DEPARTAMENTO DE ESTATÍSTICA, ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

Nonparametric tests to compare the first-order structure of inhomogeneous spatial point processes

I. Fuentes-Santos, W. González-Manteiga and J. Mateu

Modelos de Optimización, Decisión, Estadística y Aplicaciones

Santiago de Compostela, 10-12-2019

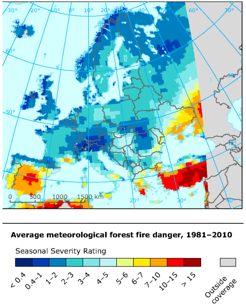
Outline

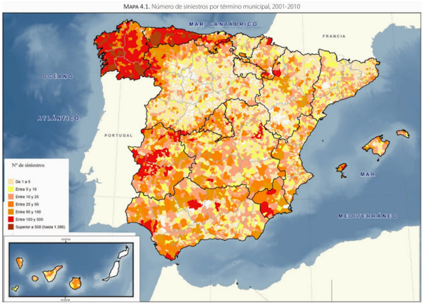
Introduction

Real data problems

1. Introduction
   - Real data problems
   - Spatial Point Processes

Wildfires in Galicia

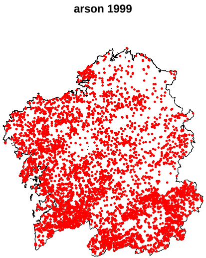


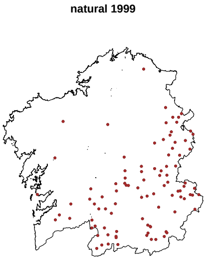


- High wildfire incidence in the North-West of Spain.
- Climate conditions are not the reason.
- **Particular features**
  - $>$ **80**% of wildfires are **arson**.
  - $<$ 5% of wildfires have natural cause.
  - $\approx$ 70% of wildfires affected < 1 ha.
- **Aim**: understand the behavior of wildfires to improve fire prevention and fighting plans.

Wildfire dataset

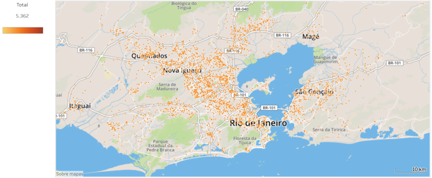
- Spatial location and time of ignition of the **104225** wildfires registered in Galicia in the period 1999-2014. Classified by burned area and cause of ignition.
- **Do arson and natural wildfires have the same spatial distribution?**





Gun violence in Rio de Janeiro

- Rio de Janeiro Metropolitan area have suffered a continuous increase in violent crimes during the last decades.
- More than 5000 deaths by firearms in the MR during 2017.

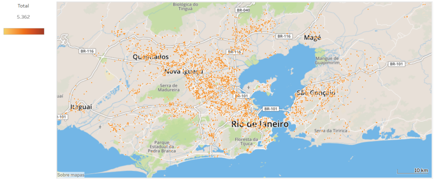


Source: Instituto de Seguranza Publica de Rio de Janeiro (ISP-RJ)

(http://www.isp.rj.gov.br/)

Gun violence in Rio de Janeiro

- Rio de Janeiro Metropolitan area have suffered a continuous increase in violent crimes during the last decades.
- More than 5000 deaths by firearms in the MR during 2017.



Source: Instituto de Seguranza Publica de Rio de Janeiro (ISP-RJ)

(http://www.isp.rj.gov.br/)

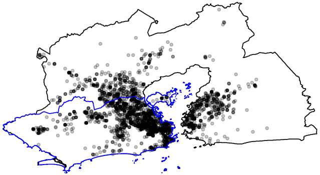
- The **Fogo cruzado** app was released in 2016 with two aims
  - Help Rio residents to avoid stray bullets.
  - Create a database with the gunfire reports collected by *Fogo cruzado*.

https://fogocruzado.org.br Link

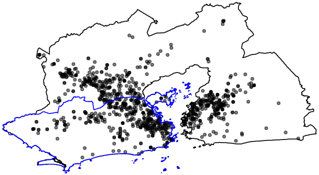
FOGO CRUZADO records 2017

- 5945 gunfires recorded in the Río de Janeiro metropolitan area during 2017.
- Information provided:
  - GPS coordinates, date and time of occurrence.
  - Indicator of police presence.
  - Number of police and civil mortal or injured victims, if any.
- **Do gunfire with and without mortal victims have the same spatial distribution?**

Without mortal victims



With mortal victims

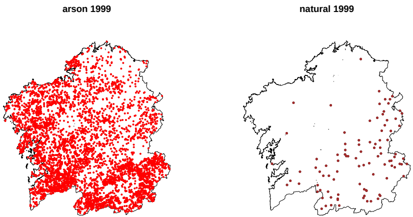


1. Introduction
   - Real data problems
   - Spatial Point Processes

Spatial Point Processes

- A **spatial point processes (SPP)** is a stochastic process governing the location of a finite number of **events** $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ in $\mathbb{R}^2$.
- A **spatial point pattern** is a realization of a SPP, commonly observed on a bounded domain.
- A SPP can be **marked** and/or depend on **covariates**.
- A **MULTITYPE SPP** is a SPP with categorical marks defining different types of events.

Example: Arson and natural wildfires in Galicia (1999-2014).



First-order intensity function

- Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ be a spatial point pattern.
- The **first-order intensity function**

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E\left[N(dx)\right]}{|dx|} \right\}$$

- Intuitively, $\lambda(x)|dx|$ is the probability for dx to contain exactly one event of $\mathbf{X}$.
- A point process is **homogeneous** if its first-order intensity is constant, $\lambda(x) = \lambda > 0$, and **inhomogeneous** otherwise.
- **Inhomogeneous spatial Poisson point process (IPP)**
  - Inhomogeneous intensity.
  - Independent events (Poisson)

Kernel density of event locations

The kernel intensity estimator is **inconsistent**.

SPP intensity	Density in $\mathbf{R}^2$
$\hat{\lambda}_H(x) = p_H(x)^{-1}\|H\|^{-1/2}\sum_{i=1}^{N} k\left(H^{-1/2}(x-X_i)\right)$	$\hat{f}_H(x) = N^{-1}\|H\|^{-1/2}\sum_{i=1}^{N} k\left(H^{-1/2}(x-X_i)\right)$

Consistent estimator

- Cuccala (2006) defined the **density of event locations** as $\lambda_0(x) = \lambda(x)/m$, where $m = \int_W \lambda(x)dx$. The kernel estimator of $\lambda_0(\cdot)$ is

$$\hat{\lambda}_{0,H}(x) = (p_H(x)\mathbf{N})^{-1}|H|^{-1/2}\sum_{i=1}^{N} k\left(H^{-1/2}(x-\mathbf{x_i})\right)\mathbf{I}[\mathbf{N} \neq \mathbf{0}]$$

- The Bandwidth matrix, H, is selected by a **PLUG IN** algorithm (Fuentes-Santos *et al*, 2016).

CUCALA, L. (2006). *Espacements bidimensionnels et données entachées d´erreurs dans l´analyse des processus ponctuels spatiaux*. PhD thesis, Université des Sciences Sociales, Toulouse I.

FUENTES-SANTOS, I., GONZÁLEZ-MANTEIGA, W., MATEU, J. (2016) Consistent Smooth Bootstrap Kernel Intensity Estimation for Inhomogeneous Spatial Poisson Point Processes. Scandinavian Journal of Statistics, 43(2): 416-435 .

- The intensities of two point processes, $\mathbf{X_1}$ and $\mathbf{X_2}$, with the same spatial structure are proportional,

$$\mathcal{H}_0 : \lambda_1(x) = \omega\lambda_2(x)$$

- **Nonparametric tests**
  - Kolmogorov-Smirnov type test (Zhang and Zhuang, 2017)
  - Cramer von Mises type statistic (Fuentes-Santos *et al.*, 2017).
  - Regression test based on the relative risk function. Analogous to the log-ratio based separability test (Fuentes Santos *et al.*, 2018).

Fuentes-Santos, I., González-Manteiga, W., Mateu, J. (2017). A nonparametric test for the comparison of first-order structures of spatial point processes. Spatial Statistics, 22(2): pp, 240-260.

Fuentes-Santos, I., González-Manteiga, W., Mateu, J. (2018) A first-order ratio-based nonparametric separability test for spatiotemporal point processes. Environmetrics, 29(1), e2482.

Zhang, T., Zhuang, R. (2017). Testing proportionality between the first-order intensity functions of spatial point processes. Journal of Multivariate Analysis, 155, 72-82.

Nonparametric comparison of SPP

Kolmogorov-Smirnov test

2. Nonparametric comparison of SPP
   - Kolmogorov-Smirnov test
   - Cramer von Mises test
   - Relative-risk based regression test

Kolmogorov-Smirnov test

- If two point processes, $\mathbf{X_1} = \{\mathbf{x_i}\}_{i=1}^{N_1}$ and $\mathbf{X_2} = \{\mathbf{x_j}\}_{j=N_1+1}^{N}$, have the same spatial structure, then $\exists \omega > 0$

$$\mathcal{H}_0 : \lambda_1(x) = \omega \lambda_2(x); \ \forall x \in W$$

- Equivalently, under $\mathcal{H}_0$ there exists a $\omega > 0$ such that for any Borel set $A \in \mathcal{B}(W)$, $\mathbb{E}[N_1(A)] = \omega \mathbb{E}[N_2(A)]$.
- The proportionality parameter ω in $\mathcal{H}_0$ can be estimated as $\hat{w} = N_1(W)/N_2(W)$
- Let

$$D_{\hat{w}}(A) = N_1(A) - \hat{\omega} N_2(A) = N_1(W)\left[\frac{N_1(A)}{N_1(W)} - \frac{N_2(A)}{N_2(W)}\right]$$

under $\mathcal{H}_0$, $|D_{\hat{w}}(A)|$ is close to 0 for any $A \in \mathcal{B}(W)$.

The test statistic

- For a given $\pi-$system, $\mathcal{P}$, we define the **test statistic**

$$\hat{T} = \frac{1}{\zeta}\sqrt{\frac{N_1(W)\,N_2(W)}{N_1(W)+N_2(W)}}\sup_{A\in\mathcal{P}}\left|\frac{N_1(A)}{N_1(W)}-\frac{N_2(A)}{N_2(W)}\right|$$

where the normalizing constant ζ is estimated as.

$$\hat{\zeta}^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left[\frac{\left(N_1(W_i)-\hat{N}_1(W_i)\right)^2}{\hat{N}_1(W_i)} + \frac{\left(N_2(W_i)-\hat{N}_2(W_i)\right)^2}{\hat{N}_1(W_2)}\right]$$

- for a partition $\{W_i\}_{i=1}^{K}$, $\hat{N}_1(W_i) = \hat{\omega}\,(N_1(W_i) - N_2(W_i))\,/\,(1+\hat{\omega})$, and $\hat{N}_2(W_i) = \hat{N}_1(W_i)\,/\hat{\omega}$.
- Zhang and Zhuang (2017) proved that the null distribution of $\hat{T}$ converges to a Brownian bridge.

Zhang, T., Zhuang, R. (2017). Testing proportionality between the first-order intensity functions of spatial point processes. Journal of Multivariate Analysis, 155, 72-82.

Cramer von Mises test

2. Nonparametric comparison of SPP
   - Kolmogorov-Smirnov test
   - Cramer von Mises test
   - Relative-risk based regression test

Cramer von Mises test

- If two point processes, $\mathbf{X_1}$ and $\mathbf{X_2}$, have the same spatial structure $\Rightarrow$ their densities of event locations are equal. Thus

$$\mathcal{H}_0 : \lambda_{01}(x) = \lambda_{02}(x)$$

- Conditional on $N_1 = n_1$ and $N_2 = N - N_1 = n_2$, $\mathbf{X_1}$ and $\mathbf{X_2}$ are random samples of the bivariate distributions with densities $\lambda_{01}(\cdot)$ and $\lambda_{02}(\cdot)$.
- Following Duong *et al.*, (2012) we propose a squared discrepancy measure as test statistic (Fuentes-Santos *et al.*, 2017)

$$T = \int_W \left(\lambda_{01}(x) - \lambda_{02}(x)\right)^2 dx = \psi_{0,1} + \psi_{0,2} - (\psi_{0,12} + \psi_{0,21})$$

where $\psi_{0,j} = \int_W \lambda_{0j}(x)^2\, dx$ and $\psi_{0,ij} = \int_W \lambda_{0i}(x)\, \lambda_{0j}(x)\, dx$, for $j = 1, 2$.

Duong, T., Goud, B., and Schauer, K. (2012) Closed-form density-based framework for automatic detection of cellular morphology changes.. Proceedings of the National Academy of Sciences of the United States of America, 109(22): 8382-8387

Fuentes-Santos, I., González-Manteiga, W., Mateu, J. (2017) A nonparametric test for the comparison of first-order structures of spatial point processes. Spatial Statistics, 22(2): pp. 240-260.

Test statistic

- Our test statistic is:

$$\hat{T} = \hat{\psi}_{0,1} + \hat{\psi}_{0,2} - \left(\hat{\psi}_{0,12} + \hat{\psi}_{0,21}\right)$$

where

$$\hat{\psi}_{0,1} = \frac{1}{n_1^2}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_1} k_{G_1}\left(\mathbf{x_{i_1}} - \mathbf{x_{i_2}}\right), \; \hat{\psi}_{0,2} = \frac{1}{n_2^2}\sum_{j_1=n_1+1}^{n}\sum_{j_2=n_1+1}^{n} k_{G_2}\left(\mathbf{x_{j_1}} - \mathbf{x_{j_2}}\right)$$

$$\hat{\psi}_{0,12} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} k_{G_1}\left(\mathbf{x_i} - \mathbf{x_j}\right), \; \hat{\psi}_{0,21} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} k_{G_2}\left(\mathbf{x_i} - \mathbf{x_j}\right)$$

- **Calibration**:
  - $\hat{T} \to N(\mu_T, \sigma_T)$ under $\mathcal{H}_0$.
  - Smooth bootstrap.

Bootstrap calibration

- **Algorithm**

  1. Compute the test statistic $\hat{T}$ for the observed patterns $\mathbf{X_1}$ and $\mathbf{X_2}$.
  2. Estimate the first-order intensity, $\hat{\lambda}_H(x)$, of the unmarked pattern $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}\}$.
  3. For $b = 1, \ldots, B$:
     - 3.1 Generate a bivariate spatial point process $\mathbf{X_b^*} = \{\mathbf{X_{1,b}^*}, \mathbf{X_{2,b}^*}\}$ where for $j = 1, 2$, $\mathbf{X_{j,b}^*}$ are realizations of spatial Poisson point processes with first-order intensities $n_j \hat{\lambda}_{0,H}(x)$, being n_j the number of event in $\mathbf{X_j}$.
     - 3.2 Compute $\hat{T}_b^*$.
  4. Obtain the **empirical p-value** according to the relative position of $\hat{T}$ in the ordered sample $\hat{T}_{(b)}^*$, $b = 1, \ldots, B$.

- Use plug-in algorithms to obtain the bandwidth matrices H and $G_j, j = 1, 2$.

Relative-risk based regression test

2. Nonparametric comparison of SPP
   - Kolmogorov-Smirnov test
   - Cramer von Mises test
   - Relative-risk based regression test

Relative-risk based regression test

- Let $\mathbf{X}$ be a bivariate point process with type 1 (cases), $\mathbf{X_1} = \{\mathbf{x_i}\}_{i=1}^{N_1}$, and type 2 (controls), $\mathbf{X_2} = \{\mathbf{x_j}\}_{j=N_1+1}^{N}$, events.
- If $\mathbf{X_1}$ and $\mathbf{X_2}$ have the same first-order structure, then the relative risk of observing a case ($\mathbf{x} \in \mathbf{X_1}$) is spatially invariant.

$$r(x) = \frac{\lambda_{01}(x)}{\lambda_{02}(x)} = \frac{f(x)}{g(x)}$$

Conditional on $N_j = n_j$, $\mathbf{X_j}$ and $\mathbf{X_2}$ are random samples of the distributions with densities f and g, respectively.

- The log relative risk functions $\log \rho(x) = \log(f(x)/g(x))$ can be estimated as follows

$$\hat{\rho}(x) = log\left(\frac{\hat{f}_h(x) + \delta}{\hat{g}_h(x) + \delta}\right),$$

where $\hat{f}_h(x)$ and $\hat{g}_h(x)$ are kernel estimators, with bandwidth h, and δ is a stabilizing constant.

The test statistic

- In the regression framework (Bowman & Azzalini, 1997)
  - Response variable $\{y_i = \hat{\rho}(\mathbf{x_i});\ i = 1, \ldots, n\}$
  - Explanatory variable $\{x_i;\ i = 1, \ldots, n\}$
- We should discriminate between the following competing hypotheses

$$\mathcal{H}_0 : E(y_i) = \mu \rightarrow \bar{y} = \sum_{i=1}^{n} y_i$$

$$\mathcal{H}_1 : E(y_i) = m(x_i) \rightarrow \hat{m}(x_1, x_2) = \frac{\sum\limits_{i=1}^{n} w_{g_1}(x_{i1} - x_1) w_{g_2}(x_{i2} - x_2) y_i}{\sum_{i=1}^{n} w_{g_1}(x_{i1} - x_1) w_{g_2}(x_{i2} - x_2)}$$

- **Test statistic**

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/df_1}$$

  - RSS_0 and RSS_1 are the residual sums of squares for $\bar{y}$ and $\hat{m}(x)$.
  - df_0 and df_1 denote the degrees of freedom for error under each hypothesis.

Bowman, A. W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Statistical Science Series 18.

Implementation

- Bandwidth selection for $\hat{\rho}(x)$ (LSCV).
- Bandwidth selection for $\hat{m}(x)$ (CV).
- Calibration method
  - χ^2 approximation if the errors are normal.
  - Permutation test: under $\mathcal{H}_0$ the pairing of any particular x and y is completely random.
  - Smooth bootstrap.

Permutation test

- Simulate random pairings of the observed values of X and Y.
- Compute F for each simulated pairing.
- The **empirical p-value** of the test is the proportion of simulated F-statistics larger than that obtained from the observed data.

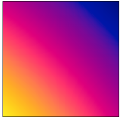
Simulation study

Simulation study

- 1000 realizations of multitype inhomogeneous Poisson and non-Poisson point processes with $m = 500$.
- Test different degrees of departure from $\mathcal{H}_0$.
- Test whether the asymmetry in the number of events in $\mathbf{X_1}$ and $\mathbf{X_2}$ affects the test.
- Two $\pi-$systems in the KS-test for any $W = [u_1, u_2] \times [v_1, v_2]$:
  - KS_1: $[u_1, t_1] \times [v_1, t_2]$ for $(t_2, t_2) \in W$
  - KS_2: $[u_1, t] \times [v_1, t]$ for $(t, t) \in W$
- Type 1 error under $\mathcal{H}_0$ and power under $\mathcal{H}_1$ with $\alpha = 0.05$.
- **Software**: R-packages `spatstat`, *ks* and *sm*.

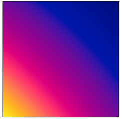
Simulation study -Poisson point process

A



B

C



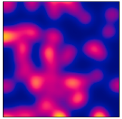
$W = [-10, 10]^2$

		KS test		T test	F test	
		KS_1	KS_2	SB	PT	SB
$m_1 = 250$	A - A	0.035	0.040	0.048	0.060	0.060
$m_2 = 250$	A - B	0.134	0.120	0.104	0.310	0.368
	A - C	0.356	0.423	0.546	0.358	0.728
$m_1 = 500/3$	A - A	0.040	0.041	0.042	0.058	0.064
$m_1 = 1000/3$	A - B	0.141	0.122	0.088	0.166	0.394
	A - C	0.432	0.400	0.334	0.870	0.762

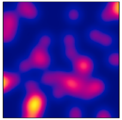
- The KS-test (K $= 6$) is slightly conservative.
- Better calibration with the T-test for symmetric designs.
- The $F-$test with bootstrap calibration outperforms its competitors in terms of power.

Simulation study - clustered point process

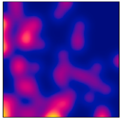
A



B



C



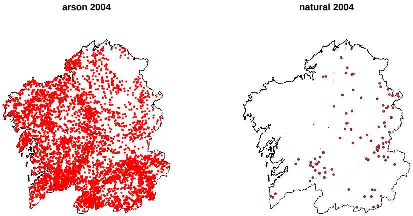
		KS test		T test	F test	
		KS_1	KS_2	SB	PT	SB
$m_1 = 250$	A - A	0.046	0.042	0.094	0.044	0.040
$m_2 = 250$	A - B	0.004	0.002	1.000	1.000	1.000
	A - C	0.392	0.376	1.000	1.000	1.000
$m_1 = 500/3$	A - A	0.040	0.041	0.042	0.058	0.064
$m_1 = 1000/3$	A - B	0.642	0.798	1.000	1.000	1.000
	A - C	0.022	0.004	1.000	1.000	1.000

- The T-test for symmetric designs is anticonservative.
- The KS-test fails in the detection of some alternative hypothesis.
- High power with the T and F tests.

$W = [-10, 10]^2$

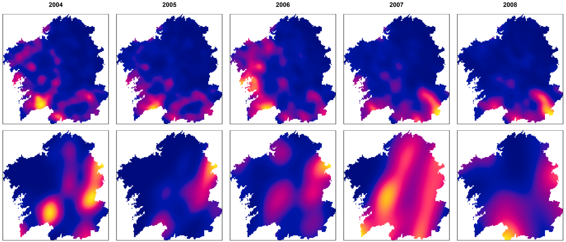
Real data problems

Comparison of wildfire patterns



- **Data**: Arson and natural wildfires from 2004 to 2008
- **Null hypotheses:**:the spatial distribution of wildfires does not depend on their cause.
- **To implement the tests**:
  - **KS**: Adapt partitions in $\hat{\zeta}$, K, to the sparseness of events.
  - **T**: Kernel estimators with 2-stages plug-in bandwidth matrices.
  - **F**: CV bandwidths in the relative-risk and kernel regression estimators.
  - **T and F**: $B = 200$ for calibration.

Arson vs Natural

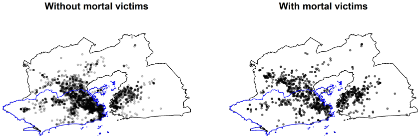


	2004	2005	2006	2007	2008
KS_1	>0.05	>0.05	>0.05	>0.05	>0.05
T	<0.005	<0.005	<0.005	<0.005	<0.005
F	<0.005	<0.005	<0.005	<0.005	<0.005

- The KS-test (K = 3) does not detect differences between arson and natural fires.
- T and F reject the null hypothesis.

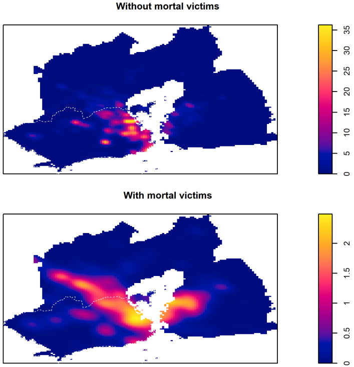
Comparison of gunfire patterns

- **5945** gunfires recorded in the Río de Janeiro metropolitan area during 2017. 1141 with mortal victims.



- **Null hypotheses**: gunfire with and without mortal victims have the same spatial distribution.
- **To implement the tests**:
  - **KS**: Adapt partitions in $\hat{\zeta}$, K, to the sparseness of events.
  - **T**: Kernel estimators with 2-stages plug-in bandwidth matrices.
  - **F**: CV bandwidths in the relative-risk and kernel regression estimators.
  - **T and F**: $B = 200$ for calibration.

Comparison of gunfire patterns



KS_1	KS_2	T	F
<0.05	>0.05	<0.005	<0.005

- The results of the KS test depend on the $\pi-$system used.
- T and F reject the null hypothesis.

Conclusions

Conclusions

- **Simulation study**
  - The three tests have a good performance under $\mathcal{H}_0$
  - The **KS-test** does not detect some alternative hypothesis in non-Poisson point processes.
  - The **relative risk based test** is the best in terms of power.
- **Application to real data**
  - Data sparseness limits the performance of the **KS-test**.
  - The spatial distribution of wildfires depends on their cause.
  - Gunfire with and without mortal victims have different spatial structure.
- **Potential limitations:**
  - We need to check and improve, if needed, the performance of T and F for small point processes.
  - T and F have **HIGH COMPUTATIONAL COST** for large datasets.

DEPARTAMENTO DE ESTATÍSTICA,
ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

Nonparametric tests to compare the first-order structure of inhomogeneous spatial point processes

I. Fuentes-Santos, W. González-Manteiga and J. Mateu

Modelos de Optimización, Decisión, Estadística y Aplicaciones

UNIVERSITAT JAUME I

Santiago de Compostela, 10-12-2019