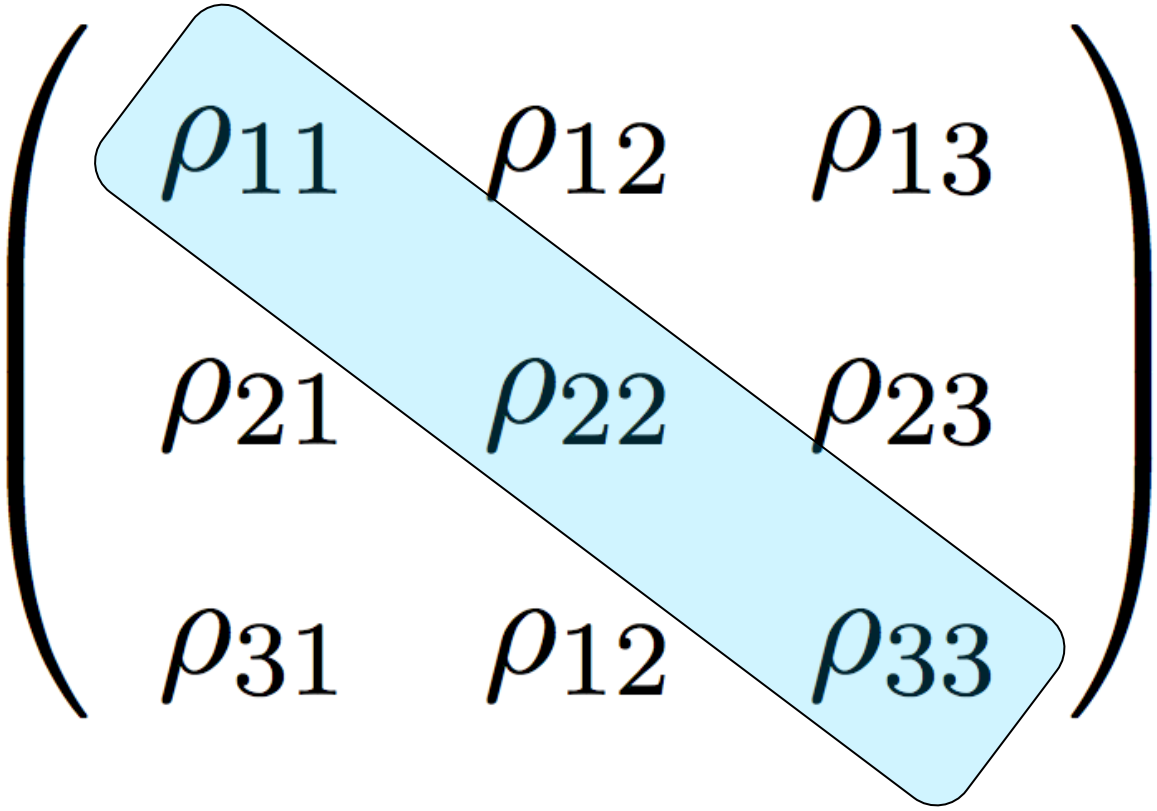
Optimal bandwidth (CV criteria) with adaptative gaussian kernel for distance and LR 2010 kernel for submarkets :

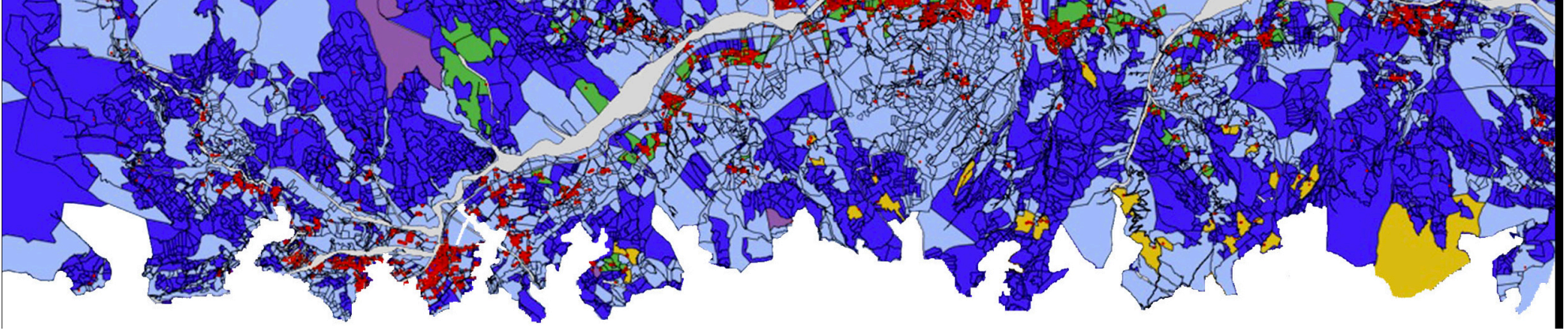
$H^* = (h_d = 520$ neighbours, $\rho 1 = 0.821$, $\rho 2 = 0.662$, $\rho 3 = 0.891$, $\rho 4 = 0.085)$

RESULTS on Developable land Sales

MODEL	MAPE	CV (leave-one-out)	SSR
SAR	32.36 %	1412.12	1286.27
GWR	30.71 %	1376.69	1213.11
GWR + TIME	30.75 %	1377.084	1217.09
GWR + Spatial Dependence (GWRSAR)	29.54 %	1263.224	1164.58
GWRSAR + categorical submarkets	24.11 %	862.38	947.32
GWRSAR with independent estimation of each segment	28.65 %	1093.35	828.35

Next Step for « shared spatial heterogenity » idea





Thanks for your attention