19-21 de Octubre 2022 | Granada

INTERNATIONAL CONFERENCE ON REGIONAL SCIENCE

Challenges, policies and governance of the territories in the post-covid era

Desafíos, políticas y gobernanza de los territorios en la era post-covid

XLVII REUNIÓN DE ESTUDIOS REGIONALES

XIV CONGRESO AACR

EXTENDED ABSTRACT

Title:

spqdep: An R package to test for spatial dependence in categorical spatial data

Authors and e-mail of all: López F.A. (Fernando.lopez@upct.es); Mínguez R. (Roman.Minguez@uclm.es); Páez A. (paezha@gmail.com); Ruiz Marín, M. (manuelruiz.spain@gmail.com)

Department: Facultad de CC de la Empresa

University: Universidad Politécnica de Cartagena

Subject area: *(please, indicate the subject area which corresponds to the paper)*

Abstract: *(minimum1500 words)*

Qualitative spatial variables are important in many fields of research. However, unlike the decades-worth of research devoted to the spatial association of quantitative variables, the exploratory analysis of spatial qualitative variables is relatively less developed. The objective of the present paper is to present a new R-package to test for spatial dependence in categorical spatial data. Several tests have been proposed, namely, the classical joint count statistics, the Q-test based on symbolic dynamics, the Scan-test based on scan methodology and a new spatial test based on spatial-runs. All tests can be applied to categorical spatial cross-section data with two or more categories and asymptotic and bootstrap permutation distribution are implemented. The R package is completely documented, including several examples and an user-guide is available as a vignette. The package spqdep is available in CRAN and is ideal for research and teaching activities.

Keywords: *qualitative data; spatial process; Q-test; local spatial tests*

JEL codes:

Introduction

This guide show the functionalities of the **spqdep** package to test spatial dependence on qualitative dataset.

Datasets

Two data sets will be used as examples in this guide:

- **provinces_spain**: The division of Spain into provinces. It is a multypolygon geometry with isolated provinces (islands without neighbouring provinces). See by example @paez2021.
- **FastFood.sf**: The data set used as example in @ruiz2010. It is a geometry of points.

The package is install like usual and the dataset can be loaded using the next code

```
library(spqdep)
data("provinces_spain", package = "spqdep")
data("FastFood.sf", package = "spqdep")
```

Data Generating Process (DGP)

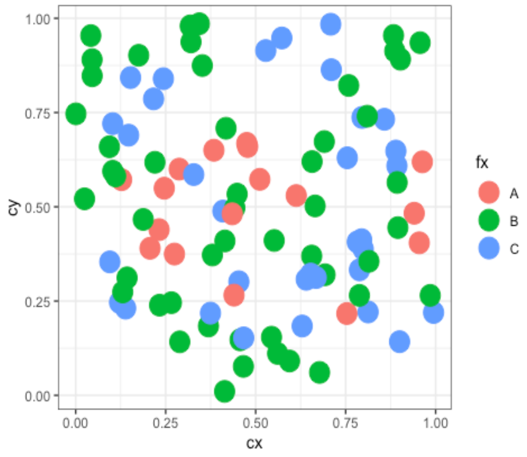
Additional to the two dataset available in the **spqdep** package. The user can generate structured spatial processes using the function. The DGP generate with this function defined in @ruiz2010.

The next code show how to generate a random process on a set of random points localized in a square 1x1. In this case, the connectivity criteria is based on the 4 near neighborhood.

```
set.seed(123)
N <- 100
cx <- runif(N)
cy <- runif(N)
coor <- cbind(cx,cy)
p <- c(1/6,3/6,2/6)
rho = 0.5
listw <- spdep::nb2listw(knn2nb(knearneigh(coor, k = 4)))
fx <- dgp.spq(list = listw, p = p, rho = rho)
```

The next plot show the qualitative spatial process defined.

```
ggplot(data.frame(fx = fx, cx = cx, cy = cy), aes(x = cx, y = cy, color = fx)) +
   geom_point(size = 6) +
   theme_bw()
```



Q-test

- The Q-test [@ruiz2010] is based on m-surroundings
- Before to apply the Q-test it is necessary define a set of the m-surroundings
- The function generate a set of m-surrounding.
- The user can tuning several parameters to obtain a congruent set of m-surroundings.

m-surroundings

- **m.surround()** is the function to generate m-surroundings.
- The output of this function is a object of the class **m_surr**
- Using the method the user can explore the coherence of m-surroundings.

By example. the next code obtain m-surroundings with length m = 3 and degree of overlapping r = 1:

```
m = 3
r = 1
mh <- m.surround(x = cbind(cx,cy), m = m, r = r)
class(mh)

## [1] "m_surr" "list"
```

Methods for the m_surr class

The **spqdep** have three methods that can be apply to this class: , and

- list the m-surroundings

```
print(mh)

##      [,1] [,2] [,3]
## [1,]    1   19   17
## [2,]    2   65   53
## [3,]    3   42   77
## [4,]    4   11   26
## [5,]    6   18   85
## [6,]    7   79   93
##  [7,]    8   21   31
##  [8,]   13   58   68
##  [9,]   15   98   74
## [10,]   17   76   96
## [11,]   20   87   24
## [12,]   22   82   92
## [13,]   25    9   61
## [14,]   26   14    7
## [15,]   27   72   13
## [16,]   31    5   20
## [17,]   37   73   89
## [18,]   38   30   15
## [19,]   41   91   62
## [20,]   46   47   49
```

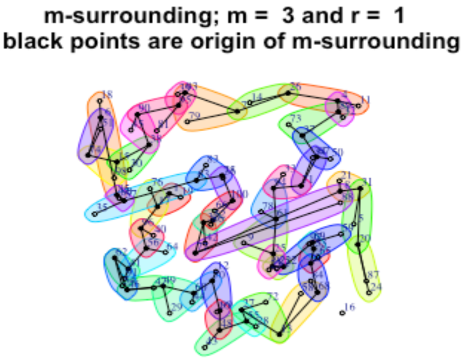
- generate a summary of some characteristics of the m-surroundings

```
summary(mh)
```

```
##
## Characteristics of m-surrounding:
##
## Number of m-surrounding (R): 49
## Length of m-surrounding (m): 3
## Number no-symbolized observations: 1
##
## List of no-symbolized observations:
## 16
##
## List of the degree overlaping:
##     There are 2 m-surrounding that have intersection with 1 m-surrounding
##     There are 47 m-surrounding that have intersection with 2 m-surrounding
## Mean degree of overlaping: 1.9592
```

- show the spatial structure of the m-surroundings

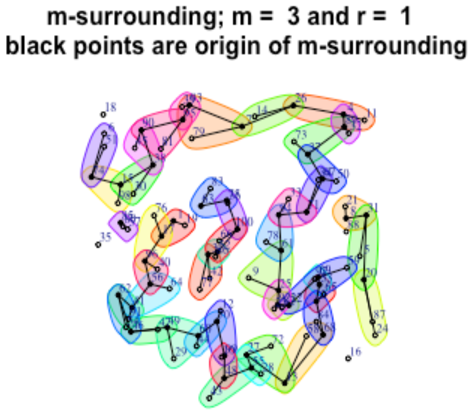
```
plot(mh, type = 1)
```



- With the argument **control** the user can tuning some characteristics of the m-surroundings.

By example, with **control** argument, the user can ‘prune’ non-coherent m-surroundings.

```
control <- list (dtmaxknn = 10)
mh.prune <- m.surround(x = coor, m = m, r = r, control = control)
plot(mh.prune)
```



The Q-test

- The function obtain the Q-test for a spatial process develop in @ruiz2010.

The user must select the longitude of the m-surroundings (m) and the overlapping degree (r). In the next code example, the Q-test is obtain for the DGP spatial process (fx) obtain with the . The coordinates **coor** must be included as argument.

```
q.test <- Q.test(fx = fx, coor = coor, m = 3, r = 1)
```

- The output is a list with the result for symbols based on permutations (standard) and combinations (equivalent).

- The output of this function is an object of the **spqtest** class.

Distribution of Q-test

- The asymptotic distribution is the default distribution to obtain the significance of Q-test [@ruiz2010].

- Alternatively, the Monte Carlo method can be used to obtain the significance of the test. The paper @lopez2012 describe this approach.

```
q.test.mc <- Q.test(fx = fx, coor = coor, m = 3, r = 1, distr = "mc")
summary(q.test.mc)
```

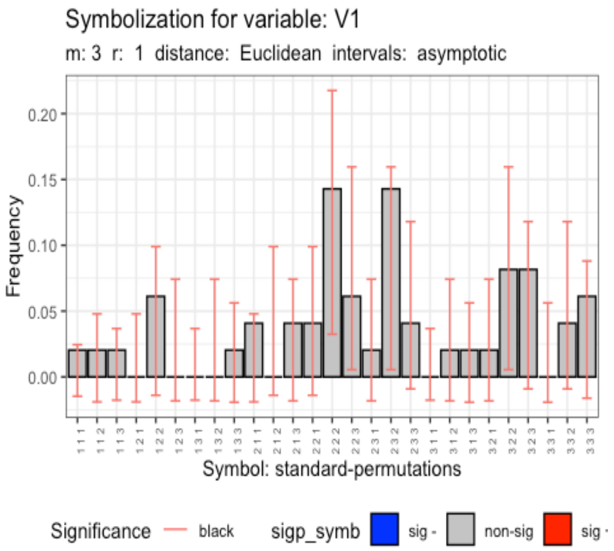
Methods for the spqtest class

A summary can be apply to an object of the spqtest class:

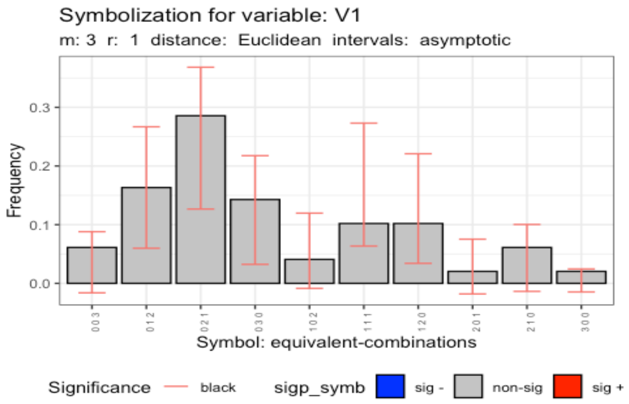
```
summary(q.test)
```

The histogram of the number of symbols is obtain appling the plot method.

```
plot(q.test)

## [[1]]
```



```
##
## [[2]]
```



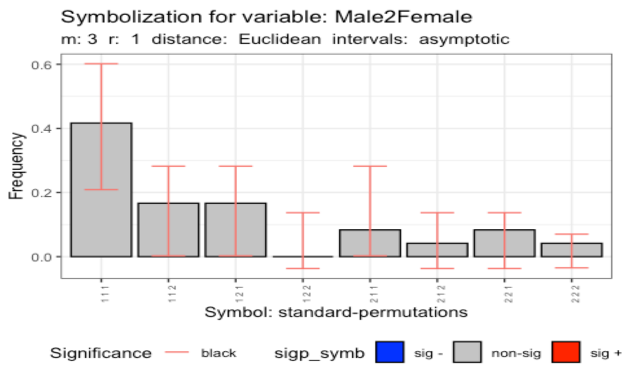
The Q-test using a sf object

- A sf object [@pebesma2018] or a data frame can be used as input of the function:

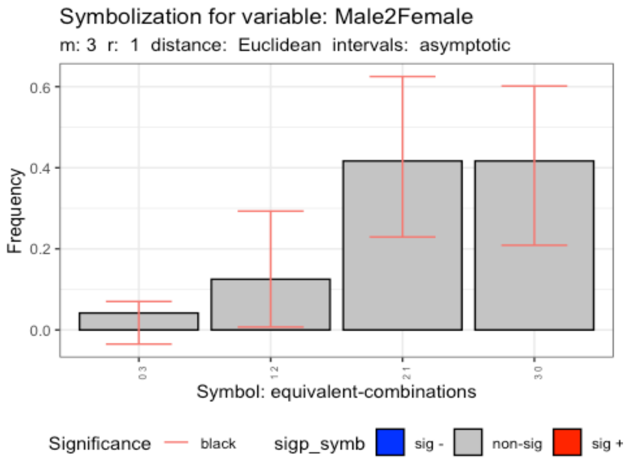
```
# Case 3: With a sf object with isolated areas
sf_use_s2(FALSE)

## Spherical geometry (s2) switched off

provinces_spain$Male2Female <- factor(provinces_spain$Male2Female > 100)
levels(provinces_spain$Male2Female) = c("men","woman")
f1 <- ~ Male2Female
q.test.sf <- Q.test(formula = f1, data = provinces_spain, m = 3, r = 1)
```

- The method show the histogram of the number of symbols

```
plot(q.test.sf)

## [[1]]
```



```
##
## [[2]]
```



Maps comparison. The QMap-test

- The function obtain the test for maps comparison publish in @Ruiz2012b

The QMap test

The next code generate two qualitative spatial process with different levels of spatial dependence and the Q-Map is apply.

```
p <- c(1/6,3/6,2/6)
rho = 0.5
QY1 <- dgp.spq(p = p, listw = listw, rho = rho)
rho = 0.8
QY2 <- dgp.spq(p = p, listw = listw, rho = rho)
dt = data.frame(QY1,QY2)
m = 3
r = 1
formula <- ~ QY1 + QY2
control <- list(dtmaxknn = 10)
qmap <- Q.map.test(formula = formula, data = dt, coor = coor, m = m, r = r, type
="combinations", control = control)
```

```
## Warning in Q.map.test(formula = formula, data = dt, coor = coor, m = m, : The
## ratio between the number of symbolized observations and the number of symbols is
## lower than 5.
```

- The output of id an object of the classes **qmap** and **htest**

Methods for qmap class

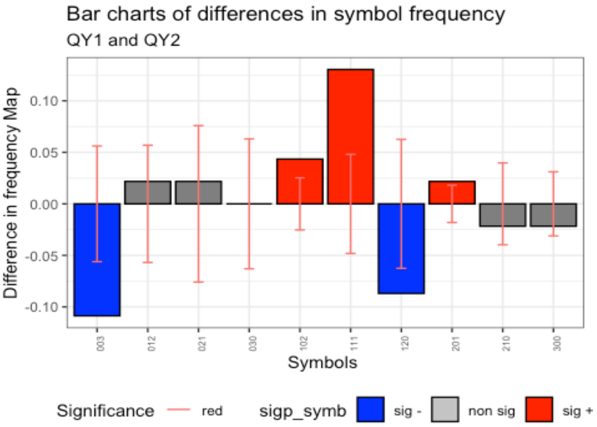
- The qmap object is a list with two elements. Each element is an object of the class **htext**

```
print(qmap[[1]])
```

```
##
##  Q-Map test of Equivalence for qualitative data.
##
##  Symbols type: combinations
##
##  Ratio Symbolized observations/Num symbols = 4.6
##
## data:  QY1 and QY2
## QE = 140.95, df = 9, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

- The method obtains the distribution of symbols with the confidence intervals specified by the user.

```
plot(qmap, ci=.6)
```



Runs tests

The runs test [@ruiz2021] have global and local versions

Global Runs test

- The function **sp.runs.test** obtain the spatial runs test.

```
listw <- knearneigh(coor, k = 3)
srq <- sp.runs.test(fx = fx, listw = listw)
```

- The output of this function is a object of the classes **sprunstest** and **htest**

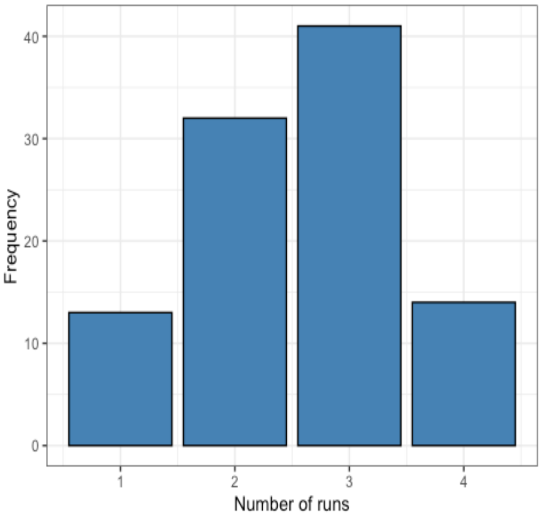
Methods for spruntest class

- The **spqdep** has two methods for this class y

```
print(srq)
```

```
##
##  Runs test of spatial dependence for qualitative data. Asymptotic
##  version
```

```
##
## data:  mxf
## sp.runs test = -2.5384, p-value = 0.01114
## alternative hypothesis: two.sided
## sample estimates:
##          Total runs     Mean total runs Variance total runs
##            256.0000            285.5152            135.1986

plot(srq)
```



The local Runs test

- The function **local.sp.runs.test** obtain the local test based on runs.

Asymptotic version

- Asymptotic version

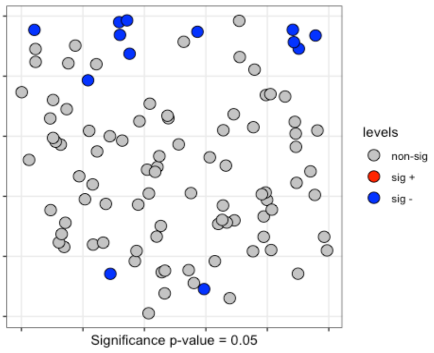
```
lsrq <- local.sp.runs.test(fx = fx, listw = listw, alternative = "less")
```

- The method list the statistic of each observation (point or region)

```
print(lsrq)

##     runs.i      E.i     Std.i    z.value    p.value
## 1        4 2.855152 0.8722689  1.3124950 0.90532341
## 2        4 2.855152 0.8722689  1.3124950 0.90532341
## 3        3 2.855152 0.8722689  0.1660594 0.56594492
## 4        1 2.855152 0.8722689 -2.1268116 0.01671787
## 5        4 2.855152 0.8722689  1.3124950 0.90532341
## 6        2 2.855152 0.8722689 -0.9803761 0.16345026
## 7        3 2.855152 0.8722689  0.1660594 0.56594492
## 8        3 2.855152 0.8722689  0.1660594 0.56594492
## 9        3 2.855152 0.8722689  0.1660594 0.56594492
## 10       3 2.855152 0.8722689  0.1660594 0.56594492
## 11       1 2.855152 0.8722689 -2.1268116 0.01671787
## 12       4 2.855152 0.8722689  1.3124950 0.90532341
## 13       2 2.855152 0.8722689 -0.9803761 0.16345026
## 14       1 2.855152 0.8722689 -2.1268116 0.01671787
## 15       2 2.855152 0.8722689 -0.9803761 0.16345026
## 16       2 2.855152 0.8722689 -0.9803761 0.16345026
## 17       3 2.855152 0.8722689  0.1660594 0.56594492
## 18      1 2.855152 0.8722689 -2.1268116 0.01671787
## 19      2 2.855152 0.8722689 -0.9803761 0.16345026
## 20      4 2.855152 0.8722689  1.3124950 0.90532341
## 21      3 2.855152 0.8722689  0.1660594 0.56594492
## 22      2 2.855152 0.8722689 -0.9803761 0.16345026
## 23      2 2.855152 0.8722689 -0.9803761 0.16345026
## 24      3 2.855152 0.8722689  0.1660594 0.56594492
## 25      3 2.855152 0.8722689  0.1660594 0.56594492
## 26      2 2.855152 0.8722689 -0.9803761 0.16345026
## 27      3 2.855152 0.8722689  0.1660594 0.56594492
## 28      1 2.855152 0.8722689 -2.1268116 0.01671787
## 29      1 2.855152 0.8722689 -2.1268116 0.01671787
## 30      2 2.855152 0.8722689 -0.9803761 0.16345026
## 31      3 2.855152 0.8722689  0.1660594 0.56594492
## 32      1 2.855152 0.8722689 -2.1268116 0.01671787
## 33      2 2.855152 0.8722689 -0.9803761 0.16345026
## 34      4 2.855152 0.8722689  1.3124950 0.90532341
## 35      2 2.855152 0.8722689 -0.9803761 0.16345026
```

- The method identify the localization with values of local test significant.

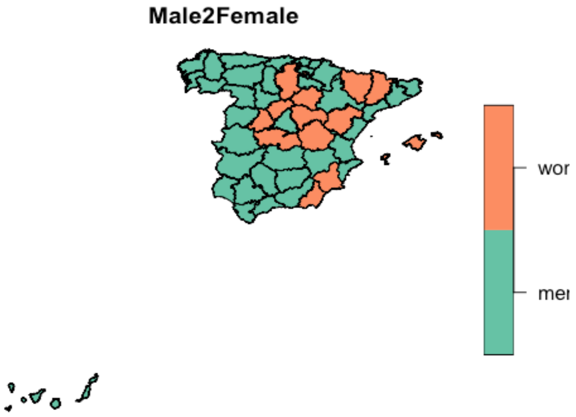
```
plot(lsrq, sig = 0.05)
```



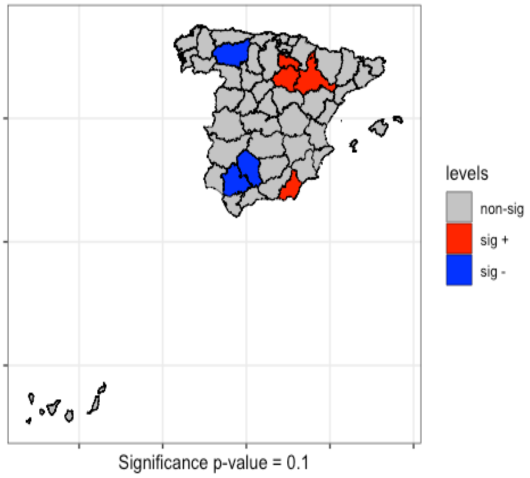
Monte Carlo local runs test

- The Monte Carlo distribution ot the local test using a sf object

```
data("provinces_spain")
listw <- spdep::poly2nb(as(provinces_spain,"Spatial"), queen = FALSE)

## although coordinates are longitude/latitude, st_intersects assumes that they are planar

provinces_spain$Male2Female <- factor(provinces_spain$Male2Female > 100)
levels(provinces_spain$Male2Female) = c("men","woman")
plot(provinces_spain["Male2Female"])
```



```
formula <- ~ Male2Female
# Boots Version
lsrq <- local.sp.runs.test(formula = formula, data = provinces_spain, listw = listw, distr
="bootstrap", nsim = 199)
plot(lsrq, sf = provinces_spain, sig = 0.10)
```



The scan test

- Two of the scan tests to identify clusters can be apply to test spatial structure in qualitative spatial processes.
- The scan test don't need pre-define the classical W conectivity matrix.
- See @Kanaroglou2016
- The scan tests contrasts the null of independence of a spatial qualitative process and give additional information indicating one (or perhaps more) spatial cluster(s).
- The scan tests don't have asymptotic distribution. The significance is obtained by permutational resampling.
- The output of the scan function is an object of the classes **scantest** and **htest**

Scan bernoulli

- For qualitative spatial process with two categories the bernoulli scan test is obtain with the next code

```
formula <- ~ Male2Female
scan.spain <- spqdep::scan.test(formula = formula, data = provinces_spain, case="men",
nsim = 99, distr = "bernoulli")
print(scan.spain)

##
##  Scan test. Distribution: bernoulli
##
## data:  Male2Female
## scan-loglik = 6.0359, p-value = 0.07
## alternative hypothesis: High
## sample estimates:
##
## Total observations in the MLC =   17.00
## Expected cases in the MLC =       11.84
## Observed cases in the MLC =       16.00
```

scan multinomial

- In case of a spatial process with three or more categories

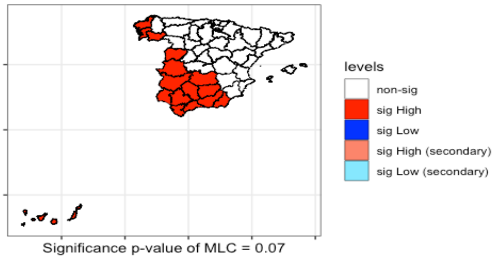
```
data(FastFood.sf)
formula <- ~ Type
scan.fastfood <- scan.test(formula = formula, data = FastFood.sf, nsim = 99, distr =
"multinomial", windows = "elliptic",
                         nv = 50)
print(scan.fastfood)

##
##  Scan test. Distribution: multinomial
##
## data:  Type
## scan-loglik = 15.506, p-value < 2.2e-16
## sample estimates:
##                  H     P     S Sum
## cases.expect 13.48 14.86 14.66  43
## cases.observ 16.00  1.00 26.00  43
```

Methods for scan test

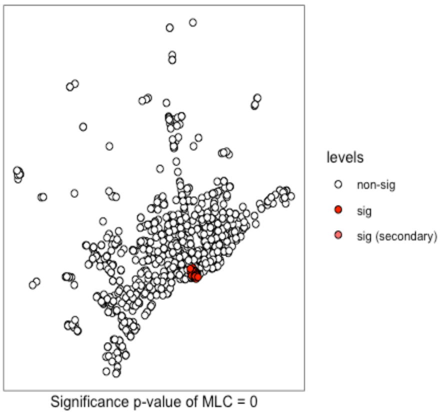
- Two method can be used with **scantest** objects: and

```
summary(scan.fastfood)

##
## Summary of data:
## Distribution....................: multinomial
## Number of locations.............: 877
## Total number of cases...........: 877
## Names of cathegories...........: H P S
## Total cases per category........: 275 303 299
## Percent cases per category......: 0.31 0.35 0.34
##
## Scan statistic:
## Total cases in the MLC.........: 43
## Names of cathegories...........: H P S
## Observed cases in the MLC......: 13.48 14.86 14.66
## Expected cases in the MLC......: 16 1 26
## Value of statistic (loglik ratio)....: 15.5058
## p-value........................: 0
##
## IDs of cluster detect:
## Location IDs included.....:  68 849 152 499 630 763 827 765 617 600 607 48 58 588
743 843 74 122 750 115 645 61 226 796 876 699 610 597 596 721 751 53 186 659 778
63 106 229 585 738 612 131 208
##
##
## Secondary Scan statistic. Number 1
## Total cases in secondary cluster......:  16
## Names of cathegories.................: H P S
## Percent per category total...........: 0.31 0.35 0.34
## Percent per category inside cluster..: 0.25 0.5 0.25
## Value of statisitic (loglik ratio)....: 11.5285
## p-value.........................: 0.1
## Location IDs included..................:  677 311 781 128 108 436 551 576 21 374 319
717 561 6 629 547
##
##
## Secondary Scan statistic. Number 2
## Total cases in secondary cluster......:  7
## Names of cathegories.................: H P S
## Percent per category total...........: 0.31 0.35 0.34
## Percent per category inside cluster..: 0.14 0.29 0.57
## Value of statisitic (loglik ratio)....: 7.0038
## p-value.........................: 0.96
## Location IDs included..................:  158 709 335 801 749 545 856
##
##
## Secondary Scan statistic. Number 3
## Total cases in secondary cluster......:  17
## Names of cathegories.................: H P S
## Percent per category total...........: 0.31 0.35 0.34
## Percent per category inside cluster..: 0.35 0.35 0.29
## Value of statisitic (loglik ratio)....: 6.8747
## p-value.........................: 0.98
## Location IDs included..................:  521 782 162 643 297 220 267 265 104 530 312
523 783 157 531 848 680
##
##
## Secondary Scan statistic. Number 4
## Total cases in secondary cluster......:  17
## Names of cathegories.................: H P S
## Percent per category total...........: 0.31 0.35 0.34
## Percent per category inside cluster..: 0.47 0.29 0.24
## Value of statisitic (loglik ratio)....: 6.7961
## p-value.........................: 0.98
## Location IDs included.................:  190 837 555 711 646 216 17 390 742 563 307 4
353 197 254 192 66
##
```

```
##
## Secondary Scan statistic. Number 5
## Total cases in secondary cluster......:  10
## Names of cathegories.................: H P S
## Percent per category total...........: 0.31 0.35 0.34
## Percent per category inside cluster..: 0.4 0.2 0.4
## Value of statisitic (loglik ratio)....: 6.5951
## p-value.........................: 0.98
## Location IDs included..................:  78 365 668 228 170 857 306 708 651 187

plot(scan.spain, sf = provinces_spain)
```



```
plot(scan.fastfood, sf = FastFood.sf)
```



Similarity test

The @Farber2014 paper develop the similarity test

Simiarity test

The function calculates the similarity test for both asymptotic distribution and permutational resampling.

```
coor <- st_coordinates(st_centroid(FastFood.sf))
listw <- spdep::knearneigh(coor, k = 4)
formula <- ~ Type
similarity <- similarity.test(formula = formula, data = FastFood.sf, listw = listw)
print(similarity)
##
##  Similarity test of spatial dependence for qualitative data.
##  Distribution: asymptotic
```

```
##
## data:  Type
## Similarity-test = -5.4476, p-value = 5.105e-08
## alternative hypothesis: two.sided
```

join-count tests

- The functions of the **spdep** R-package have been **wrapped** for Bernoulli and Multinomial distributions. Asymptotic or Monte Carlo distributions (permutations) can be used to evaluate the signification of the tests.

Asyntotic distribution

```
provinces_spain$Older <- cut(provinces_spain$Older, breaks = c(-Inf,19,22.5,Inf))
levels(provinces_spain$Older) = c("low","middle","high")
f1 <- ~ Older + Male2Female
jc1 <- jc.test(formula = f1, data = provinces_spain, distr = "asymptotic", alternative =
"greater", zero.policy = TRUE)

## although coordinates are longitude/latitude, st_intersects assumes that they are planar

summary(jc1)
```

Monte Carlo distribution

```
jc1 <- jc.test(formula = f1, data = provinces_spain, distr = "mc", alternative = "greater",
zero.policy = TRUE)

## although coordinates are longitude/latitude, st_intersects assumes that they are planar

summary(jc1)
```

8 References

Farber, S., Marin, M. R., & Páez, A. (2015). Testing for spatial independence using similarity relations. Geographical Analysis, 47(2), 97-120.

Paez, A., Lopez, F. A., Menezes, T., Cavalcanti, R., & Pitta, M. G. D. R. (2021). A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain. Geographical analysis, 53(3), 397-421.

Kanaroglou, Pavlos. 2016. *Spatial Analysis in Health Geography*. Routledge. https://doi.org/10.4324/9781315610252.

López, Fernando A, and Antonio Páez. 2012. “Distribution-Free Inference for q (m) Based on Permutational Bootstrapping: An Application to the Spatial Co-Location Pattern of Firms in Madrid.” *Estadística Española* 54 (177): 135–56.

Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Ruiz, Manuel, Fernando López, and Antonio Páez. 2010. “Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics.” *Journal of Geographical Systems* 12 (3): 281–309. https://doi.org/10.1007/s10109-009-0100-1.

———. 2012. “Comparison of Thematic Maps Using Symbolic Entropy.” *International Journal of Geographical Information Science* 26 (3): 413–39. https://doi.org/10.1080/13658816.2011.586327.

———. 2021. “A Test for Global and Local Homogeneity of Categorical Data Based on Spatial Runs.” *Working Paper*.