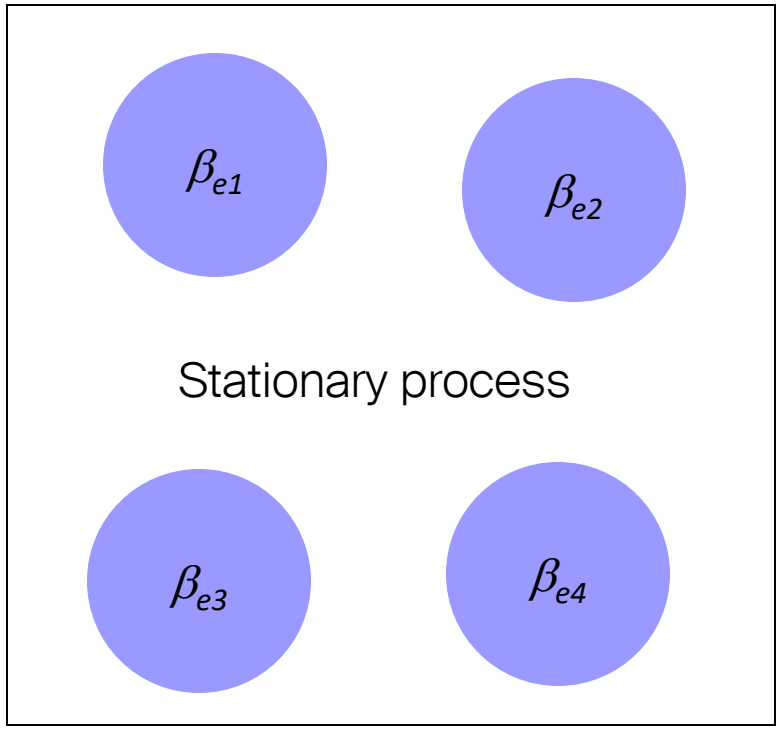
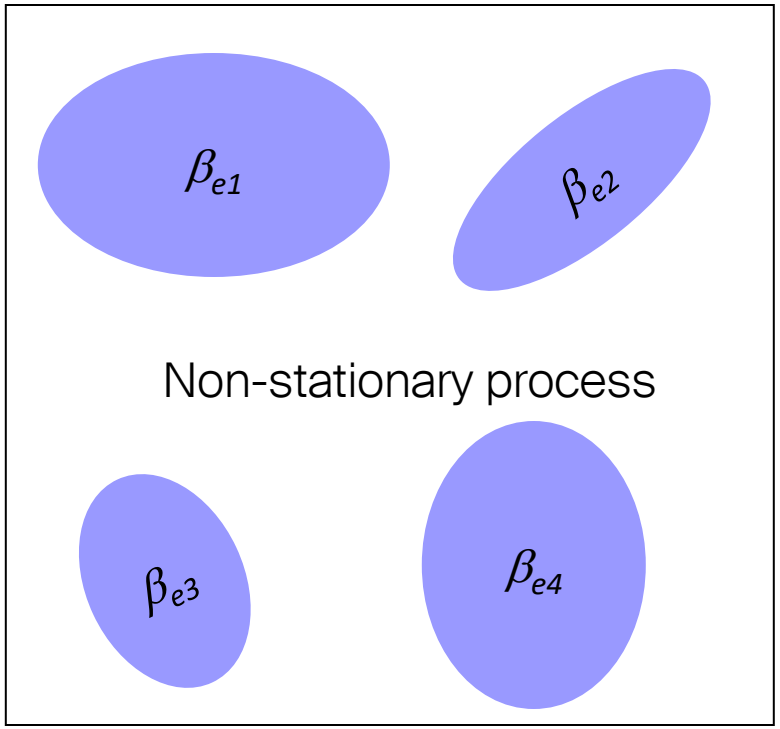
Exploratory data analysis in environmental health

Stéphane Joost & Mayssam Nehme

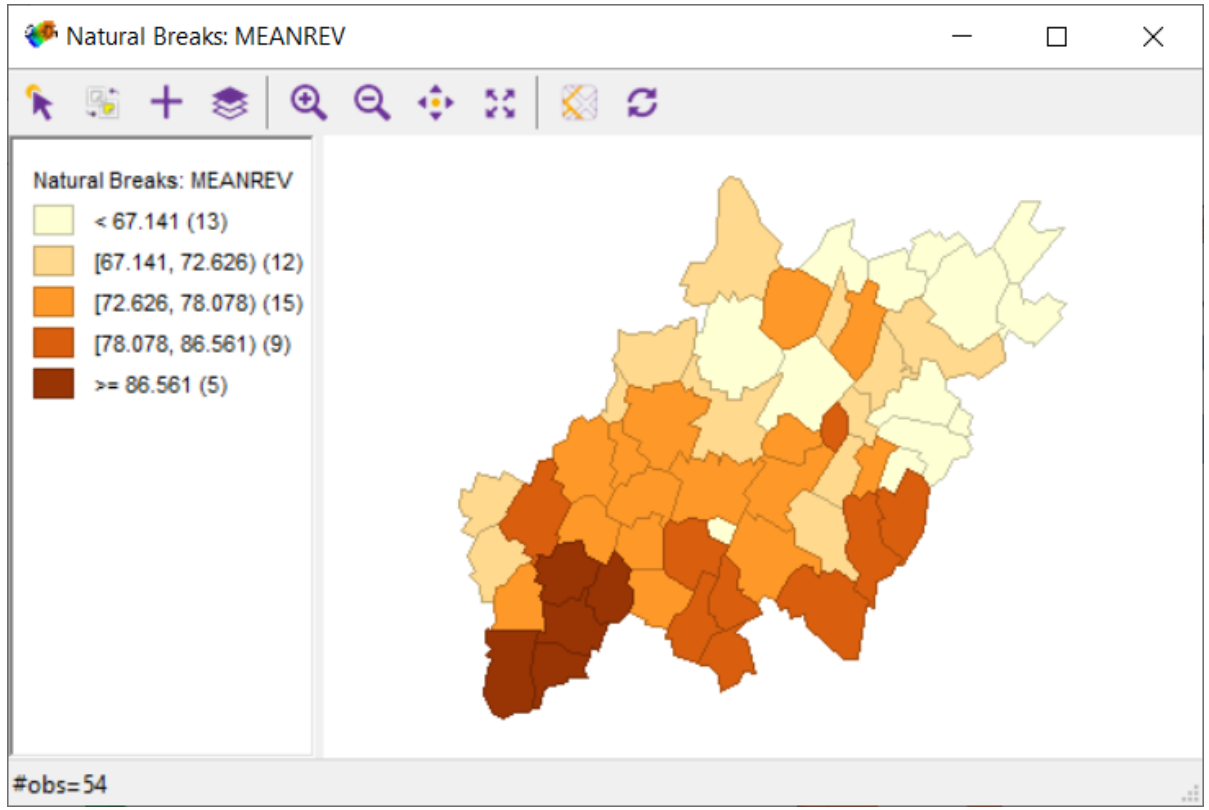
Geographically Weighted Regression

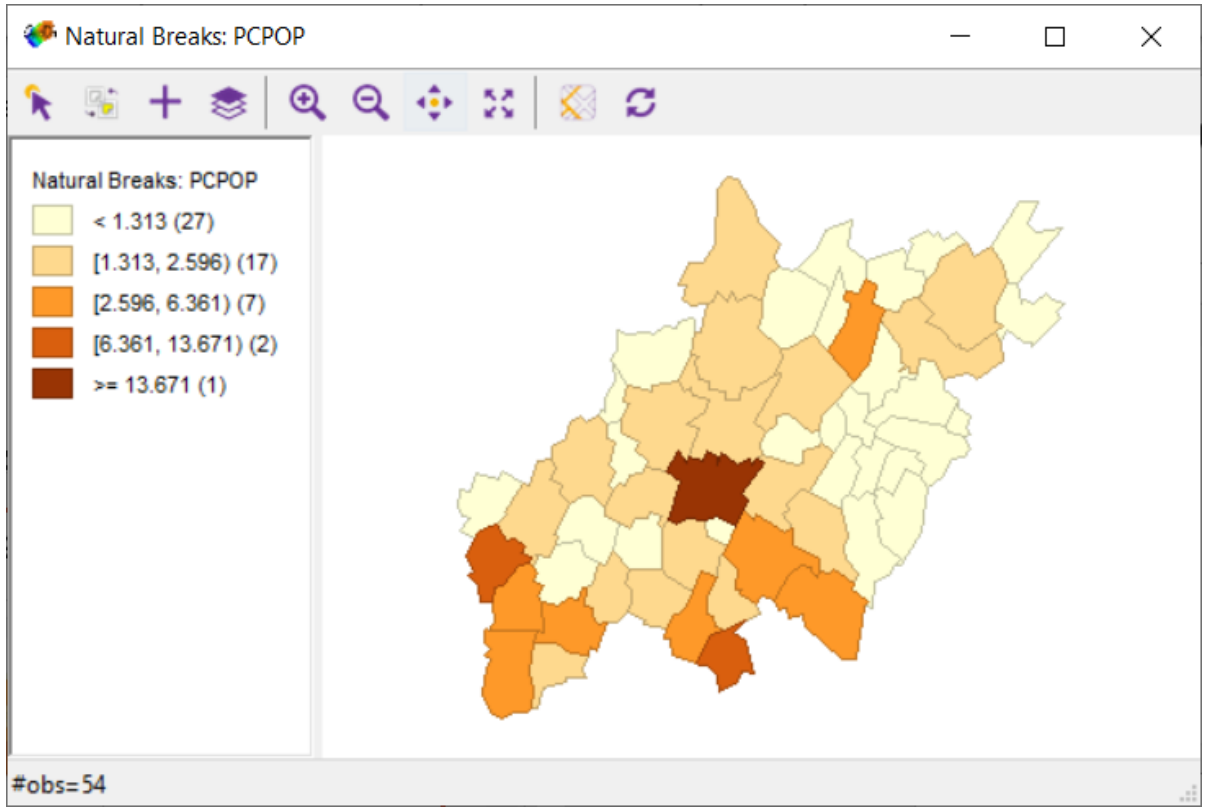
Stationarity versus non-stationarity





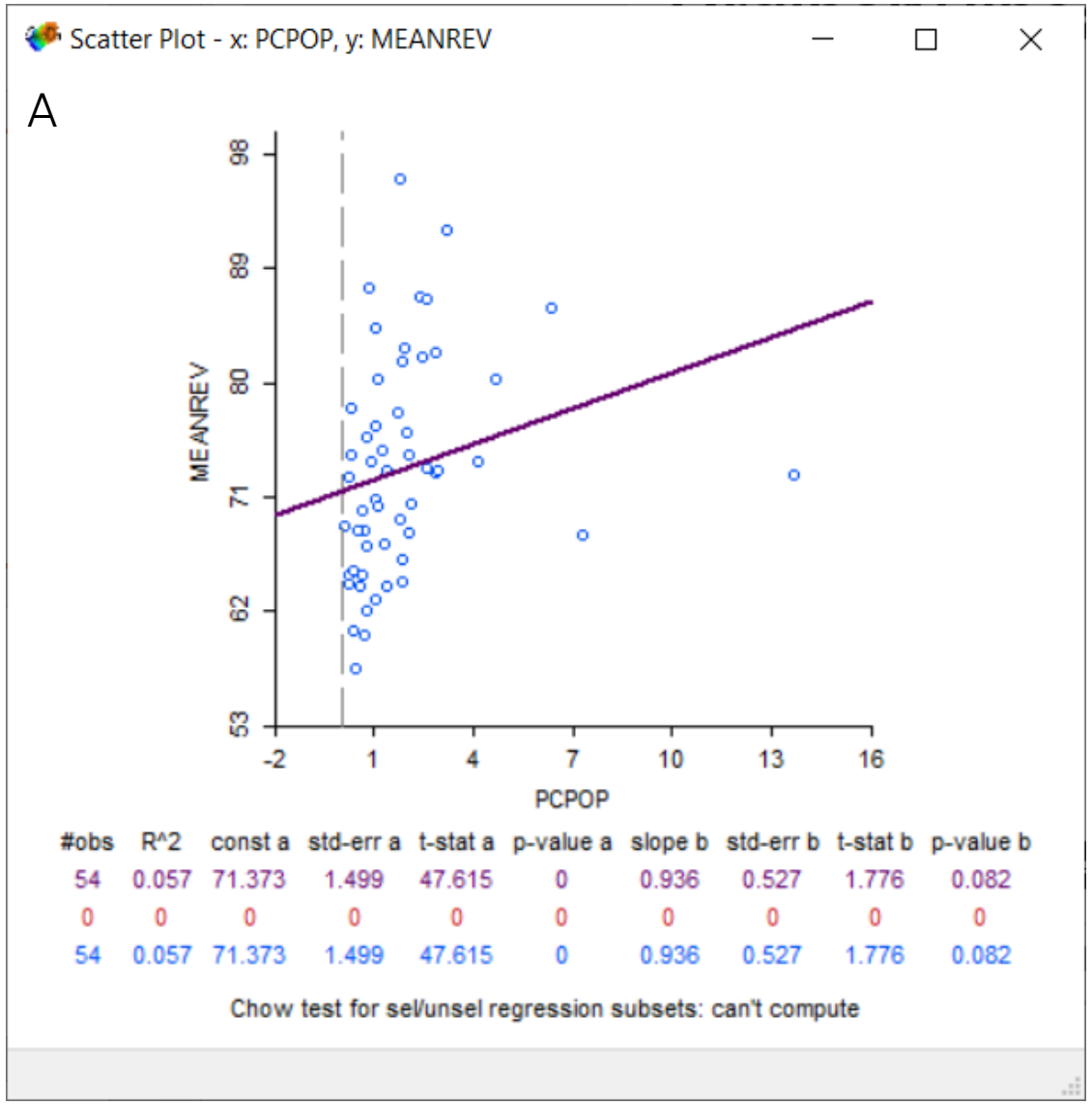
Ordinary linear regression





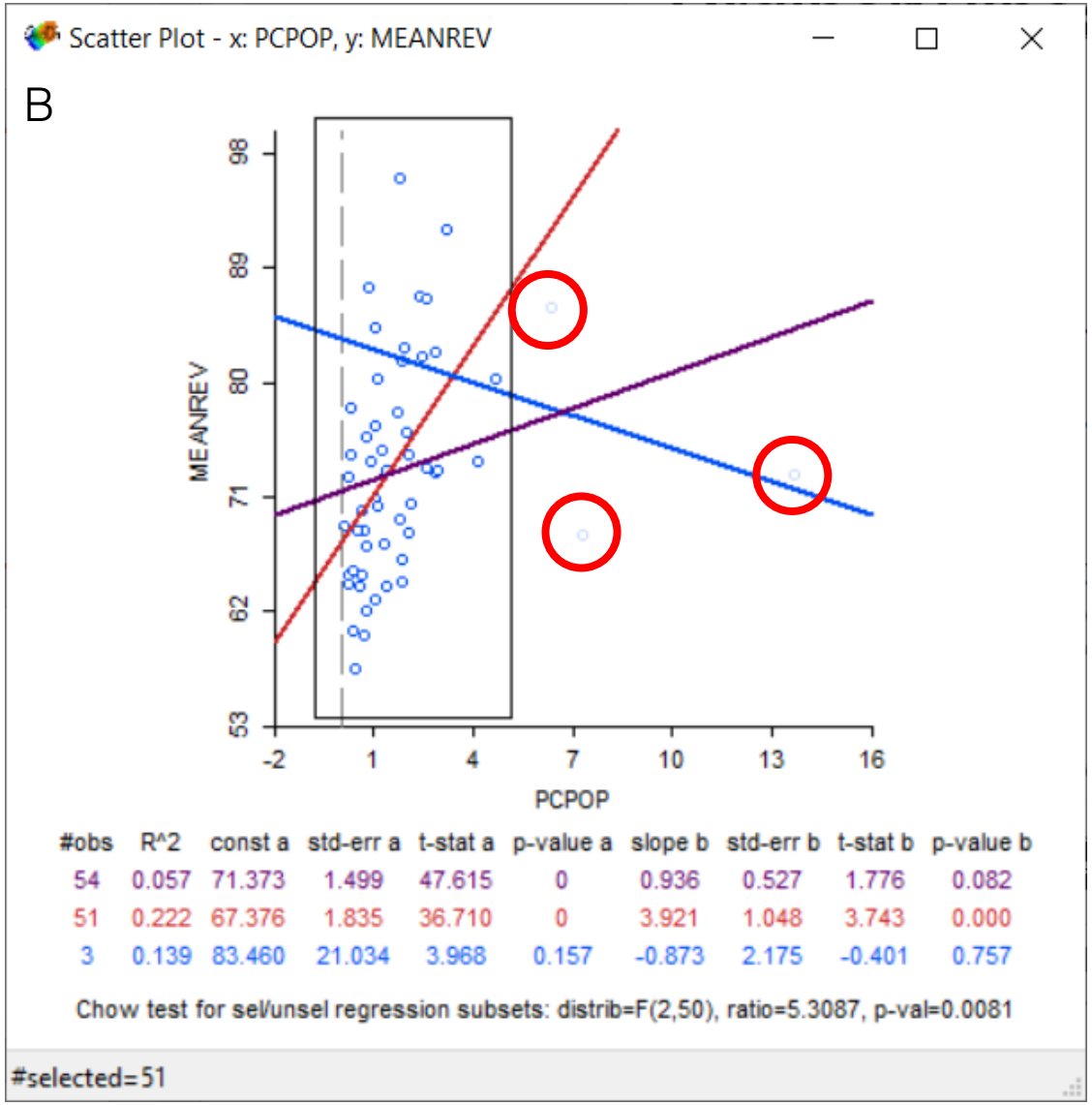
Relationship between the mean income (left) and the percentage of the population living in the whole region per municipality. The hypothesis is that we may find higher revenues where a higher number of inhabitants will translate a higher level of economic activity, with more specialized services

Ordinary linear regression (OLS)



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
54	0.057	71.373	1.499	47.615	0	0.936	0.527	1.776	0.082
0	0	0	0	0	0	0	0	0	0
54	0.057	71.373	1.499	47.615	0	0.936	0.527	1.776	0.082

Chow test for sel/unsel regression subsets: can't compute

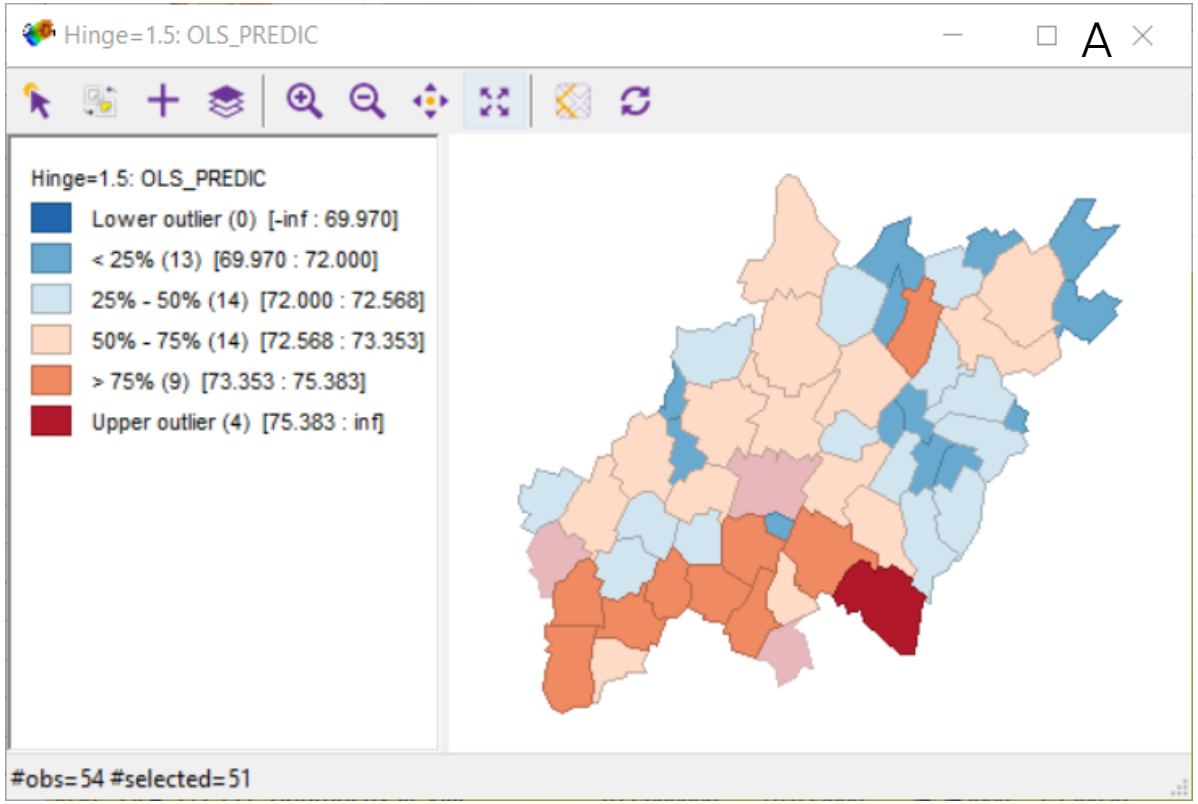


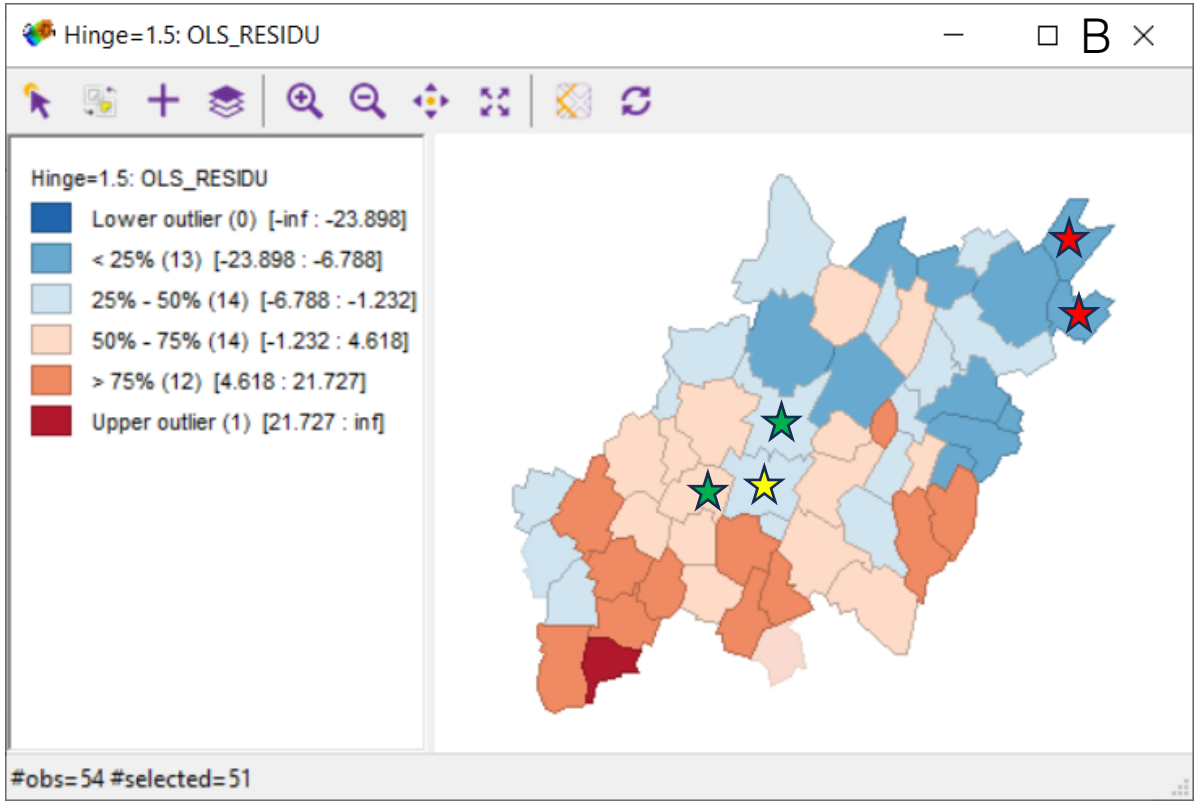
#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
54	0.057	71.373	1.499	47.615	0	0.936	0.527	1.776	0.082
51	0.222	67.376	1.835	36.710	0	3.921	1.048	3.743	0.000
3	0.139	83.460	21.034	3.968	0.157	-0.873	2.175	-0.401	0.757

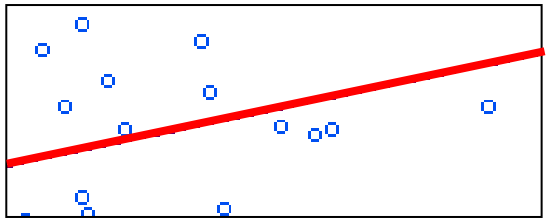
Chow test for sel/unsel regression subsets: distrib=F(2,50), ratio=5.3087, p-val=0.0081

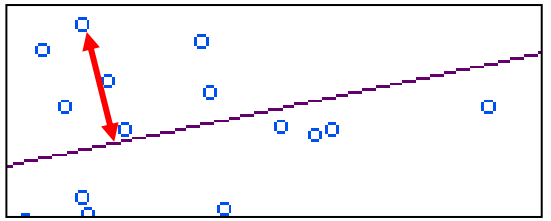
#selected=51

Mapping the results of an OLS



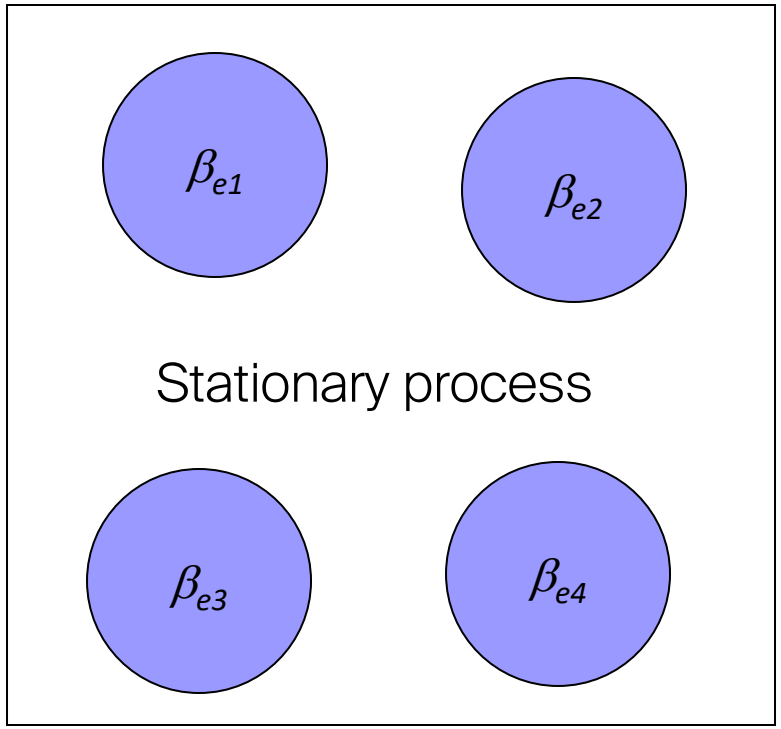




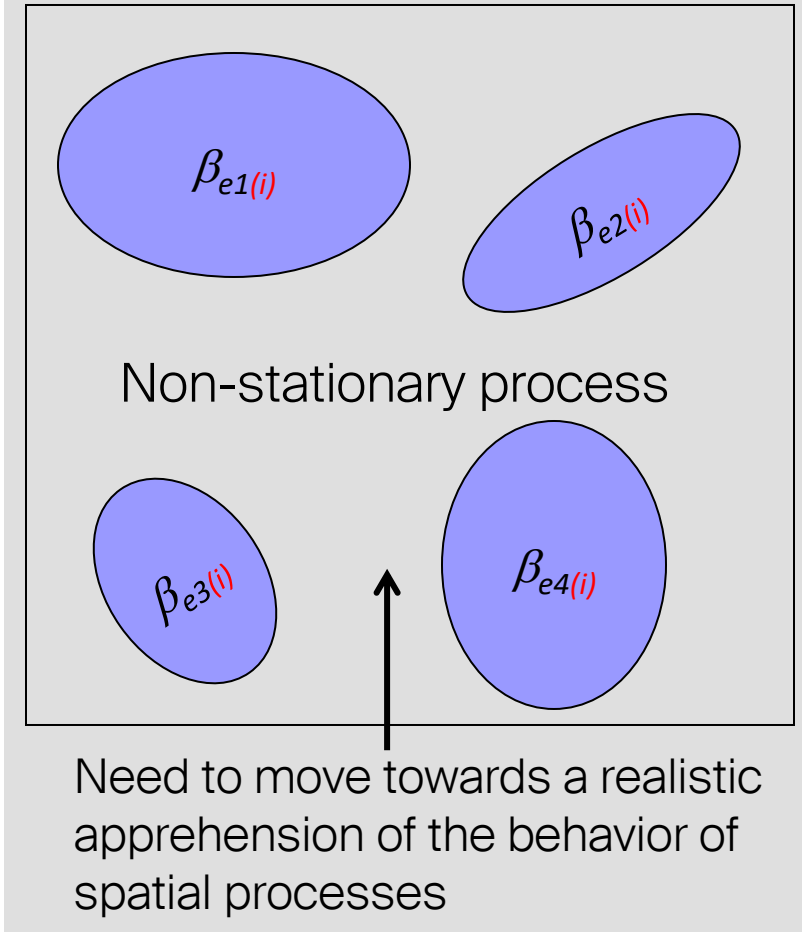


Closer observations should have a greater weight

$$y_i = \beta_0 + \beta_1 x_{1i}$$



$$y_i = \beta_{i0} + \beta_{i1} x_{1i}$$



Brunsdon C, Fotheringham AS and Charlton M (1996) Geographically weighted regression: a method for exploring spatial non-stationarity, Geographical Analysis, 28(4), 281-298

Geographic Weighted Regression

- The basic hypothesis is spatial heterogeneity
- We directly take into account non-stationarity and we admit that relationships in the geographical space will vary:

$$y(i) = \beta_0(i) + \beta_1(i)\, x_1 + \beta_2(i)\, x_2 + \ldots + \ldots + \beta_n(i)\, x_n + \varepsilon(i)$$

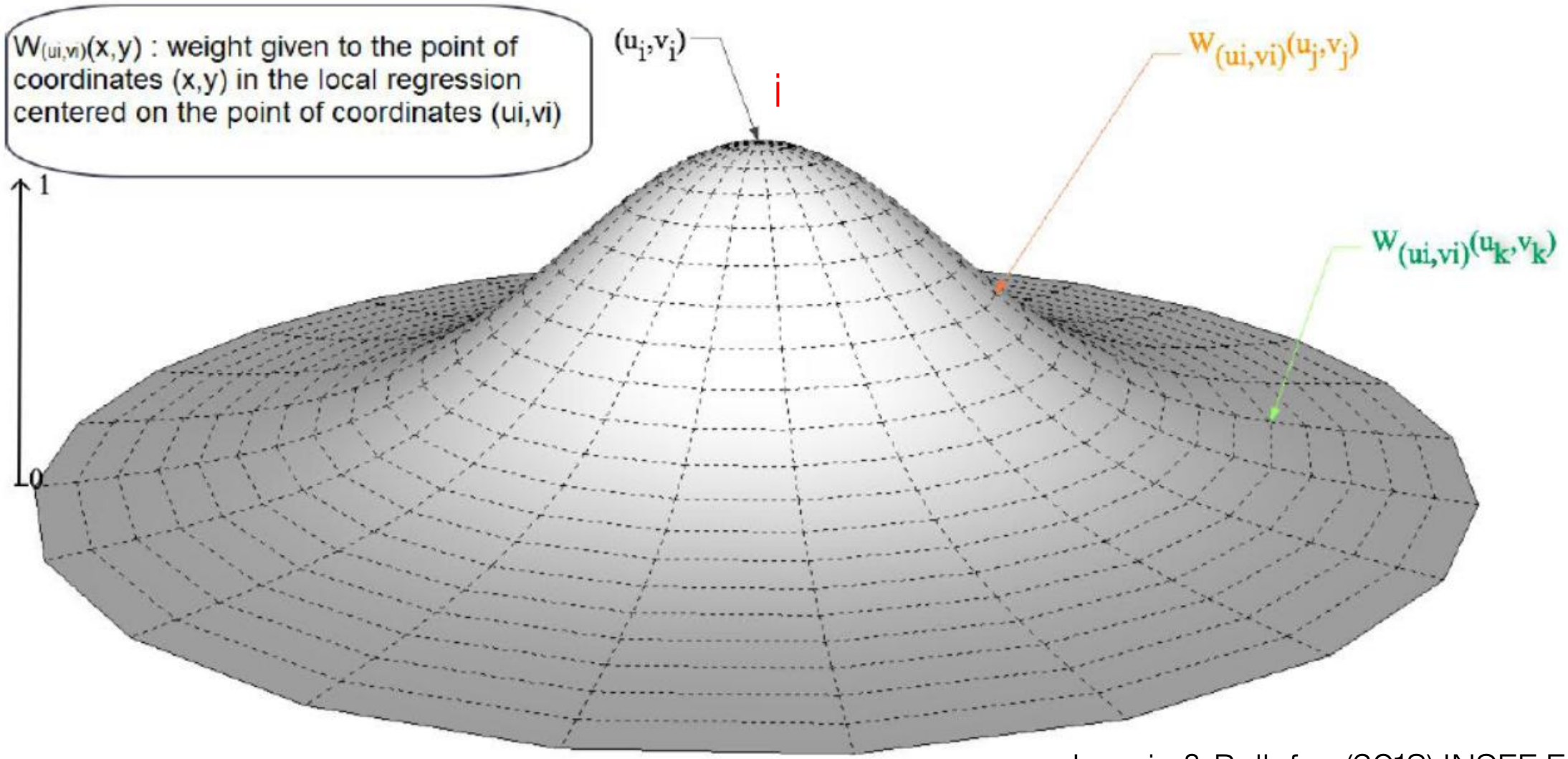
Where (i) refers to a distinct location where parameters are estimated

Influence of explanatory variables

- The closer two observations, the more similar the influence, the closer the coefficients of explanatory parameters
- Regression with close observations of location i only ?
  - No, because the more points in the sample, the lowest the variance
  - It does not remove the bias, even with few neighboring points (measure of 1 global effect instead of several local effects)

- Solution ?

Reduce importance of most remote observations



Loonis & Bellefon (2018) INSEE Eurostat

W Matrix

Parameters are calculated with the following estimator:

$$\beta' = (X^T W(i)\, X)^{-1} X^T W(i)\, Y$$

One "matrix" per i location, i.e. per regression point

$$W(i) = \begin{pmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 0 & 0 & 0 & w_{in} \end{pmatrix}$$

W Matrix

A

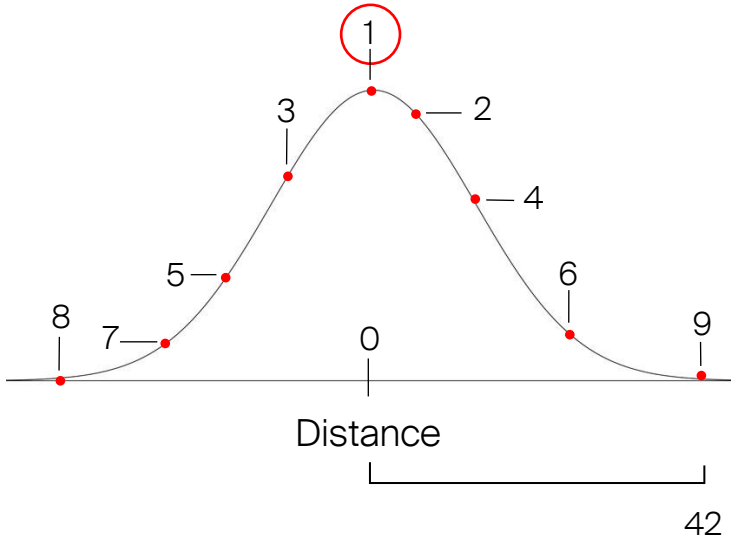
$$W(x_i) = \begin{bmatrix} 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9950 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8825 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7261 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2780 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0889 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0340 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0060 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0002 \end{bmatrix}$$

Lloyd, 2009

B

No.	x coordinate	y coordinate	Variable 1 (y)	Variable 2 (z)	Distance (d_{ij})	Geog. wt. (w_{ij})
i → 1	25.00	45.00	12	6	0.00	1.0000
2	25.51	44.14	34	52	1.00	0.9950
3	21.87	48.90	32	41	5.00	0.8825
4	27.60	52.57	12	25	8.00	0.7261
5	16.69	31.33	11	22	16.00	0.2780
6	42.52	35.35	14	9	20.00	0.0889
7	9.20	65.65	56	43	26.00	0.0340
8	29.23	76.72	75	67	32.00	0.0060
9	61.37	66.01	43	32	42.00	0.0002

C

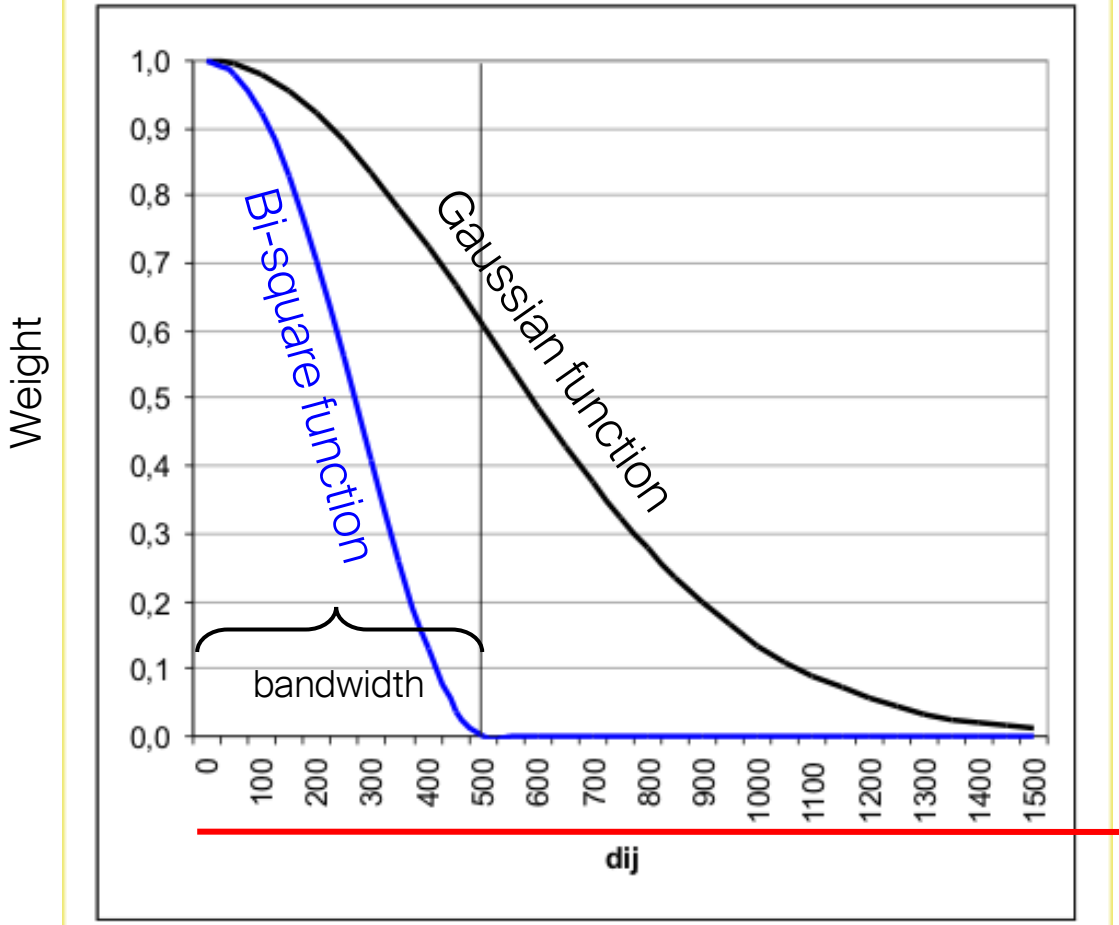


D

No	Var 2	Weight	Var 2 in regression
1	6	1	6
2	52	0.99	51.48
3	41	0.88	36.08
4	25	0.72	18
5	22	0.27	5.94
6	9	0.08	0.72
7	43	0.03	1.29
8	67	0.006	0.402
9	32	0.0002	0.0064

Weighted value of variable 2 in the regression

Kernel functions



Key parameters

1. Shape of the kernel
   a) Gaussian
   b) Bi-square
   c) Etc.
2. Fixed versus adaptive kernel
   a) Distance
   b) Number of neighbors
3. Size of the bandwidth

As many equations as there are geo-units

- A GWR will generate 1 equation for each spatial unit *i*
- E.g. 50 spatial units
  = 50 equations = 50 sets of estimated parameters
- Sets (range) of local β_0 , local β_j, local significance tests (e.g. Student's T), and local r_i^2 values

OLS versus GWR

	OLR	GWR
Predictor variables	Global parameter estimate	GWR parameter estimates
Total population, β_1	0.24 x10^{-4}	0.14 to 0.28 x10^{-4}
% rural, β_2	-0.044	-0.06 to -0.03
% elderly, β_3	-0.06 (not signif.)	-0.26 to -0.06
% foreign born, β_4	1.26	0.51 to 2.42
% poverty, β_5	-0.15	-0.20 to –0.00
% black, β_6	0.022 (not signif.)	-0.04 to 0.08
Intercept, β_0	14.78	12.62 to 16.49
Diagnostics		
Residual SS	1816 →	1506
Adjusted R^2	0.63 →	0.68
AICc	855.4	839.2

Parameter estimate

Range of parameter estimates

Akaike Information Criterion is a measure of the relative goodness of fit →

$$AIC = 2k - 2\ln(L)$$

Residuals = (sum of the local residuals)^2

Coefficient of determination (% of variance explained)

R2 = Coefficient of determination

The global estimation of the quality of a model is based on the equation of variance analysis:

$$SCT = SCE + SCR$$

where:

- *SCT* = total sum of squares = total variance
- *SCE* = sum of explained squares = variance explained by the model
- *SCR* = sum of residual squares = variance not explained by the model

Source de variation	Somme des carrés
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$

And the coefficient of determination is:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Global R2 for GWR

- The overall R2 value for GWR is calculated the exact same way as for OLS
- It is the proportion of the variability explained by the model
- For both it is:
  $1 - (\Sigma(\text{predicted-observed})^2/\Sigma(\text{observed-mean of all observed})^2)$
- In other words, it is 1 minus the variance of the residuals divided by the variance of the input data
- In both OLS and GWR, the residuals are the predicted minus the observed

Diagnostic information (global statistics)

OLS

```
  < Diagnostic information >
Residual sum of squares:                   3636.112485
Number of parameters:                            2
 (Note: this num does not include an error variance term
ML based global sigma estimate:                8.205816
Unbiased global sigma estimate:                8.362131
Log-likelihood:                              380.568424
Classic AIC:                                 386.568424
AICc:                                        387.048424
BIC/MDL:                                     392.535376
CV:                                           82.105037
R square:                                      0.057164
Adjusted R square:                             0.020190
```

GWR

```
  Diagnostic information
Residual sum of squares:                   1718.519412
Effective number of parameters (model: trace(S)):
Effective number of parameters (variance: trace(S'S)):
Degree of freedom (model: n - trace(S)):
Degree of freedom (residual: n - 2trace(S) + trace(S'S)):
ML based sigma estimate:                       5.641315
Unbiased sigma estimate:                       6.099078
Log-likelihood:                              340.098016
Classic AIC:                                 353.541354
AICc:                                        355.784413
BIC/MDL:                                     366.910647
CV:                                           53.935680
R square:                                      0.554391
Adjusted R square:                             0.477475
```

Arizona State University

MGWR

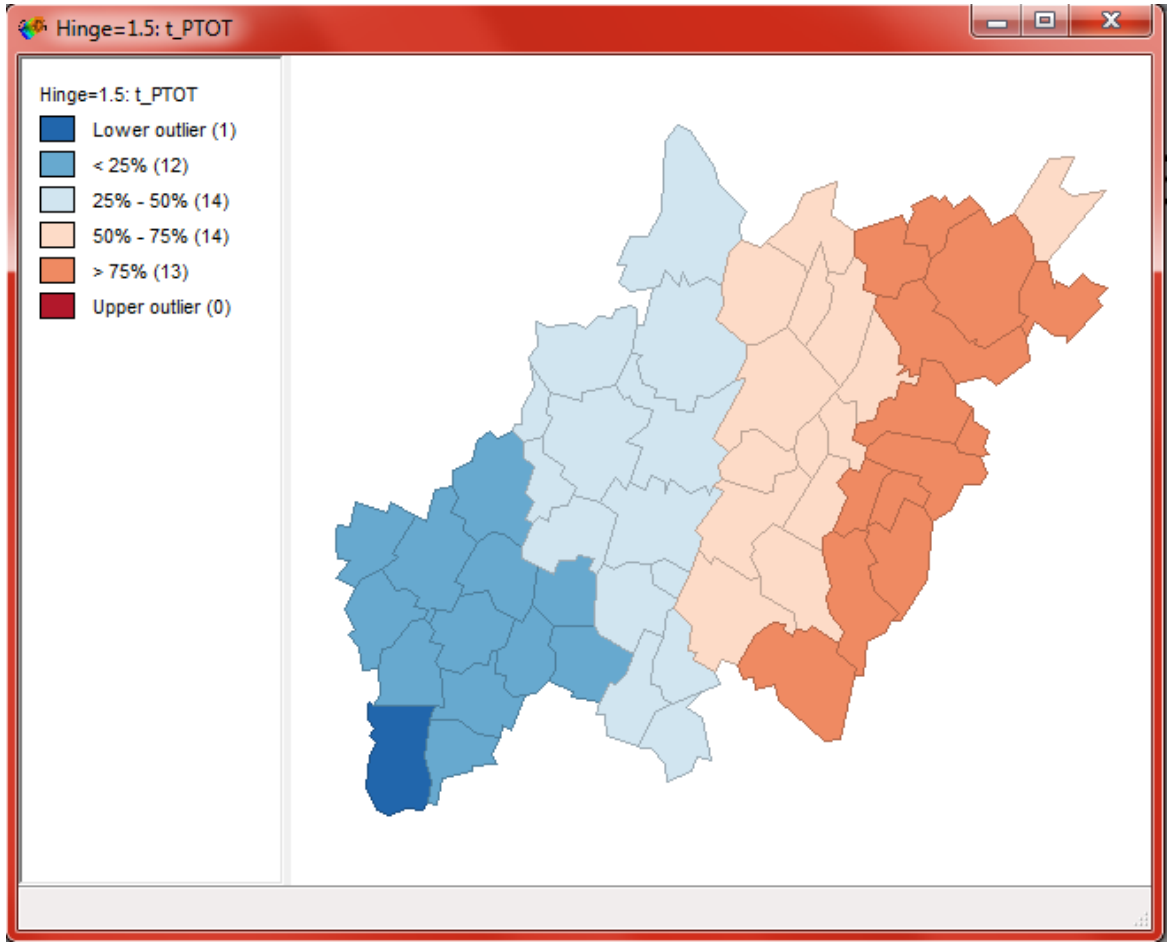
School of Geographical Sciences & Urban Planning

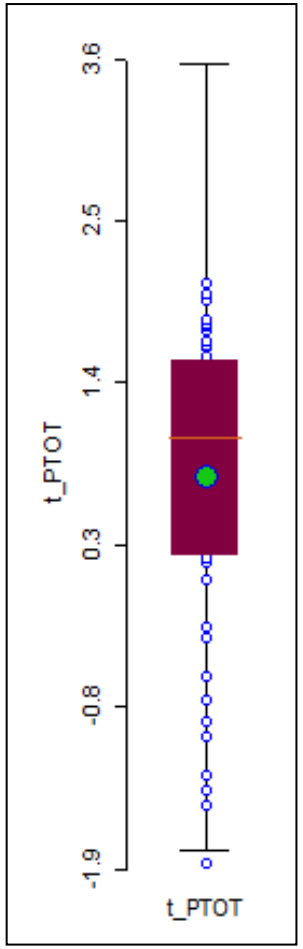
https://sgsup.asu.edu/sparc/mgwr

Local statistics

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Area_key	x_coord	y_coord	est_Intercept	se_Intercept	t_Intercept	est_PTOT	se_PTOT	t_PTOT	y	yhat	residual	std_residual	localR2	influence	CooksD
1	533869.5513	163747.8536	77.436322	1.597498	48.4735	-0.000355	0.001385	-0.256509	76.6417	77.319102	-0.677402	-0.115238	0.376025	0.071096	0.000178
2	532922.1075	162150.8069	79.438641	1.731887	45.868268	-0.001123	0.001497	-0.749842	87.3991	79.156884	8.242216	1.420243	0.348336	0.094613	0.03684
3	534473.1107	161640.4585	78.826897	1.642424	47.994242	-0.000458	0.001384	-0.330697	86.6725	78.486859	8.185641	1.382928	0.348497	0.058159	0.02064
4	532074.6162	164422.9523	77.817779	1.728277	45.026206	-0.000904	0.001526	-0.592259	82.6386	77.277226	5.361374	0.911412	0.371737	0.069761	0.010887
5	530163.1204	164671.952	78.493839	1.951297	40.226491	-0.001686	0.001875	-0.899351	70.7993	77.942431	-7.143131	-1.249798	0.359115	0.12185	0.03788
6	532520.2008	158357.6681	83.699563	2.101443	39.829569	-0.002648	0.001805	-1.466791	95.9672	82.251264	13.715936	2.393639	0.243209	0.117317	0.133091
7	530183.4226	162519.8456	80.752933	2.043233	39.522137	-0.00256	0.002021	-1.266995	67.9182	74.908017	-6.989817	-1.595266	0.343115	0.483896	0.417022
8	531126.3823	161113.6843	81.754357	2.003263	40.810599	-0.002584	0.001895	-1.363313	73.8325	78.428558	-4.596058	-0.803198	0.321258	0.119767	0.015341
9	533048.0332	160304.9959	81.145333	1.838353	44.140227	-0.001565	0.001568	-0.998102	86.5608	79.872824	6.687976	1.137739	0.309158	0.071086	0.017313
10	530803.4731	158561.3138	85.112597	2.331331	36.508158	-0.00405	0.002176	-1.861054	92.0691	81.066781	11.002319	1.920717	0.246674	0.11791	0.086187
11	537589.7716	162960.3057	75.69708	1.43137	52.884372	0.000717	0.00137	0.523644	81.9781	76.249394	5.728706	0.963378	0.373385	0.049415	0.008432
12	543972.0901	171280.4365	67.641072	1.369716	49.383278	0.00254	0.001685	1.508023	72.9163	69.87146	3.04484	0.532437	0.306679	0.120847	0.006811
13	535846.9672	163314.8022	76.583384	1.494086	51.257684	0.000235	0.001359	0.172626	73.7992	76.650945	-2.851745	-0.483837	0.378007	0.066114	0.002897
14	540623.0571	162883.9766	73.766033	1.375682	53.621421	0.001541	0.00138	1.117027	73.0555	75.179407	-2.123907	-0.358169	0.351922	0.054708	0.001298
15	539086.911	161461.6647	75.615569	1.466645	51.556825	0.001219	0.001383	0.881637	81.7319	76.302085	5.429815	0.918435	0.356803	0.060395	0.009476
16	539199.6	159116.9638	76.86166	1.617971	47.504963	0.001475	0.001491	0.988902	85.9219	79.799578	6.122322	1.154408	0.325434	0.24389	0.075129
17	543324.6357	165975.9712	70.569128	1.337583	52.758701	0.002002	0.001407	1.423063	67.1409	71.05163	-3.91073	-0.659349	0.322568	0.054292	0.004362
18	538580.0272	165247.3977	73.558466	1.309579	56.169536	0.000941	0.001399	0.672429	72.6973	77.58573	-4.88843	-1.892294	0.367482	0.820596	2.862536
19	534894.61	166626.9474	74.635945	1.489027	50.123986	0.000308	0.001394	0.221168	74.361	74.661544	-0.300544	-0.051528	0.391613	0.08547	0.000043
20	538825.4819	174476.8098	68.69367	1.562051	43.976579	0.001707	0.001725	0.989676	70.4458	69.825425	0.620375	0.109271	0.226946	0.133496	0.000322
21	536327.3115	161410.6288	77.661558	1.559516	49.798498	0.000293	0.001362	0.214942	73.1911	77.895697	-4.704597	-0.793813	0.355024	0.055764	0.006504
22	541690.8328	169143.1372	69.736835	1.271786	54.833768	0.001565	0.001439	1.087116	64.0425	70.397192	-6.354692	-1.068014	0.300782	0.048285	0.010114
23	542682.0919	160671.1831	73.687029	1.51819	48.536112	0.002454	0.001461	1.679834	80.2088	77.252245	2.956555	0.523328	0.325611	0.141983	0.007921
24	536601.8263	167853.8124	72.928502	1.377467	52.943931	0.000678	0.001425	0.475449	76.053	73.342466	2.710534	0.457222	0.369342	0.055227	0.002136
25	534578.6108	168903.9423	73.175594	1.555864	47.032115	0.000496	0.001436	0.345391	68.6083	73.192461	-4.584161	-0.797468	0.376037	0.11169	0.013975
26	538707.0612	163728.0121	74.462472	1.360562	54.729201	0.000993	0.001382	0.71848	64.7667	74.523034	-9.756334	-1.662994	0.369441	0.074741	0.039044
27	537872.6526	160526.8514	77.089128	1.554612	49.587383	0.000868	0.001393	0.62297	82.3043	77.859878	4.444422	0.752387	0.346427	0.061963	0.006535
28	543024.592	167636.4446	69.853103	1.300503	53.71236	0.001855	0.001425	1.301896	78.0778	70.021889	8.055911	1.363007	0.313401	0.060916	0.021062
29	533952.1121	166369.4479	75.286409	1.557345	48.342794	0.000087	0.001392	0.062329	73.0652	75.323457	-2.258257	-0.383281	0.391848	0.06678	0.001837
30	541346.5515	172159.1396	68.530738	1.36992	50.025359	0.001838	0.001558	1.179695	74.7324	69.245655	5.486745	0.929587	0.274853	0.063471	0.010236
31	536372.3464	170600.2131	71.373725	1.471191	48.514232	0.000817	0.001465	0.557713	70.208	71.653131	-1.445131	-0.247272	0.319819	0.0818	0.000952

Mapping local significance





Net primary production (NPP) in China

NPP is the rate at which energy is stored as biomass by plants

Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2005) **14**, 379–393

RESEARCH PAPER

Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems

Quan Wang[1]*, Jian Ni[2,3] and John Tenhunen[1]

[1]*Department of Plant Ecology, University of Bayreuth, D-95440 Bayreuth, Germany;* [2]*Laboratory of Quantitative Vegetation Ecology, Institute of Botany, Chinese Academy of Sciences, 100093 Beijing, China; and* [3]*Max Planck Institute for Biogeochemistry, PO Box 100164, D-07701 Jena, Germany*

ABSTRACT

Aim The objective of this paper is to obtain a net primary production (NPP) regression model based on the geographically weighted regression (GWR) method, which includes spatial non-stationarity in the parameters estimated for forest ecosystems in China.

Location We used data across China.

Methods We examine the relationships between NPP of Chinese forest ecosystems and environmental variables, specifically altitude, temperature, precipitation and time-integrated normalized difference vegetation index (TINDVI) based on the ordinary least squares (OLS) regression, the *spatial lag* model and GWR methods.

NPP data from more than 1200 sites in 29 Provinces of China.

Net primary production (NPP) in China

Table 1 The independent variables of each model candidates designed for both GWR and OLS methods

Models	Independent variables
Model 1	Altitude, Temperature, Precipitation, and TINDVI
Model 2	Altitude, Temperature, and Precipitation
Model 3	Altitude, Temperature, and TINDVI
Model 4	Altitude, Precipitation, and TINDVI
Model 5	Altitude, Temperature, and TINDVI
Model 6	Temperature, and Precipitation
Model 7	Precipitation, and TINDVI
Model 8	Temperature, and TINDVI
Model 9	Altitude
Model 10	Temperature
Model 11	Precipitation
Model 12	TINDVI

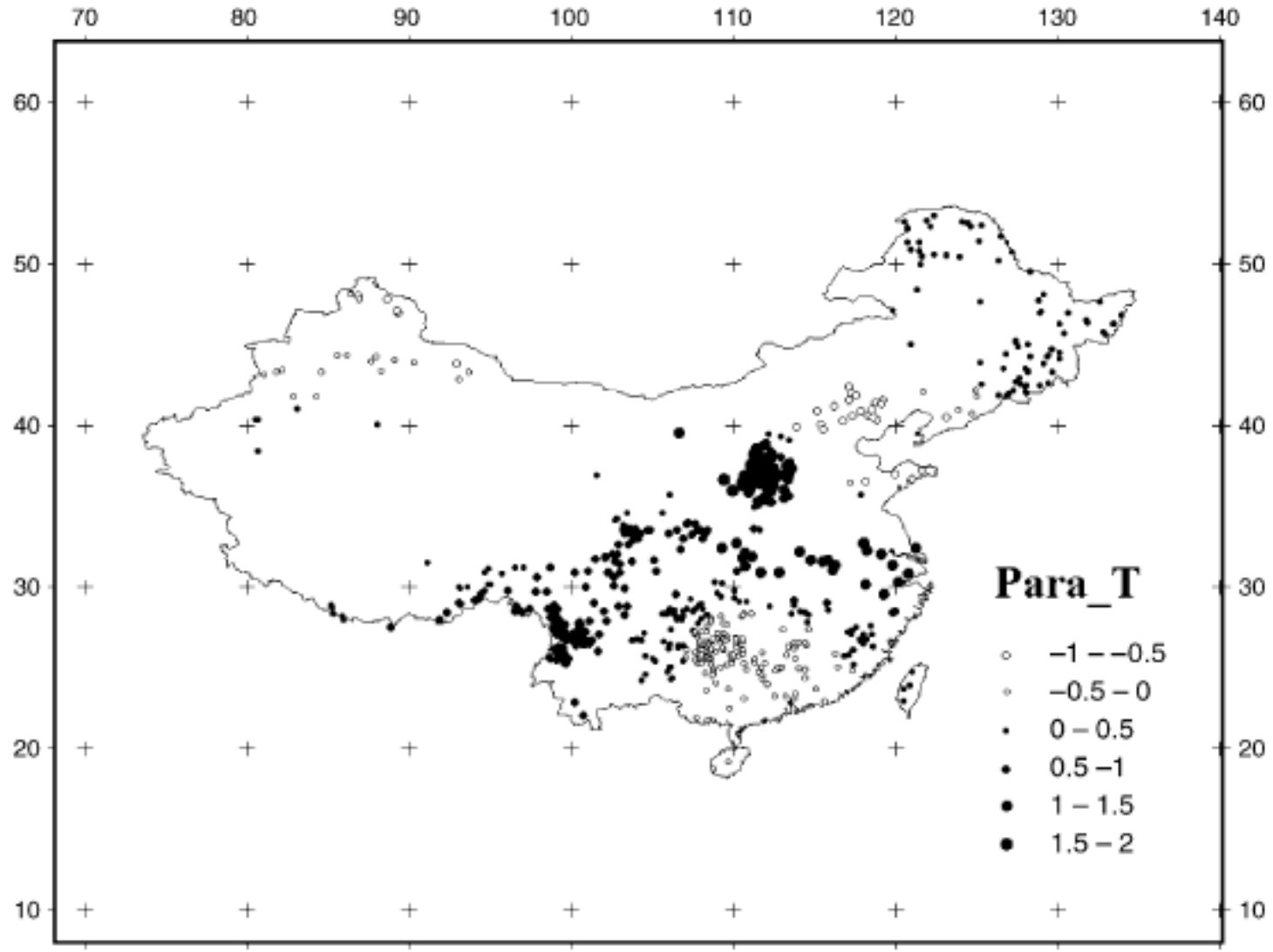
Net primary production (NPP) in China

Table 3 Descriptive statistics of the parameter estimates for the Model 1 both by OLS and GWR methods

Methods	Statistics	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	Estimate	6.1387	−0.0009	0.0998	0.0075	−0.0087
	Standard error	0.7959	0.0001	0.0343	0.0004	0.0051
	Lower limit of 95% CI	4.5767	−0.0012	0.0325	0.0067	−0.0188
	Upper limit of 95% CI	7.7007	−0.0007	0.1671	0.0083	0.0014
	b − 1 SD	5.3428	−0.0011	0.0655	0.0071	−0.0139
	b + 1 SD	6.9346	−0.0008	0.1341	0.0079	−0.0036
GWR	Mean	−0.5225	0.0021	0.5160	0.0020	0.0143
	Minimum	−19.2328	−0.0052	−0.8503	−0.0152	−0.0640
	25% quartile	−12.5593	−0.0015	0.0606	−0.0004	0.0035
	Median	3.4685	0.0004	0.4752	0.0023	0.0235
	75% quartile	8.0914	0.0075	0.9816	0.0063	0.0337
	Maximum	20.4342	0.0142	1.8626	0.0114	0.0692

Net primary production (NPP) in China

The beta parameter for temperature, showing the effect of temperature to explain the amount of net annual primary production



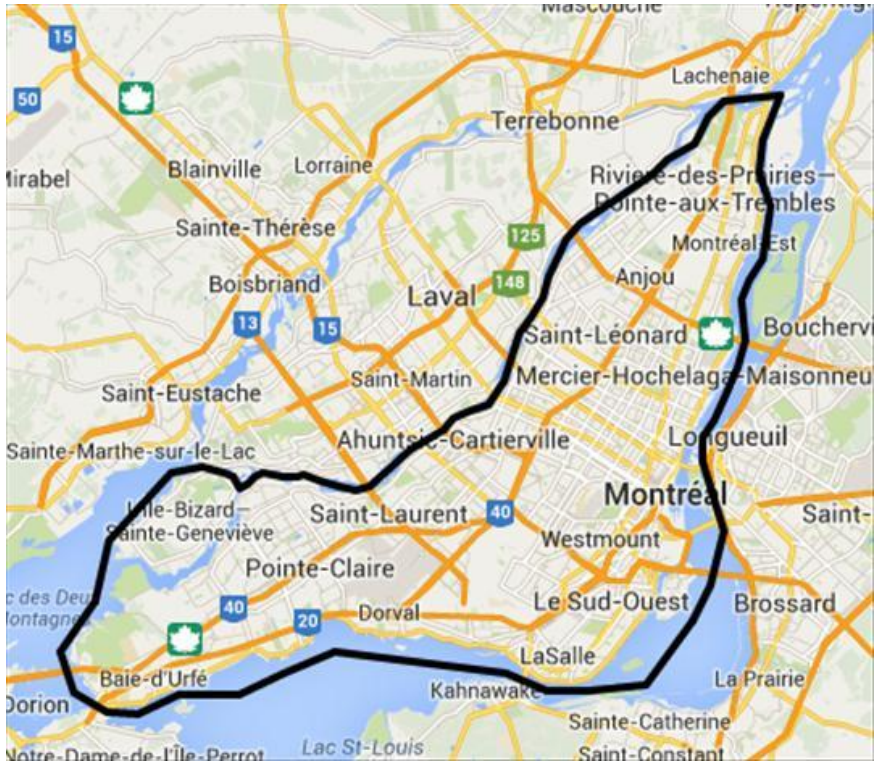
Poverty in Montreal

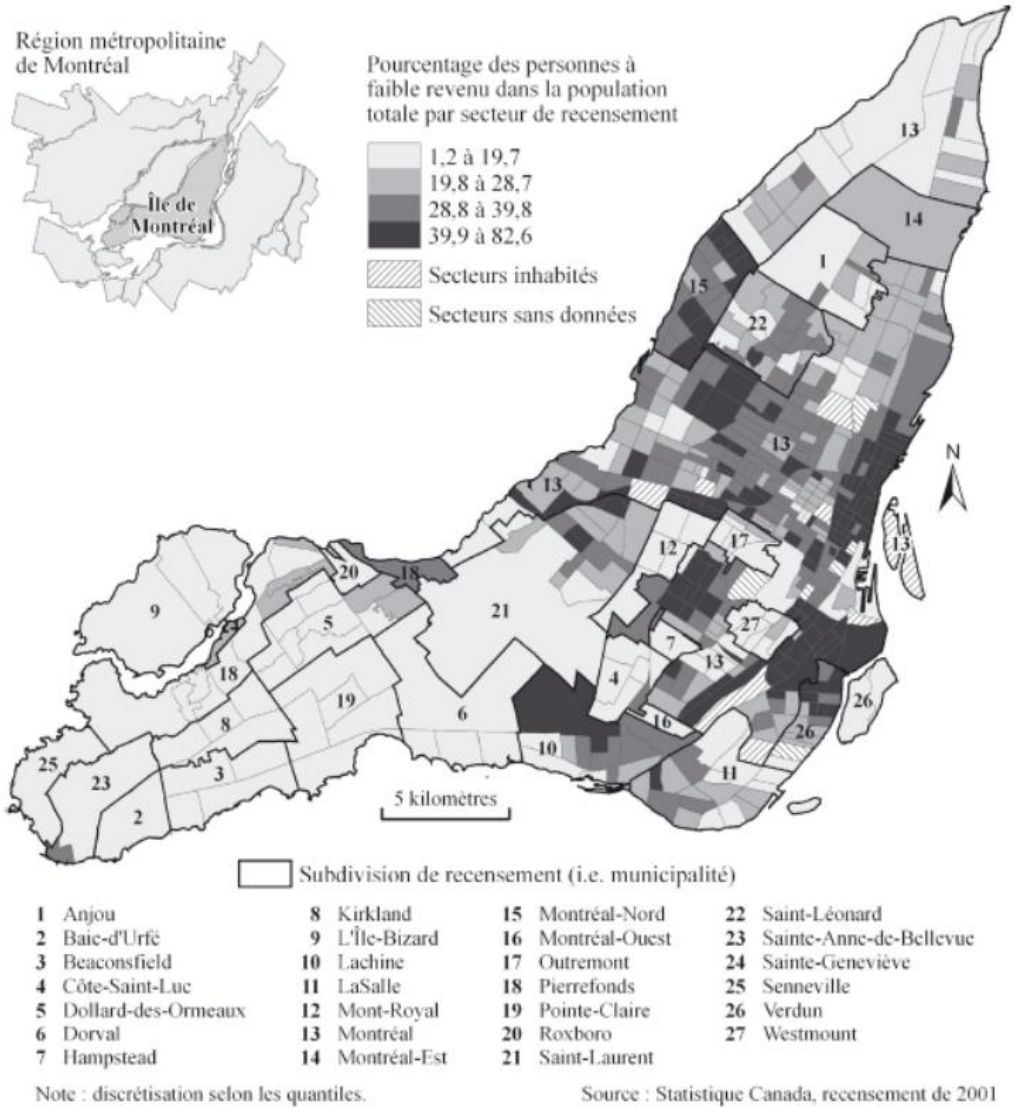
Research | Full Access

Modélisation spatiale de la pauvretéà Montréal: apport méthodologique de la régression géographiquement pondérée

PHILIPPE APPARICIO, ANNE-MARIE SÉGUIN, XAVIER LELOUP

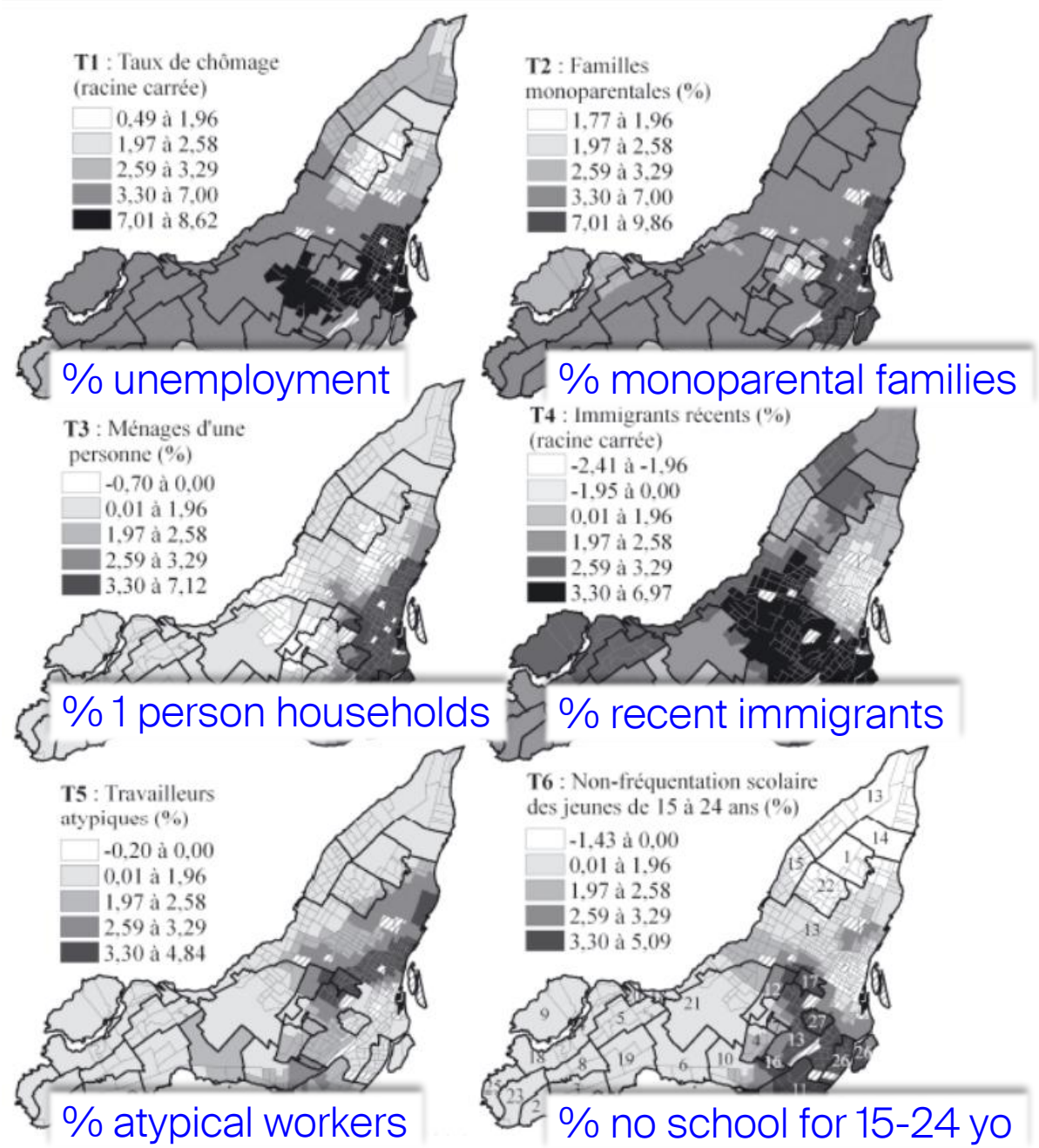
First published: 15 May 2017 | https://doi.org/10.1111/j.1541-0064.2007.00189.x | Citations: 12





Poverty in Montreal

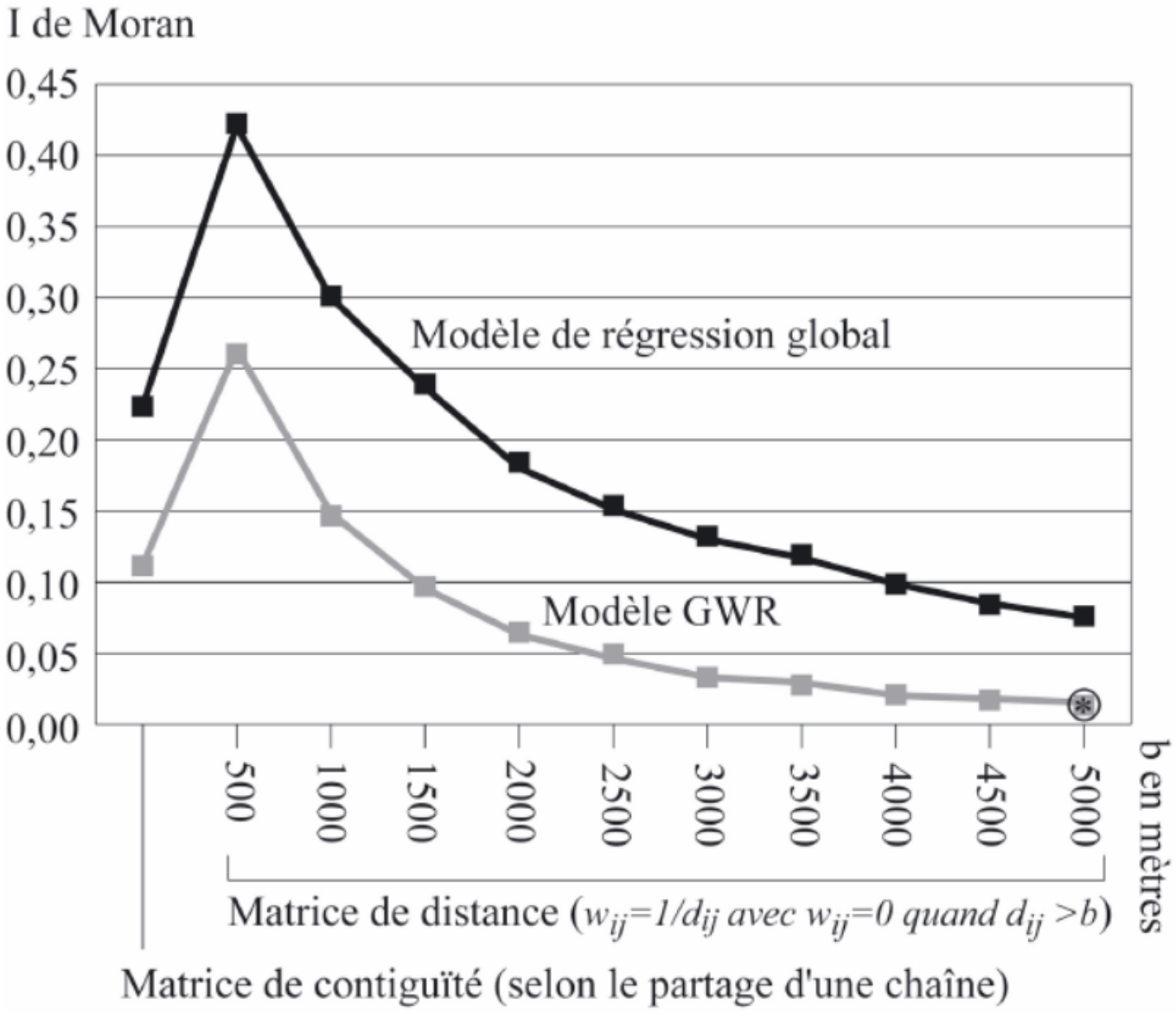
- Mapping the local significance (here Student's T) of independent variables
- A significance test provides the probability to reject H0 for a given variable
- A high T means that the corresponding variable is significant



Poverty in Montreal

Tableau 4
Analyse de variance entre les modèles de régression global et GWR

Modèle de régression	Sommes des carrés	Degrés de liberté	Carrés Moyens	F de Fisher
Résidus du modèle global	19976,5	7,00		
Modèle GWR	6514,7	53,19	122,48	
Résidus du modèle GWR	13461,8	445,81	30,20	4,0561



Conclusions on GWR

- Spatially explicit approach to measure the relationship between a dependent and a set of independent variables
- This approach was designed in order to follow the first law of geography of Tobler, while respecting assumptions of classic statistics.
- GWR employs a spatial weighting function with the assumption that near places look more similar than distant ones - **GEOGRAPHY MATTERS**
- Residuals produced by GWR are generally much lower, less autocorrelated
- The outputs are location specific, hence mappable for further analysis
- This type of local analysis makes it possible to understand georeferenced phenomena into much more details and less bias than with OLS

References

- Apparicio, P., Séguin, A.-M., & Leloup, X. (2007). Modélisation spatiale de la pauvretéà Montréal: apport méthodologique de la régression géographiquement pondérée. The Canadian Geographer / Le Géographe Canadien, 51(4), 412–427. *
- Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., (1998) Geographically weighted regression - modelling spatial non- stationarity, Journal of the Royal Statistical Society, Series D-The Statistician, 47(3), 431-443
- Brunsdon, C., Fotheringham, A.S., and Charlton, M.E. (1998) Spatial nonstationarity and autoregressive models, Environment and Planning A, 30(6), 957-993
- Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., (2002) Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Chichester: Wiley.
- Lloyd, C. (2010) Spatial Data Analysis, An Introduction for GIS users, Oxford University Press.
- Loonis, V., Bellefon, M.-P. (2018) Manuel d'analyse spatiale, INSEE Eurostat, Montrouge.
- Wang, Q., Ni, J., & Tenhunen, J. (2005). Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. Global Ecology and Biogeography, 14(4), 379–393. *


Thank you for your attention!