Spatial autocorrelation: robustness of measures and tests

Marie Ernst and Gentiane Haesbroeck

University of Liege

London, December 14, 2015

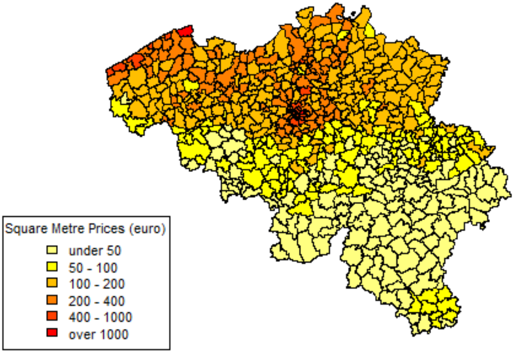
Spatial Data

Spatial data :

- geographical positions
- non spatial attributes

Example[a]

Mean price of land for sales (€/m^2) in Belgian municipalities



[a]Statistics from Belgian Federal Government - `http://statbel.fgov.be`

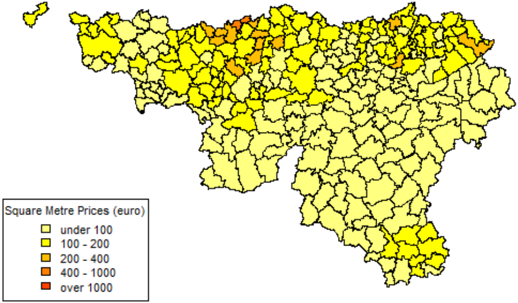
Spatial Data

Spatial data :

- geographical positions
- non spatial attributes

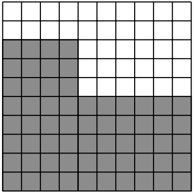
Example[a]

Mean price of land for sales (€/m^2) in Walloon municipalities

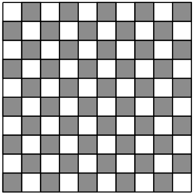


[a]Statistics from Belgian Federal Government - `http://statbel.fgov.be`

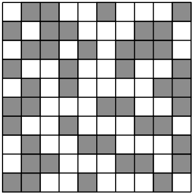
Spatial autocorrelation



Positive spatial autocorrelation



Negative spatial autocorrelation



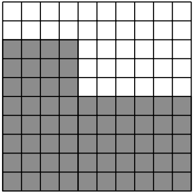
No spatial autocorrelation

Weighting matrix W

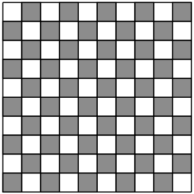
- Binary weights,
- Row-standardized,
- Globally standardized, . . .

Convention: zero diagonal and $S_0 = \sum_i \sum_j W_{ij}$

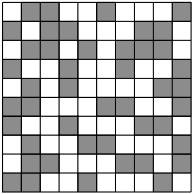
Spatial autocorrelation



Positive spatial autocorrelation



Negative spatial autocorrelation



No spatial autocorrelation

Weighting matrix W

- Binary weights,
- Row-standardized,
- Globally standardized, ...

Convention: zero diagonal and $S_0 = \sum_i \sum_j W_{ij}$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Geary's ratio (1954)

$$c(z) = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - z_j)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Geary's ratio (1954)

$$c(z) = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - z_j)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}$$

Geary's ratio (1954)

$$c(z) = \frac{n-1}{2S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(z_i - z_j)^2}{\sum_{i=1}^n (z_i - \bar{z})^2}$$

Getis and Ord's statistics (1992)

$$G(z) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} z_i z_j}{\sum_{i=1}^n \sum_{j=1, j \neq i}^n z_i z_j}$$

Measures of spatial autocorrelation

The values $z_1, \ldots, z_n$ are observed at locations $s_1, \ldots, s_n$

Moran's Index (1950)

$$I(z) = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Geary's ratio (1954)

$$c(z) = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - z_j)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Getis and Ord's statistics (1992)

$$G(z) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} z_i z_j}$$

Tests based on asymptotic normality

Without spatial autocorrelation, I, c and G are **asymptotically Gaussian** under normality (N) and/or randomisation (R) assumption.

Package spdep in R:

- `moran.test`: under N and R assumption
- `geary.test`: under N and R assumption
- `globalG.test`: under R assumption

Example

Statistic	Expected value	Variance	p-value
$I = 0.53$	-0.0038	0.0006	$< 2 \times 10^{-16}$
$c = 0.42$	1	0.0008	$< 2 \times 10^{-16}$
$G = 0.046$	0.038	2×10^{-7}	$< 2 \times 10^{-16}$

Tests based on asymptotic normality

Without spatial autocorrelation, I, c and G are **asymptotically Gaussian** under normality (N) and/or randomisation (R) assumption.

Package `spdep` in R:

- `moran.test`: under N and R assumption
- `geary.test`: under N and R assumption
- `globalG.test`: under R assumption

Example

Statistic	Expected value	Variance	p-value
$I = 0.53$	-0.0038	0.0006	$< 2 \times 10^{-16}$
$c = 0.42$	1	0.0008	$< 2 \times 10^{-16}$
$G = 0.046$	0.038	2×10^{-7}	$< 2 \times 10^{-16}$

Tests based on asymptotic normality

Without spatial autocorrelation, I, c and G are **asymptotically Gaussian** under normality (N) and/or randomisation (R) assumption.

Package `spdep` in R:

- `moran.test`: under N and R assumption
- `geary.test`: under N and R assumption
- `globalG.test`: under R assumption

Example

Statistic	Expected value	Variance	p-value
$I = 0.53$	-0.0038	0.0006	$< 2 \times 10^{-16}$
$c = 0.42$	1	0.0008	$< 2 \times 10^{-16}$
$G = 0.046$	0.038	2×10^{-7}	$< 2 \times 10^{-16}$

Inference

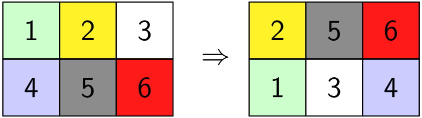
Permutation tests

Computation of the index using **random permutations** of the variable on the spatial domain.

Inference

Permutation tests

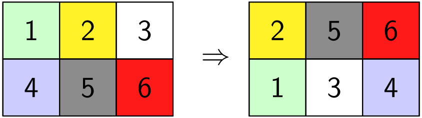
Computation of the index using **random permutations** of the variable on the spatial domain.



Inference

Permutation tests

Computation of the index using **random permutations** of the variable on the spatial domain.



$$\text{p-value} = \frac{\#\{I_{permuted} \geq I_{obs}\}}{\#\{\text{simulations}\} + 1}$$

Inference

Permutation tests

Computation of the index using **random permutations** of the variable on the spatial domain.

1	2	3
4	5	6

$\Rightarrow$

2	5	6
1	3	4

$$\text{p-value} = \frac{\#\{I_{permuted} \geq I_{obs}\}}{\#\{\text{simulations}\} + 1}$$

- **Benefits**: No assumption on the distribution
- **Drawbacks** : Simulations (randomness, computational cost,...)

Inference

Permutation tests

Computation of the index using **random permutations** of the variable on the spatial domain.

$$\text{p-value} = \frac{\#\{I_{permuted} \geq I_{obs}\}}{\#\{\text{simulations}\} + 1}$$

- **Benefits**: No assumption on the distribution
- **Drawbacks** : Simulations (randomness, computational cost,...)

Example (nsim=5000)

Statistic	Observed rank	p-value	Package `spdep` in R:
$I = 0.53$	5001	2×10^{-4}	`moran.mc`
$c = 0.42$	1	2×10^{-4}	`geary.mc`
$G = 0.046$	5001	2×10^{-4}	/

Inference

Dray's test based on I

Decomposition of positive / negative influence on autocorrelation

$$I(z) = \underbrace{\sum_{I(u_k)<E[I]} I(u_k)cor^2(u_k, z)}_{S_I^-(z)} + \underbrace{\sum_{I(u_k)>E[I]} I(u_k)cor^2(u_k, z)}_{S_I^+(z)}$$

where u_k are eigenvector of HWH, $H = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'$.

- Simultaneous positive and negative spatial autocorrelation
- Permutation test

Example (nsim=5000)

Statistic		Observed rank	p-value
$I = 0.53$	$S_I^+ = 0.57$	5001	2×10^{-4}
	$S_I^- = -0.04$		

Dray's test based on I

Decomposition of positive / negative influence on autocorrelation

$$I(z) = \underbrace{\sum_{I(u_k)<E[I]} I(u_k) cor^2(u_k, z)}_{S_I^-(z)} + \underbrace{\sum_{I(u_k)>E[I]} I(u_k) cor^2(u_k, z)}_{S_I^+(z)}$$

where u_k are eigenvector of HWH, $H = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'$.

- Simultaneous positive and negative spatial autocorrelation
- Permutation test

Example (nsim=5000)

Statistic		Observed rank	p-value
$I = 0.53$	$S_I^+ = 0.57$	5001	2×10^{-4}
	$S_I^- = -0.04$		

Inference

Moran's I	**Geary's c**	**Getis and Ord's G**
Test under R	Test under R	Test under R
Test under N	Test under N	
Permutation test	Permutation test	Permutation test
Dray's test		

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

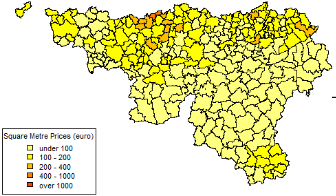
Example

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



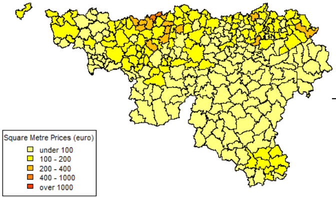
Test	p-value
I under R	$< 2 \times 10^{-16}$
I under N	$< 2 \times 10^{-16}$
I Permutation	$< 2 \times 10^{-4}$
Dray test	$< 2 \times 10^{-4}$
c under R	$< 2 \times 10^{-16}$
c under N	$< 2 \times 10^{-16}$
c Permutation	$< 2 \times 10^{-4}$
G under R	$< 2 \times 10^{-16}$
G Permutation	$< 2 \times 10^{-4}$

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



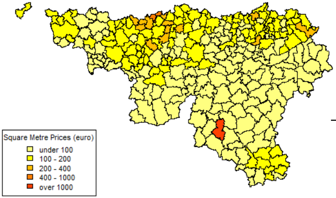
Test	Decision
I under R	RH_0
I under N	RH_0
I Permutation	RH_0
Dray test	RH_0
c under R	RH_0
c under N	RH_0
c Permutation	RH_0
G under R	RH_0
G Permutation	RH_0

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



Paliseul: 690€/ m^2

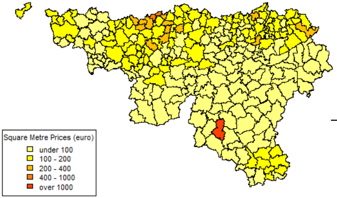
Test	Decision
I under R	RH_0
I under N	RH_0
I Permutation	RH_0
Dray test	RH_0
c under R	RH_0
c under N	RH_0
c Permutation	FTR
G under R	RH_0
G Permutation	RH_0

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



Paliseul: 730€/ m^2

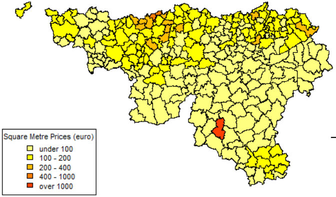
Test	Decision
I under R	RH_0
I under N	RH_0
I Permutation	RH_0
Dray test	RH_0
c under R	FTR
c under N	RH_0
c Permutation	FTR
G under R	RH_0
G Permutation	RH_0

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



Paliseul $\approx$ max(Belgium) (1200€)

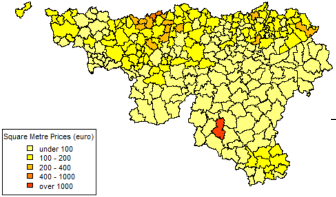
Test	Decision
I under R	FTR
I under N	RH_0
I Permutation	RH_0
Dray test	RH_0
c under R	FTR
c under N	FTR
c Permutation	FTR
G under R	RH_0
G Permutation	RH_0

Robustness of the tests

How contamination influences the result of the tests?

Impact on the decision of the test

Example



Paliseul: 1610€/ m^2

Test	Decision
I under R	FTR
I under N	FTR
I Permutation	FTR
Dray test	FTR
c under R	FTR
c under N	FTR
c Permutation	FTR
G under R	FTR
G Permutation	FTR

Robustness of the tests

Resistance of a test (Ylvisaker, 1977)

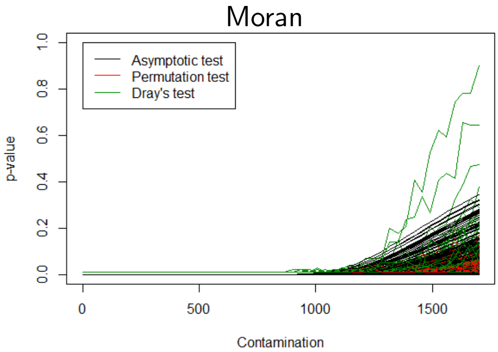
- **Resistance to acceptance**: smallest proportion of fixed observations which will always implies the *non rejection* of H_0.
- **Resistance to rejection**: smallest proportion m_0 of fixed observations which will always implies the *rejection* of H_0.

Result: Resistance to rejection of considered tests: $\frac{1}{n}$

Robustness of the tests

Impact on the p-value

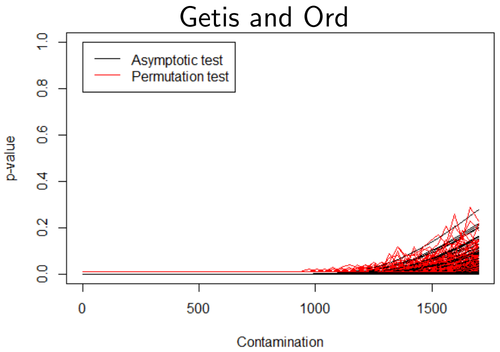
Hairplot: $\xi \mapsto$ p-value$(z + \xi e_i)$ for $i = 1, \ldots, n$



Robustness of the tests

Impact on the p-value

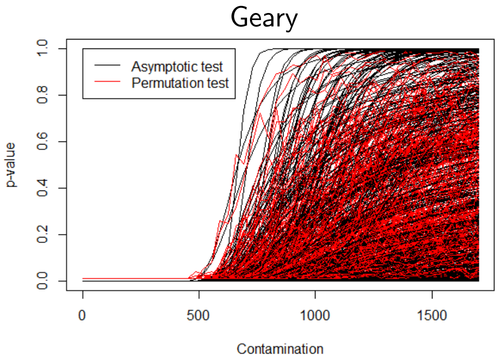
Hairplot: $\xi \mapsto \text{p-value}(z + \xi e_i)$ for $i = 1, \ldots, n$



Robustness of the tests

Impact on the p-value

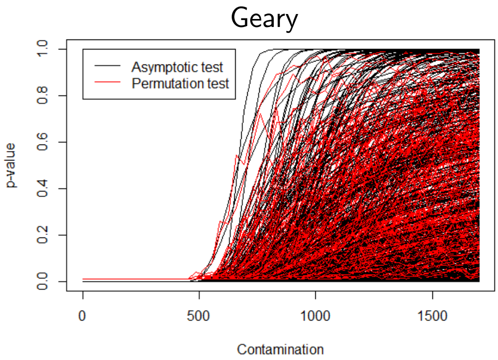
Hairplot: $\xi \mapsto \text{p-value}(z + \xi e_i)$ for $i = 1, \ldots, n$



Robustness of the tests

Impact on the p-value

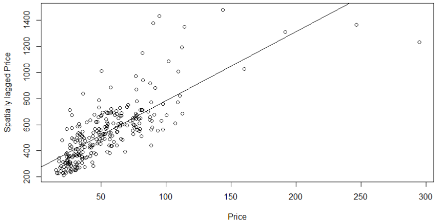
Hairplot: $\xi \mapsto$ p-value$(z + \xi e_i)$ for $i = 1, \ldots, n$



$\Rightarrow$ **robust alternative must be considered**

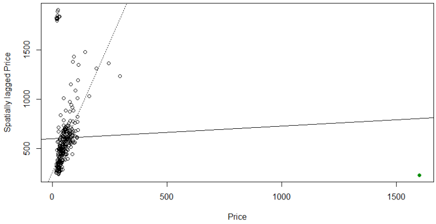
Moran scatterplot

Moran can be interpreted as the slope in a LS regression of Wz over z



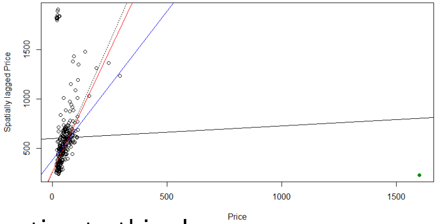
Moran scatterplot

Moran can be interpreted as the slope in a LS regression of Wz over z



Moran scatterplot

Moran can be interpreted as the slope in a LS regression of Wz over z



Idea: Robustly estimate this slope

Robust techniques

- Least Trimmed Squares (Rousseeuw, 1984)
- M-estimator (Huber, 1973)

Robust regression

Regression: $y = ax + b$ and residuals: $r_i = y_i - \hat{y}_i$

Least squares regression

Minimisation of $\sum_{i=1}^{n} r_i^2$

Least Trimmed Squares (Rousseeuw, 1984)

Minimisation of the trimmed sum: $\sum_{i=1}^{n-h} r_{(i)}^2$

M-estimator (Huber, 1973)

Minimisation of another function: $\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right)$
e.g., $\rho(r) = r^2$, $\rho(r) = |r|$

Robust regression

Regression: $y = ax + b$ and residuals: $r_i = y_i - \hat{y}_i$

Least squares regression

Minimisation of $\sum_{i=1}^{n} r_i^2$

Least Trimmed Squares (Rousseeuw, 1984)

Minimisation of the trimmed sum: $\sum_{i=1}^{n-h} r_{(i)}^2$

M-estimator (Huber, 1973)

Minimisation of another function: $\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right)$
e.g., $\rho(r) = r^2$, $\rho(r) = |r|$

Robust regression

Regression: $y = ax + b$ and residuals: $r_i = y_i - \hat{y}_i$

Least squares regression

Minimisation of $\sum_{i=1}^{n} r_i^2$

Least Trimmed Squares (Rousseeuw, 1984)

Minimisation of the trimmed sum: $\sum_{i=1}^{n-h} r_{(i)}^2$

M-estimator (Huber, 1973)

Minimisation of another function: $\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right)$
e.g., $\rho(r) = r^2$, $\rho(r) = |r|$

Robust regression

Regression: $y = ax + b$ and residuals: $r_i = y_i - \hat{y}_i$

Least squares regression

Minimisation of $\sum_{i=1}^{n} r_i^2$

Least Trimmed Squares (Rousseeuw, 1984)

Minimisation of the trimmed sum: $\sum_{i=1}^{n-h} r_{(i)}^2$

M-estimator (Huber, 1973)

Minimisation of another function: $\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right)$
e.g., $\rho(r) = r^2$, $\rho(r) = |r|$

Robust inference based on Moran

Impact on the decision of the test: resistance

Linked to the Breakdown point of the robust regression

Least Trimmed Squares	M-estimator
h/n	≤ 0.5

Impact on the p-value: hairplot

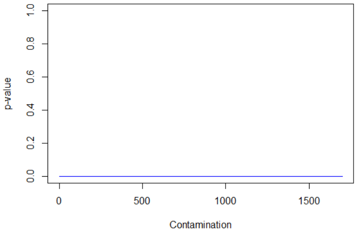
Robust inference based on Moran

Impact on the decision of the test: resistance

Linked to the Breakdown point of the robust regression

Least Trimmed Squares	M-estimator
h/n	≤ 0.5

Impact on the p-value: hairplot



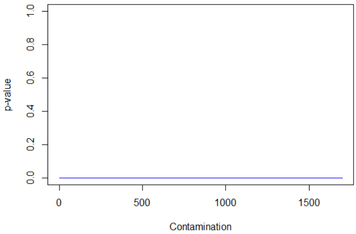
Robust inference based on Moran

Impact on the decision of the test: resistance

Linked to the Breakdown point of the robust regression

Least Trimmed Squares	M-estimator
h/n	≤ 0.5

Impact on the p-value: hairplot



What about level and power?

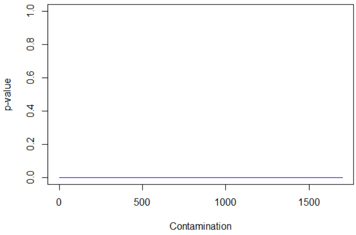
Robust inference based on Moran

Impact on the decision of the test: resistance

Linked to the Breakdown point of the robust regression

	Least Trimmed Squares	M-estimator
	h/n	≤ 0.5

Impact on the p-value: hairplot



What about level and power? $\Rightarrow$ simulations

Simulations

Model

The Spatial autoregressive (SAR) model: $Z = \rho W Z + \varepsilon$

Simulation settings (Holmberg and Lundevaller, 2015)

- Spatial domain: grid ($n = 100, 400, 900, 1600$)
- Weighting matrix: queen and rook contiguity

- Spatial correlation coefficient: $\rho = 0, 0.1, 0.2, 0.3$
- Distribution of ε: $N(0, 1)$, $\mathcal{P}(1)$, Bern(0.5)

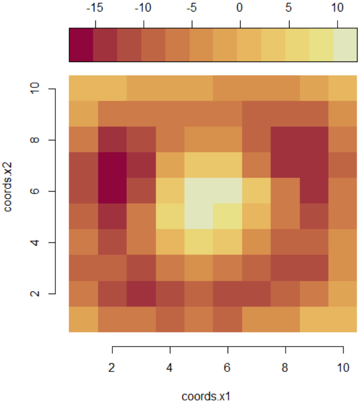
Contamination process

A number δ of observations are contaminated

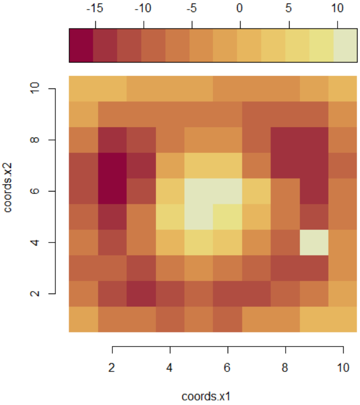
Example

$n = 100$, queen contiguity, $\rho = 0.2$ and $\varepsilon \sim N(0, 1)$

$\delta = 0$



$\delta = 1$



Comparison

1. Without contamination ($\delta = 0$)

Similar results are expected for classical and robust tests

2. With contamination ($\delta > 0$)

Robust tests should resist to contamination and be powerful

Results ($\delta = 0$)

Robust tests vs initial tests

Good correlation between the p-value of robust and classical Moran's tests ($\geq 90\%$)

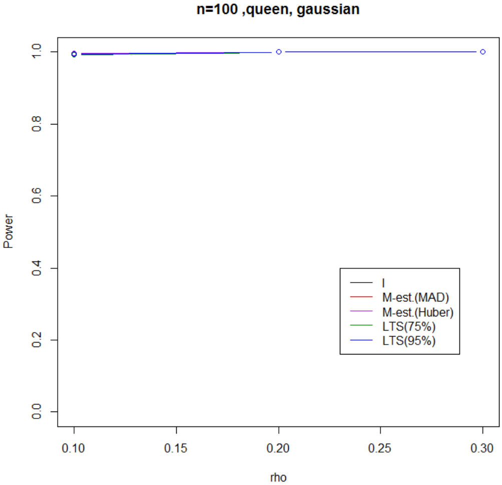
Level ($n = 100$)

ε	W	Moran	M-estimator		LTS	
			MAD	Huber	75%	95%
N(0,1)	Queen	0.055	0.061	0.062	0.072	0.073
N(0,1)	Rook	0.051	0.051	0.050	0.053	0.054
$\mathcal{P}(1)$	Queen	0.044	0.054	0.053	0.080	0.063
$\mathcal{P}(1)$	Rook	0.052	0.056	0.055	0.057	0.056
Bern(0.5)	Queen	0.067	0.086	0.087	0.091	0.091
Bern(0.5)	Rook	0.041	0.061	0.050	0.076	0.076

Results ($\delta = 0$)

Power

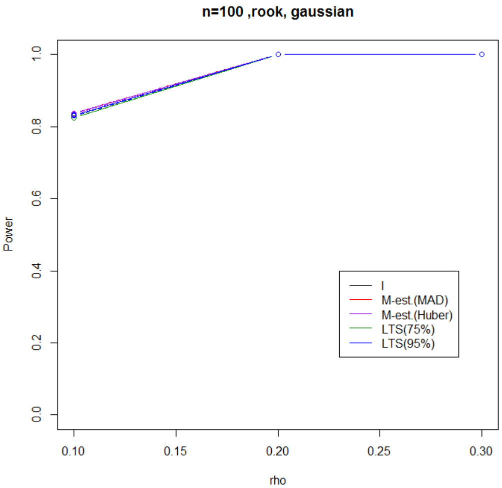
proportion of rejected null hypothesis with $\rho = 0.1$, 0.2 or 0.3



Results ($\delta = 0$)

Power

proportion of rejected null hypothesis with $\rho = 0.1$, 0.2 or 0.3



Results ($\delta > 0$)

500 simulations for $n = 100$, $\varepsilon \sim N(0,1)$

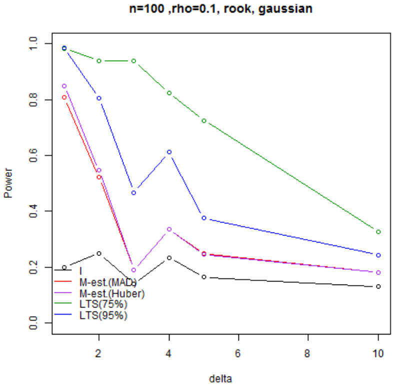
Level

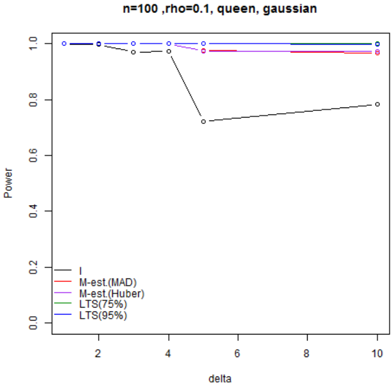
δ	W	Moran	M-estimator		LTS	
			MAD	Huber	75%	95%
1	Queen	0.064	0.072	0.072	0.140	0.126
1	Rook	0.026	0.030	0.030	0.094	0.046
2	Queen	0.050	0.068	0.068	0.134	0.090
2	Rook	0.040	0.052	0.052	0.096	0.064
3	Queen	0.054	0.062	0.060	0.104	0.090
3	Rook	0.054	0.076	0.076	0.086	0.098
4	Queen	0.026	0.048	0.048	0.092	0.064
4	Rook	0.050	0.064	0.062	0.090	0.080
5	Queen	0.058	0.072	0.070	0.124	0.126
5	Rook	0.040	0.058	0.058	0.062	0.054
10	Queen	0.052	0.060	0.060	0.072	0.070
10	Rook	0.068	0.074	0.070	0.078	0.068

Results ($\delta > 0$)

500 simulations for $n = 100$, $\varepsilon \sim N(0,1)$

Power

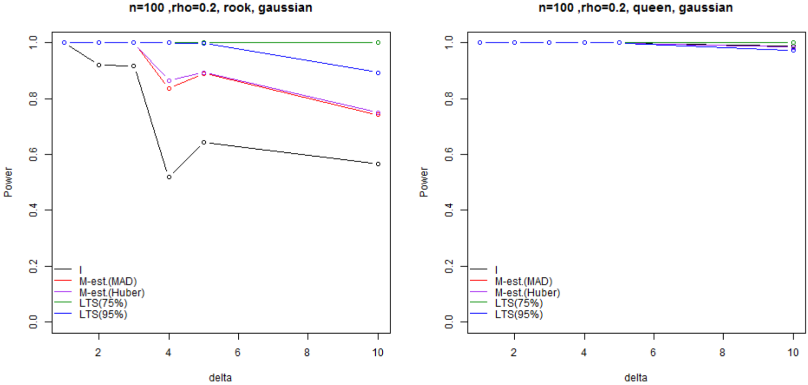




Results ($\delta > 0$)

500 simulations for $n = 100$, $\varepsilon \sim N(0, 1)$

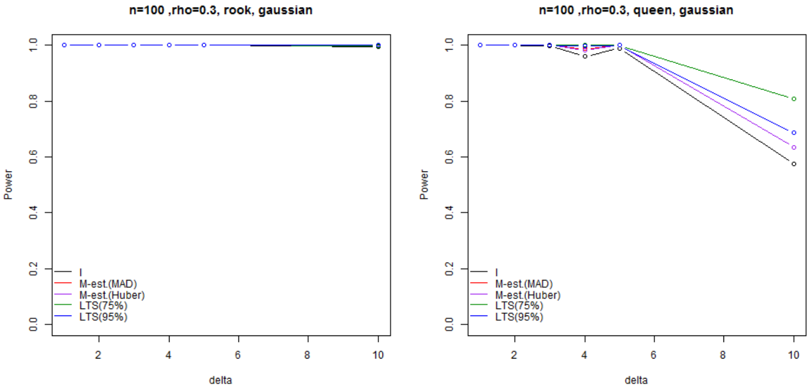
Power



Results ($\delta > 0$)

500 simulations for $n = 100$, $\varepsilon \sim N(0, 1)$

Power



Results ($\delta > 0$)

500 simulations for $n = 100$, $\varepsilon \sim N(0, 1)$

Preliminary observations

- Robust test using Least Trimmed Square gives good results
- Robust test using M-estimator is more sensitive to leverage points

Perspective

In progress

- Simulations and comparison need to pursue
- Proposition of a robust test based on the others statistics

Perspective

- Study robustness of local measures and local tests

In progress

- Simulations and comparison need to pursue
- Proposition of a robust test based on the others statistics

  For Geary's c:
  - $\hat{\gamma}(h) = S_z^2 c$ for $\hat{\gamma}(h)$ the empirical variogram
  - Robust estimation of the variogram using Q_n
  - $\hat{c} \propto \frac{1}{2}\left(\frac{Q_n(\mathcal{V}_n)}{Q_n(z)}\right)^2$

Perspective

- Study robustness of local measures and local tests

References

Spatial autocorrelation:

- Dray (2011). A new perspective about Moran's coefficient [...]. *Geogr. anal.*
- Geary (1954). The contiguity ratio and statistical mapping, *The inc. Statistician*.
- Getis and Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*.
- Holmberg and Lundevaller (2015). A test for robust detection of residual spatial autocorrelation [...]. *Spatial Statistics*.
- Moran (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*.

Robustness:

- Genton (1998). Spatial BDP of variogram estimators. *Math Geol*.
- Genton and Lucas (2003). Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations. *J R Stat Society*.
- Genton and Ruiz-Gazen (2010). Visualizing influential observations in dependent data. *Journal of computational and graphical statistics*.
- Huber (1973). Robust regression [...]. *Annals of Statistics*.
- Lark (2008). Some Results on the Spatial Breakdown Point of Robust Point Estimates of the Variogram. *Mathematical geosciences*.
- Rousseeuw (1984). Least Median of Squares Regression. *J Am stat assoc*.
- Ylvisaker (1977). Test resistance. *J Am stat assoc*.