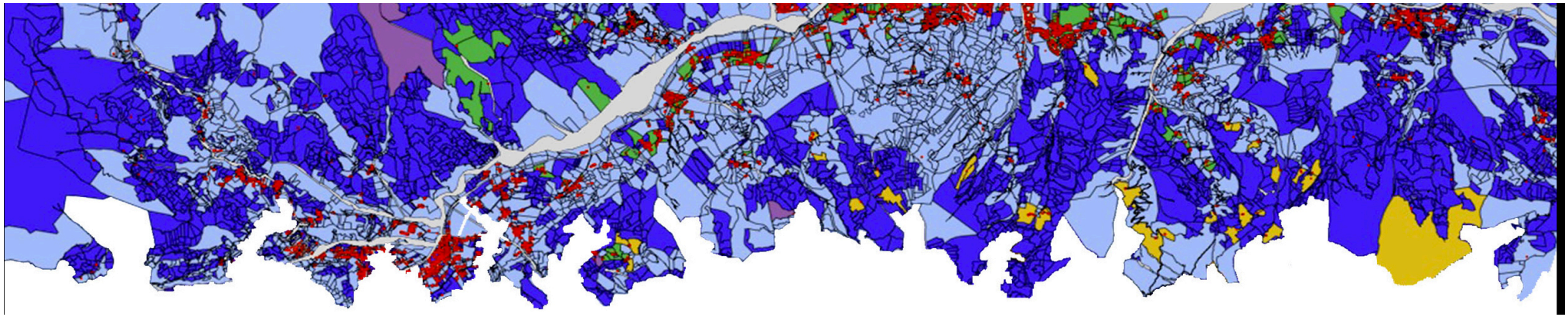
A space-time-categorical local linear smoother for predicting house prices

Ghislain Geniaux, Davide Martinetti



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

Ghislain Geniaux,
INRA Ecodeveloppement

Co-author, D. Martinetti, INRA Biosp

EC^2, Nador, 2018

Research problematics around UrbanSIMUL project

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

Research problematics around UrbanSIMUL project

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)

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

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

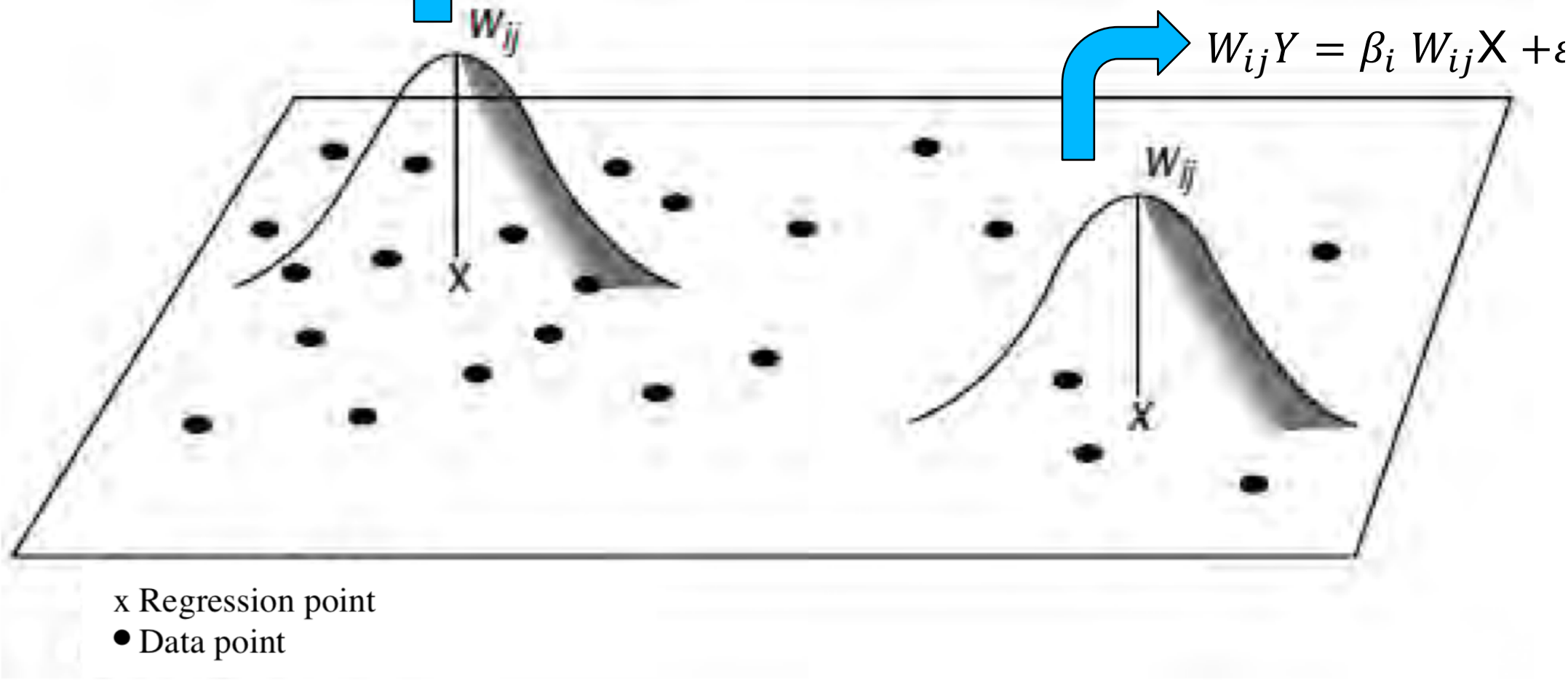
$$Y_i = \beta(u_i, v_i; h) X_i + \epsilon_i \,,$$

Each Local Regression for point *i* is based on a local subsample

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

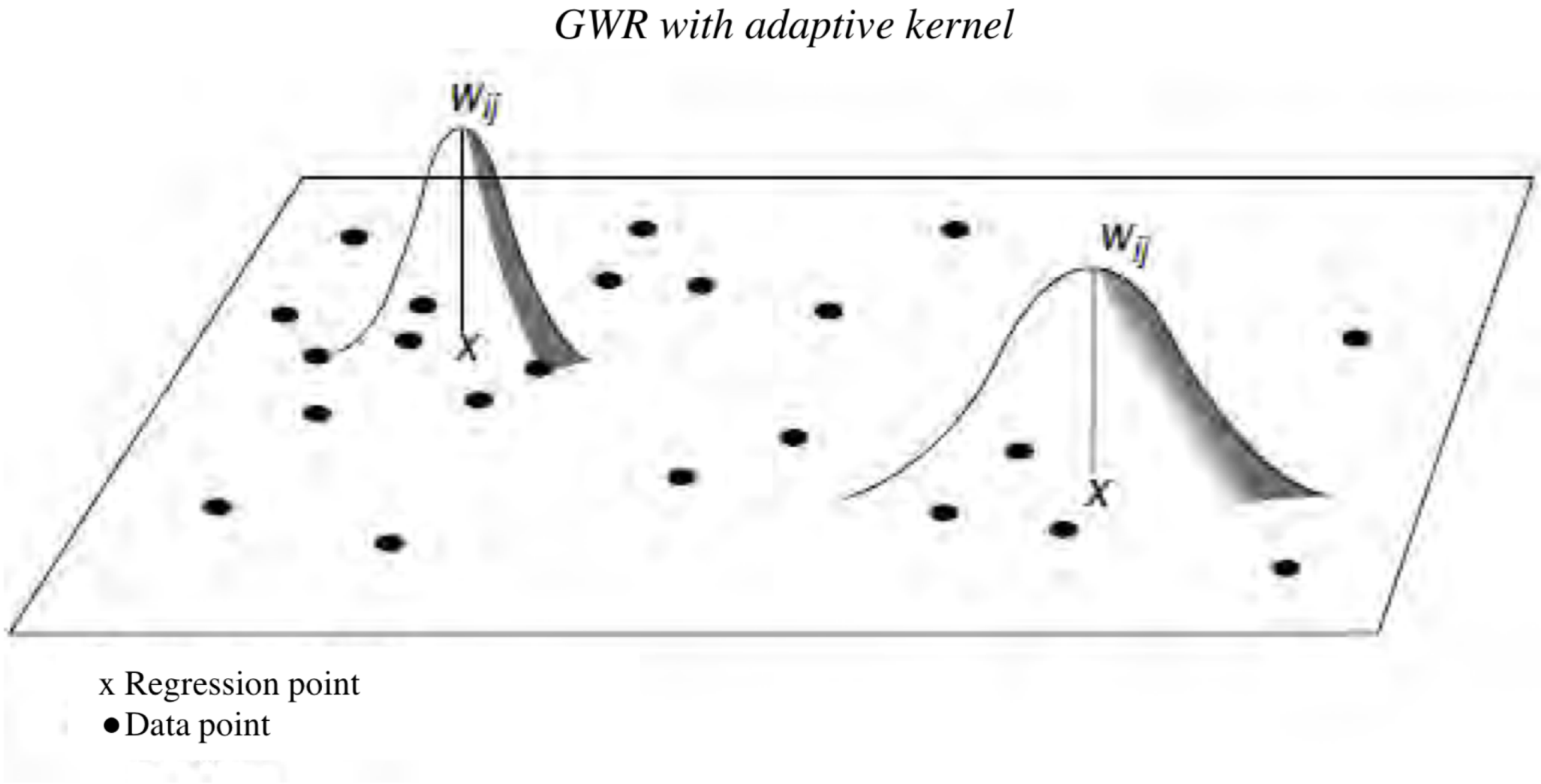
$$W_{ij}Y = \beta_i\, W_{ij}\mathrm{X} + \varepsilon$$

GWR with fixed spatial kernel



Source: Fotheringham–Brunsdon–Charlton 2002, p. 45.

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



Source: Fotheringham–Brunsdon–Charlton 2002, p. 47.

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.

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 2017* **hereafter GM2017**

+

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

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 optimization of bandwith 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

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

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

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

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

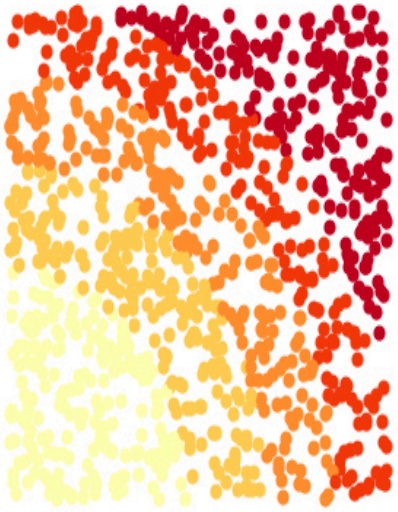
Monte Carlo

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

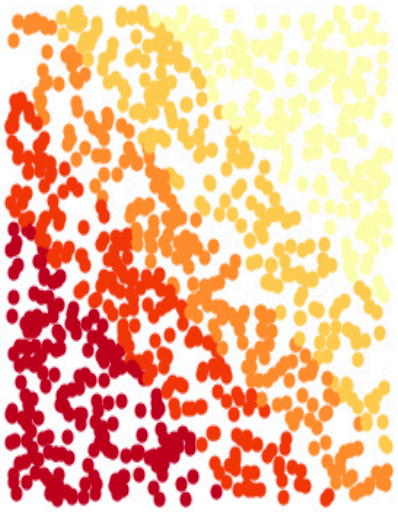


Beta1

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

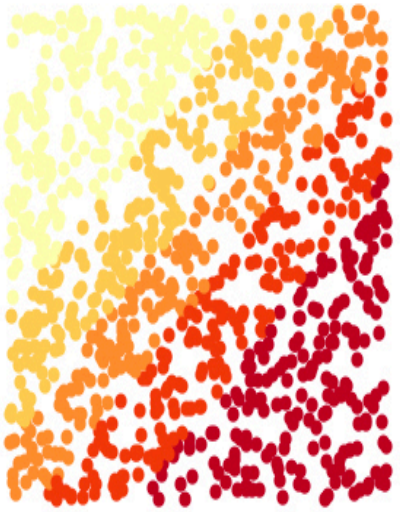
Beta2

- 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

- one submarket is design to be independent

→ 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 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 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 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

RESULTS for land Sales Data

- 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

Model	kernel type	bandwidth	LOO-CV*	In,sample RMSE
OLS (piecewise linear)	none	none	1.4961	1. 4234
SAR (W matrix)	nn	8	1.2710	1.2642
GWR (space kernel)	Gauss_adapt	1600	1.2810	1.2655
GWRSAR (W matrix)	bisq_adapt	14	1.1621	1.0896
(space kernel)	Gauss_adapt	1400		
GWRSARX (W matrix)	bisq_adapt	14	1.0613	1.0310
(surface kernel)	bisq_adapt	3600		
(space kernel	Gauss_adapt	1400		

RESULTS for land Sales Data

Developpable lands N=1531

Model	kernel type	bandwidth	LOO-CV*	In,sample RMSE
OLS (piecewise linear)	none	none	1.2941	1. 1343
SAR (W matrix)	nn	2	1.0412	0.9776
GWR (space kernel)	Gauss_adapt	140	0.9790	0.9049
GWRSAR (W matrix)	bisq_adapt	3	0.9714	0.8929
(space kernel)	Gauss_adapt	135		
GWRSARX (W matrix)	bisq_adapt	3	0.8664	0.7487
(surface kernel)	bisq_adapt	950		
(space kernel	Gauss_adapt	270		

LOO-CV GWRSARX
Ind. Developpable+Agricultural Land= 1.0432

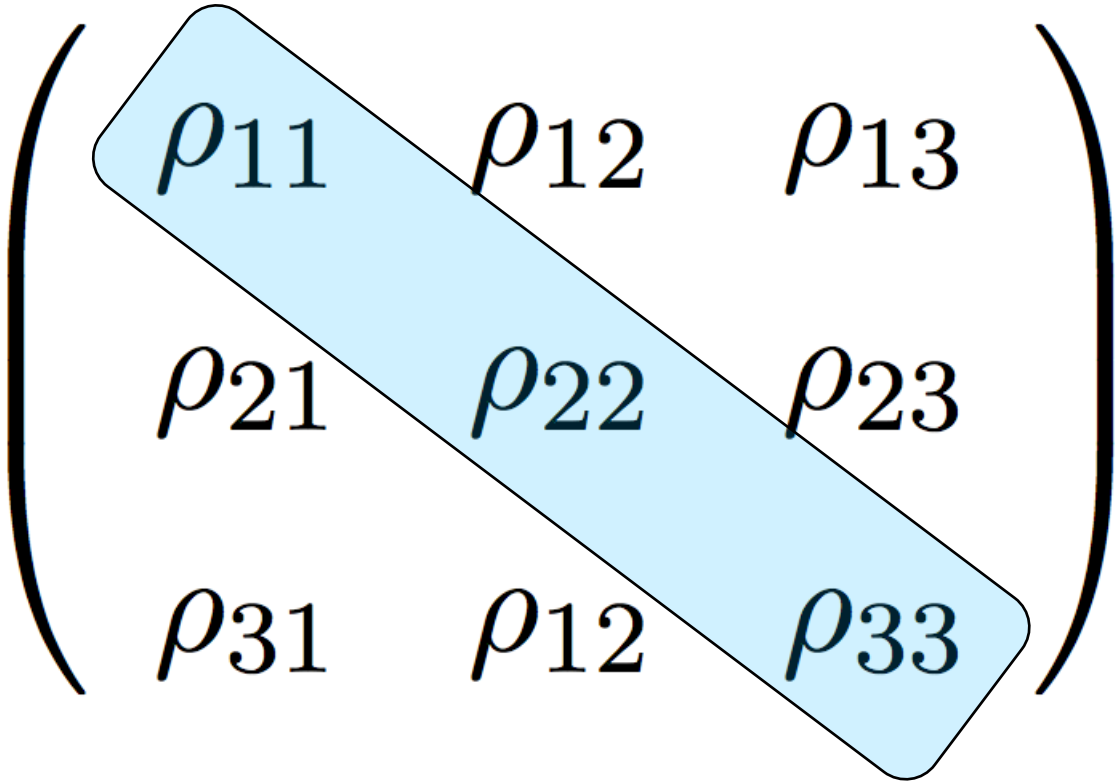
RESULTS for land Sales Data

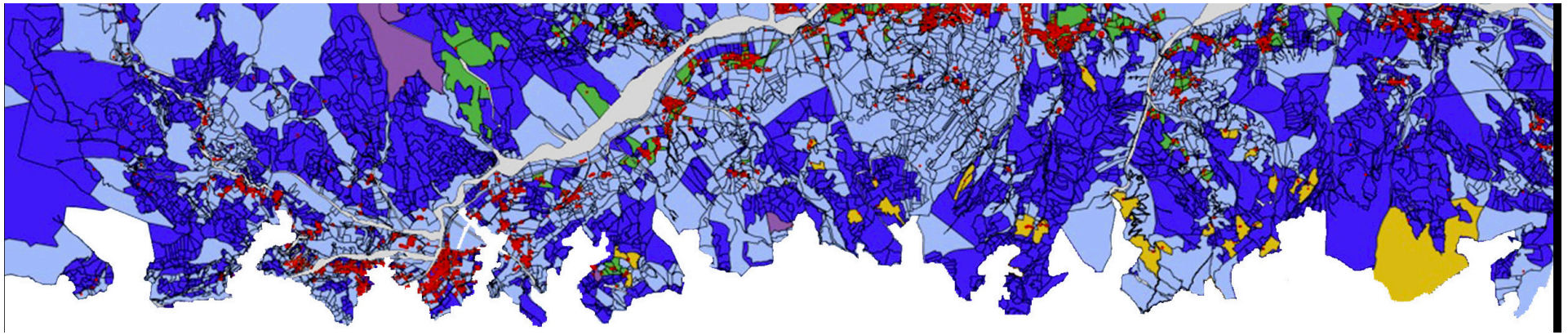
Developpable lands and Agricultural Lands N= 9542

Model	kernel type	bandwidth	LOO-CV*	In,sample RMSE
GWRSARX (W block diag matrix)	bisq_adapt	3 - 14	0.9965	0.7487
(surface kernel)	bisq_adapt	1500 - 3900		
(space kernel	Gauss_adapt	2400 - 2400		
Segment kernel	LR2010	0.03 - 0.4		

LOO-CV GWRSARXC
Developpable+Agricultural Land= 0.9965

Next Step for « shared spatial heterogenity » idea





Introducing mgwrsar R Package