IndTestPP: An R Package for Testing Independence between Point Processes in Time

Ana C. Cebrián

Dpto. Métodos Estadísticos. University of Zaragoza (Spain)
E-mail: acebrian@unizar.es

Introduction: aim of IndTestPP package

Tests of independence between point processses

- POISSON family
- CLOSE family
- CROSS family

Other tools

Application: occurrence of extreme heat events in three locations

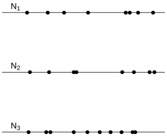
Conclusions

Introduction

Many real problems require to model the occurrence of events over time. Examples: the arrival of claims in an insurance company, the occurrence of earthquakes in a region, floods in a river, etc.

Point processes (PP) in time are a useful tool to model those phenomena. If the intensity of occurrence is constant over time, λ, the process is homogeneous, if it is a function of time, $\lambda(t)$, the process is non homogeneous. Poisson process is the most known example of PP.

Real situations often involve several PP: timing of different types of trades in stock exchange, occurrence of climate extremes in different locations.



The study of independence between PP is an importat issue in statistical modelling problems.

- To determine if processes can be modelled separately (independent PPs), or more complex models are needed (dependent PPs). Ex.: dependent multi-sites processes will require spatial models,
- To study the origin of dependence between PPs: Influence of the same or dependent variables on several PPs yield dependence, which can be modelled allowing parameters of the marginal PPS to be a function of the covariates. To analyse if the dependence is well captured by the covariates, hypothesis of independent PPs given the covariates has to be checked.

IndTestPP is an R package providing tools used in the statitical modelling of real problems involving several PPs.

The main application is to assess the independence between two or more time point processes, homogeneous or nonhomogeneous.

It includes three families of independence tests, preliminary tools to process the data and describe the dependence between PP, and functions to generate dependent PPs.

Tests of independence between point processses

Null hypothesis : independence between N_x and N_y
Alternative hypothesis: existence of any type of random dependence

Three families are implemented. They use different statistics and require different assumptions: POISSON family, CLOSE family, CROSS family.

All together they cover a wide range of situations appearing in real problems.

Poisson family

CondTest.fun

Test to check the independence between two Poisson processes, Cebrián et al. (2019)

Under independence between N_x and N_y: given that a point tx_i has occurred in N_x, the distribution of N_y does not change.

$X_i \equiv$ number of points in N_y in an interval of length l_i around tx_i

$$X_i \sim Poisson(\mu_i) \text{ with } \mu_i = \int_{l_i} \lambda_y(t)dt$$

Two options are implemented to calculate the p-value: '*Poisson*' and '*Normal*'

CLOSE family

TestIndNH.fun and *TestIndLS.fun*

Tests based on the **close points distance** (Abaurrea et al. (2015)).

Testing independence is reduced to check if we have a uniform sample (related to distances between close points). Even under the null the uniform observations may be correlated.

A KS statistic is used but we need the empirical distribution of the statistic under the null, to calculate the p-value.

To generate samples of values of the statistic under the null, by generating samples of processes $N^* = (N_x, N_y^*)$: N_x is fixed and a new independent N_y^* is generated with the same distribution as N_y.

TestIndPB.fun : It can be applied to PPs where N_y follows a parametric models with a generation algorithm. It generates N_y^* the parametric model (implemented for Poisson processes and Neyman-Scott cluster processes).

TestIndLS.fun : It can be applied only to any homogeneous PPs (it does not requiere a parametric model). It generates N_y^* using a Lotwick-Silverman approach, **Lotwick and Silverman** (1982).

- the observation period of (N_x, N_y) is wrapped onto a circumference by identifying opposite sides,
- Fixing N_x, new N_y^* is generated by translating N_y a random (usually uniform) amount on the circumference. This breaks any dependence between them and keeps marginal distributions if it does not change over time (homogeneous processes).

CROSS family

NHK.fun and *NHJ.fun*

The CROSS family, Cebrián et al. 2019, can be applied to two PPs. They do not require any parametric assumption, only to known the marginal intensities $\lambda_x(t)$ and $\lambda_y(t)$.

The tests are based on statistics inspired by the cross K and J functions, Cronie and Van Lieshout (2016), which measure dependence betweeen spatial point processes, adapted to PPs in time.

The p-values are calculated using a LoS approach. In NH processes, the problem is that random translations change the marginal distribution of the processes. However, since the statistics are adjusted for the time-varying intensity, this approach is still valid provided that the intensity is translated with the process.

Other tools

Preprocessing tools

Peak Over Threshold approach (POT): an extreme is a run of consecutive observations over an extreme threshold u.

Tools to identify occurrence points using POT: *POTevents.fun* , *CPSPpoints.fun* , *PlotMargP.fun* , *PlotMCPSP.fun* ,...

Dependence measures between point proocesses

depchi.fun : estimates the extremal dependence functions $\chi(u)$ (and others) against thresholds u (for PP obtained from a POT approach)

CountingCor : calculates the correlation coefficient between the number of points in intervals of length l, in two PPs.

DutilleulPlot.fun : plots to check independence based on a Diggle's randomization test.

Generation of dependent point processes

Types of dependent point processes which represent common dependence structures in real problems

- Vectors of Marked Poisson processes with dependent marks from a Markov chain *DepMarkedNHPP.fun*
- Common Poisson shock processes *DepNHCPSP.fun*
- Neyman-Scott cluster processes *DepNHNeyScot.fun*
- Queues of point processes in a tandem *DepNHPPqueue.fun*

In all the cases, both homogeneous and non homogeneous can be generated. They are useful to implement inference tools based on simulation.

Application: occurrence of extreme heat events in three locations

Analysis of pairwise dependence between the occurrence of extreme heat events (EHE) in three Spanish locations: Barcelona (B), Zaragoza (Z) and Huesca (H).

EHE is defined as a run of consecutive days with a temperature over an extreme threshold (95^{th} percentile). The occurrence point of the event is the day of maximum temperature.

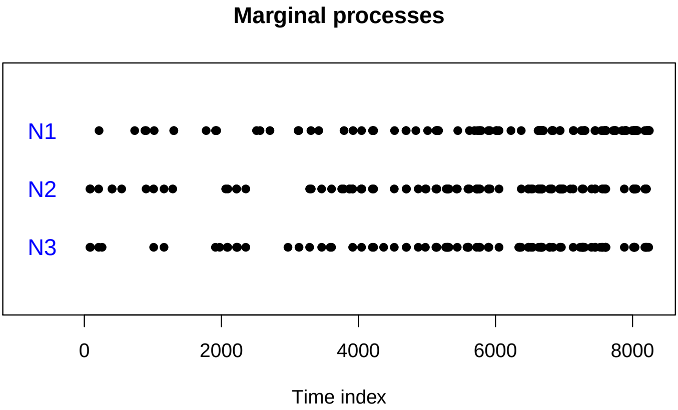
Data: daily maximum temperatures from May to September in 1951-2016
$\Rightarrow$ 3 series with 8262 observations

Identifying occurrence points of the EHEs: *posB*, *posH*, and *posZ*

```
load("F:/actual/JSSIndNHTest/JornadasR/AplicacionTxBHZ.RData")
library(NHPoisson)

T<-length(TxB)
auxB<-POTevents.fun(TxB, thres=31.3)

## Number of events:  121
## Number of excesses over threshold 31.3 : 205

auxH<-POTevents.fun(TxH, thres=36.4)

## Number of events:  106
## Number of excesses over threshold 36.4 : 188

auxZ<-POTevents.fun(TxZ, thres=37.8)

## Number of events:  104
## Number of excesses over threshold 37.8 : 176

posZ<-auxZ$Px
posB<-auxB$Px
posH<-auxH$Px
```

Plot of the occurence times of the EHE in the three locations

```
PlotMargP(list(posB,posH,posZ), T=T)
```



Dependence between extremes:

Extremal index: $\hat{\chi}_{BH}(0.95) \approx \hat{\chi}_{BZ}(0.95) \approx 0.4$. $TxH - TxZ$: $\hat{\chi}_{HZ}(0.95) \approx 0.7$.

```
depchi.fun(TxB,TxH,indgraph=FALSE,xlegend="bottomright", thresval=c(0,99)/100)
```

Correlation between number of EHEs in intervals of $ll = 10$ days.

```
 aux<-CountingCor(posB,posH, ll=10, T=T, method='kendall')

## Correlation:  0.33 . P-value:  0

 aux<-CountingCor(posB,posZ, ll=10, T=T, method='kendall')

## Correlation:  0.355 . P-value:  0

 aux<-CountingCor(posZ,posH, ll=10, T=T, method='kendall')

## Correlation:  0.676 . P-value:  0
```

There exists a pairwise dependence between all the locations and stronger between Zaragoza and Huesca.

Our aim: to check if this dependence can be explained by a set of covariates (harmonic terms and a covariate which represents the atmospheric situation.)

Step 1 . Fit of a marginal model to each series. Cebrián et al (2019): Poisson processes where the intensity is a function of the previous covariates are satisfactorily fitted to the three series.

Library NHPoisson: These models can be fitted using the function *fitPP.fun*, the covariates selected using the likelihood ratio test in *testlik.fun* and the models validated using *globalval.fun*.

Step 2 : to check if the Poisson processes are independent, given the covariates.

If the PP are independent it can be concluded that the dependence between the EHE processes is explained by the considered covariates, since once its effect is given, the processes are independent.

The three families of tests can be used, since the processes are Poisson processes.

Poisson family

Normal test is applied, since it has been proved empirically that it is more powerful than Poisson test. It requires N_y to be a Poisson process. Interval length $r = 15$, necessary to guarantee the Normal approximation of the statistic.

```
library(IndTestPP)
aux<-CondTest.fun(posZ, posB, lambda2=lambdaB, r=12)

aux$pvN

## [1] 0.66792
```

CLOSE family

The LoS can only be applied to homogeous PP, so that PaB is used. It requires a parametric marginal model, the fitted Poisson process in this case. It is implemented with 1000 runs.

```
library(parallel)
PBBZ<-TestIndNH(posZ,posB,nsim = 1000,type = "Poisson",lambdaMarg =cbind(lambdaB),
               cores=2,fixed.seed=65)
PBBZ$pv

##         D
## 0.3026973
```

Since it is a test based on simulation, the seed of generation can be fixed, and several cores can be used in the computation.

CROSS family

K test is used since it is in general more powerful than J test. It only requires to know λ_Y (or $\hat{\lambda}_Y$). It is implemented with 1000 translations and a r-grid with values from 1 to 10 (a short dependence is expected).

```
auxBZ<-NHK(lambdaZ, lambdaB, posC=posZ, posD=posB, r=c(1:20), typePlot='None',
          cores=2,fixed.seed=25)
auxBZ$pv

## [1] 0.2608392
```

Summary of the three pairwise comparisons

	B-H			B-Z			H-Z		
	Normal	PaB	*K*	Normal	PaB	*K*	Normal	PaB	*K*
pv	.30	.12	.40	.66	.30	.26	.002	.000	.002

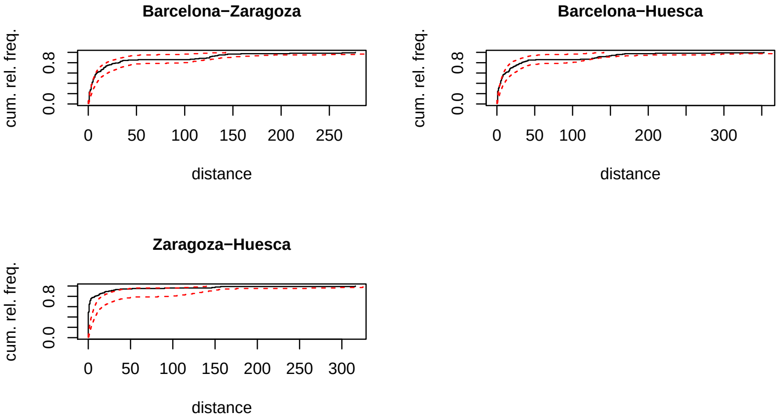
All the tests lead to the non rejection of independence between the occurrences in B-Z, and B-H and to the rejection between Z-H.

Conclusion: The considered covariates are able to explain the dependence the occurrence of the EHEs in Barcelona and Zaragoza and in Barcelona and Huesca but there still exists a dependence not explained by the covariates between Zaragoza and Huesca, which are the closest locations.

The best model for Barcelona is the previously fitted model, while the occurrence processes of Huesca and Zaragoza should be modelled taking into account the dependence between them.

The results of the tests are graphically confirmed by the Dutilleul plots given the fitted intensities,

```r
par(mfrow=c(2,2))
aux<-DutilleulPlot(posB, posZ, ModZ@lambdafit,cex.axis=1, cex.lab=1)
title("Barcelona-Zaragoza", cex.main=1)
```



Conclusions

IndTestPP is an R-package dealing providing tools to help in the modelling of vectors of point processes in time

- Three families of tests to check the independence between point processes
  - POISSON family: it can be applied to two H or NH Poisson processes
  - CLOSE family: It can be applied to two or more processes. LS to any homogeous processes and PaB to H and NH processes with a generation algorithm.
  - CROSS family: It can be applied to H or NH processes with a known intensity (without any parametric assumption).
- Some tools for data preprocessing: application of the POT approach, dependence measures, graphical tools, etc.
- Generation of point processes with different dependence structures

References

- Abaurrea, J. Asin, J. and Cebrian, A.C. (2015). A Bootstrap Test of Independence Between Three Temporal Nonhomogeneous Poisson Processes and its Application to Heat Wave Modeling. *Environmental and Ecological Statistics.*
- Cebrián A.C., Abaurrea, J. and Asin, J. (2019). Testing independence between two nonhomogeneous point processes in time. Submitted to Journal of simulation and computational statistics.
- Cronie, O. and van Lieshout, M.N.M. (2016). Summary statistics for inhomogeneous marked point processes. Ann Inst Stat Math.
- Dutilleul, P. (2011), *Spatio-temporal heterogeneity: Concepts and analyses*, Cambridge University Press.
- Lotwick, H.W. and Silverman, B.W. (1982). Methods for analysing Spatial processes of several types of points. *J.R. Statist. Soc. B*, 44(3), pp. 406-13