Package ‘WaveSampling’

January 30, 2020

Type Package

Title Weakly Associated Vectors (WAVE) Sampling

Version 0.1.1

Description Spatial data are generally auto-correlated, meaning that if two units selected are close to each other, then it is likely that they share the same properties. For this reason, when sampling in the population it is often needed that the sample is well spread over space. A new method to draw a sample from a population with spatial coordinates is proposed. This method is called wave (Weakly Associated Vectors) sampling. It uses the less correlated vector to a spatial weights matrix to update the inclusion probabilities vector into a sample. For more details see Raphaël Jauslin and Yves Tillé (2019) <arXiv:1910.13152>.

URL https://github.com/RJauslin/WaveSampling

BugReports https://github.com/RJauslin/WaveSampling/issues

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

LinkingTo RcppArmadillo, Rcpp

Imports Rcpp

Depends Matrix, R (>= 2.10)

Suggests knitr, rmarkdown, ggplot2, ggvoronoi, sampling, BalancedSampling, sp, sf, stats

VignetteBuilder knitr

NeedsCompilation yes

Author Raphaël Jauslin [aut, cre] (<https://orcid.org/0000-0003-1088-3356>), Yves Tillé [aut] (<https://orcid.org/0000-0003-0904-5523>)

Maintainer Raphaël Jauslin <raphael.jauslin@unine.ch>

Repository CRAN

Date/Publication 2020-01-30 12:00:09 UTC
R topics documented:

distUnitk .................................................. 2
IB .............................................................. 3
sb_vk .......................................................... 5
varHAJ ......................................................... 6
wave ............................................................ 7
wpik ........................................................... 9
wpikInv ....................................................... 11

Index 13

distUnitk

Squared Euclidean distances of the unit k.

Description

Calculate the squared Euclidean distance from unit k to the other units.

Usage

distUnitk(X, k, tore, toreBound)

Arguments

X matrix representing the spatial coordinates.
k the unit index to be used.
tore an optional logical value, if we are considering the distance on a tore. See Details.
toreBound an optional numeric value that specify the length of the tore.

Details

Let \( x_k, x_l \) be the spatial coordinates of the unit \( k, l \in U \). The classical euclidean distance is given by

\[
d^2(k, l) = (x_k - x_l)^\top(x_k - x_l).
\]

When the points are distributed on a \( N_1 \times N_2 \) regular grid of \( \mathbb{R}^2 \). It is possible to consider the units like they were placed on a tore. It can be illustrated by Pac-Man passing through the wall to get away from ghosts. Specifically, we could consider two units on the same column (resp. row) that are on the opposite have a small distance,

\[
d^2_T(k, l) = \min((x_{k_1} - x_{l_1})^2, (x_{k_1} + N_1 - x_{l_1})^2, (x_{k_1} - N_1 - x_{l_1})^2) + \min((x_{k_2} - x_{l_2})^2, (x_{k_2} + N_2 - x_{l_2})^2, (x_{k_2} - N_2 - x_{l_2})^2).
\]

The option toreBound specify the length of the tore in the case of \( N_1 = N_2 = N \). It is omitted if the tore option is equal to FALSE.
**Value**

A vector of length $N$ that contains the distances from the unit $k$ to all other units.

**Author(s)**

Raphaël Jauslin <raphael.jauslin@unine.ch>

**See Also**

`wpik`, `wave` and `dist`.

**Examples**

```r
N <- 5
x <- seq(1,N,1)
X <- as.matrix(expand.grid(x,x))
distUnitk(X,k = 2,tore = TRUE,toreBound = 5)
distUnitk(X,k = 2,tore = FALSE,toreBound = -1)
```

---

**Description**

This function implements the spreading measure based on Moran’s $I$ index.

**Usage**

```r
IB(W, s)
```

**Arguments**

- `W` : a stratification matrix inheriting from `sparseMatrix` that represents the spatial weights. See `wpik`.
- `s` : a vector of size $N$ with elements equal to 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for non-chosen units.

**Details**

This index is developed by Tillé et al. (2018) and measures the spreading of a sample drawn from a population. It uses a corrected version of the traditional Moran’s $I$ index. Each row of the matrix $W$ should represent a stratum. Each stratum is defined by a particular unit and its neighbouring units. See `wpik`. The spatial balance measure is equal to

$$I_B = \frac{(s - \bar{s}_w) \T W (s - \bar{s}_w)}{\sqrt{(s - \bar{s}_w) \T D (s - \bar{s}_w)(s - \bar{s}_w) \T B (s - \bar{s}_w)}}.$$ 

where $D$ is the diagonal matrix containing the $w_i$. 

**IB**

*Spreading measure based on Moran’s $I$ index*
\[ \bar{s}_w = \frac{1}{1^T W 1} s^T W 1 \]

and

\[ B = W^T D^{-1} W - \frac{W^T 1 1^T W}{1^T W 1}. \]

To specify the spatial weights uses the argument \( W \).

Value

A numeric value that represents the spatial balance. It could be any real value between -1 (spread) and 1 (clustered).

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References


See Also

wpik

Examples

```r
N <- 36
n <- 12
x <- seq(1,sqrt(N),1)
X <- expand.grid(x,x)
pik <- rep(n/N,N)
W <- wpik(as.matrix(X),pik,bound = 1,tore = TRUE,shift = FALSE,toreBound = sqrt(N))
W <- W - diag(diag(W))
s <- wave(as.matrix(X),pik,tore = TRUE,shift = TRUE,comment = TRUE)
IB(W,s)
```
Description

Calculates the $v_k$ values of the spatial balance developed by Stevens and Olsen (2004) and suggested by Grafström et al. (2012).

Usage

```
sb_vk(pik, X, s)
```

Arguments

- `pik`: vector of the inclusion probabilities. The length should be equal to $N$.
- `X`: matrix representing the spatial coordinates.
- `s`: A vector of size $N$ with elements equal 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for non-chosen unit.

Details

The spatial balance measure based on the Voronoi polygons is defined by

$$B(S) = \frac{1}{n} \sum_{k \in S} (v_k - 1)^2.$$  

The function returns the $v_k$ values and is mainly based on the function `sb` of the package `BalancedSampling` Grafström and Lisic (2019).

Value

A vector of size $N$ with elements equal to the $v_k$ values. If the unit is not selected then the value is equal to 0.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References


See Also

BalancedSampling::sb

Examples

N <- 50
n <- 10
X <- as.matrix(cbind(runif(N),runif(N)))
pik <- sampling::inclusionprobabilities(runif(N),n)
s <- wave(X, pik)
v <- sb_vk(pik,X,s)
1/n*sum((v[which(v != 0)]-1)^2)
BalancedSampling::sb(pik,X,which(s == 1))

---

varHAJ

Hajek-Rosen variance estimator

Description

Estimator of the variance of the Horvitz-Thompson estimator. It is based on the variance estimator of the conditional Poisson sampling design. See Tillé (2020, Chapter 5) for more informations.

Usage

varHAJ(y, pik, s)

Arguments

y vector of size n that represent the variable of interest.
pik vector of the inclusion probabilities. The length should be equal to n.
s index vector of size n with elements equal to the selected units.

Details

The function computes the following quantity :

\[ v_{HAJ}(\hat{Y}_{HT}) = \frac{n}{n-1} \sum_{k \in S} (1 - \pi_k) \left( \frac{y_k}{\pi_k} - \frac{\sum_{l \in S} (1 - \pi_l) / \pi_l}{\sum_{l \in S} (1 - \pi_l)} \right)^2. \]

This estimator is well-defined for maximum entropy sampling design and use only inclusion probabilities of order one.

Value

A number, the variance
References


---

**Weakly associated vectors sampling**

**Description**

Select a spread sample from inclusion probabilities using the weakly associated vectors sampling method.

**Usage**

```r
wave(
  X,
  pik,
  bound = 1,
  tore = FALSE,
  shift = FALSE,
  toreBound = -1,
  comment = FALSE,
  fixedSize = TRUE
)
```

**Arguments**

- **X**
  - matrix representing the spatial coordinates.

- **pik**
  - vector of the inclusion probabilities. The length should be equal to N.

- **bound**
  - a scalar representing the bound to reach. See Details. Default is 1.

- **tore**
  - an optional logical value, if we are considering the distance on a tore. See Details. Default is TRUE.

- **shift**
  - an optional logical value, if you would use a shift perturbation. See Details. Default is FALSE.

- **toreBound**
  - a numeric value that specify the size of the grid. Default is -1.

- **comment**
  - an optional logical value, indicating some informations during the execution. Default is FALSE.

- **fixedSize**
  - an optional logical value, if you would impose a fixed sample size. Default is TRUE.
Details

The main idea is derived from the cube method (Devill and Tillé, 2004). At each step, the inclusion probabilities vector \( p_i \) is randomly modified. This modification is carried out in a direction that best preserves the spreading of the sample.

A stratification matrix \( A \) is computed from the spatial weights matrix calculated from the function \( wpik \). Depending if \( A \) is full rank or not, the vector giving the direction is not selected in the same way.

If matrix \( A \) is not full rank, a vector that is contained in the right null space is selected:

\[
Null(A) = \{x \in \mathbb{R}^N | Ax = 0 \}.
\]

If matrix \( A \) is full rank, we find \( v, u \) the singular vectors associated to the smallest singular value \( \sigma \) of \( A \) such that

\[
Av = \sigma u, \quad A^T u = \sigma v.
\]

Vector \( v \) is then centered to ensure fixed sample size. At each step, inclusion probabilities is modified and at least one component is set to 0 or 1. Matrix \( A \) is updated from the new inclusion probabilities. The whole procedure it repeated until it remains only one component that is not equal to 0 or 1.

For more informations on the options \( \text{tore} \) and \( \text{toreBound} \), see \text{distUnitk}. If \( \text{tore} \) is set up \( \text{TRUE} \) and \( \text{toreBound} \) not specified the \( \text{toreBound} \) is equal to \( N^{1/p} \)

where \( p \) is equal to the number of column of the matrix \( X \).

For more informations on the option \( \text{shift} \), see \( wpik \).

If \( \text{fixedSize} \) is equal \( \text{TRUE} \), the weakest associated vector is centered at each step of the algorithm. This ensures that the size of the selected sample is equal to the sum of the inclusion probabilities.

Value

A vector of size \( N \) with elements equal 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for non-chosen unit.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References


See Also

\( wpik \), \text{distUnitk}. 
Examples

#---------
# Example 2D
#---------
N <- 50
n <- 15
X <- as.matrix(cbind(runif(N),runif(N)))
pik <- sampling::inclusionprobabilities(runif(N),n)
s <- wave(X,pik)

#---------
# Example 2D grid
#---------
N <- 36 # 6 x 6 grid
n <- 12 # number of unit selected
x <- seq(1,sqrt(N),1)
X <- as.matrix(cbind(rep(x,times = sqrt(N)),rep(x,each = sqrt(N))))
pik <- rep(n/N,N)
s <- wave(X,pik, tore = TRUE,shift = FALSE)

#---------
# Example 1D
#---------
N <- 100
n <- 10
X <- as.matrix(seq(1,N,1))
pik <- rep(n/N,N)
s <- wave(X,pik,tore = TRUE,shift =FALSE,comment = TRUE)

------------------------------------------------------------------------

wpik

Stratification matrix from inclusion probabilities

Description

The stratification matrix is calculated from the inclusion probabilities. It takes the distances between units into account. See Details.

Usage

wpik(X, pik, bound = 1, tore = FALSE, shift = FALSE, toreBound = -1)
Arguments

- **X**: matrix representing the spatial coordinates.
- **pik**: vector of the inclusion probabilities. The length should be equal to \( N \).
- **bound**: a scalar representing the bound to reach. Default is 1.
- **tore**: an optional logical value, if we are considering the distance on a tore. Default is FALSE.
- **shift**: an optional logical value, if you would use a shift perturbation. See Details for more informations. Default is FALSE.
- **toreBound**: a numeric value that specify the size of the grid. Default is -1.

Details

Entries of the stratification matrix indicates how the units are close from each others. Hence a large value \( w_{kl} \) means that the unit \( k \) is close to the unit \( l \). This function considers that a unit represents its neighbor till their inclusion probabilities sum up to bound.

We define \( G_k \) the set of the nearest neighbor of the unit \( k \) including \( k \) such that the sum of their inclusion probabilities is just greater than bound. Moreover, let \( g_k = \#G_k \), the number of elements in \( G_k \). The matrix \( W \) is then defined as follows,

- \( w_{kl} = \pi_l \) if unit \( l \) is in the set of the \( g_k - 1 \) nearest neighbor of \( k \).
- \( w_{kl} = \pi_l + 1 - (\sum_{t \in G_k} \pi_t) \) if unit \( l \) is the \( g_k \) nearest neighbour of \( k \).
- \( w_{kl} = 0 \) otherwise.

Hence, the \( k \)th row of the matrix represents neighborhood or stratum of the unit such that the inclusion probabilities sum up to 1 and the \( k \)th column the weights that unit \( k \) takes for each stratum.

The option **shift** add a small normally distributed perturbation \( \text{rnorm}(0, 0.01) \) to the coordinates of the centroid of the stratum considered. This could be useful if there are many unit that have the same distances. Indeed, if two units have the same distance and are the last unit before that the bound is reached, then the weights of the both units is updated. If a shift perturbation is used then all the distances are differents and only one unit weight is update such that the bound is reached.

The shift perturbation is generated at the beginning of the procedure such that each stratum is shifted by the same perturbation.

Value

A sparse matrix representing the spatial weights.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

See Also

wpikInv, distUnitk, wave.
Examples

```r
N <- 25
n <- 5
X <- as.matrix(cbind(runif(N), runif(N)))
pik <- sampling::inclusionprobabilities(runif(N), n)
W <- wpik(X, pik)
```

---

**wpikInv**  
*Stratification matrix from inverse inclusion probabilities*

**Description**

The stratification matrix is calculated from the inverse inclusion probabilities. It is a direct implementation of the spatial weights specified in Tillé et al., (2018).

**Usage**

```r
wpikInv(X, pik, tore = FALSE, shift = FALSE, toreBound = -1)
```

**Arguments**

- **X**  
  matrix representing the spatial coordinates.
- **pik**  
  vector of the inclusion probabilities. The length should be equal to N.
- **tore**  
  an optional logical value, if we are considering the distance on a tore. Default is FALSE.
- **shift**  
  an optional logical value, if you would use a shift perturbation. See Details for more informations. Default is FALSE.
- **toreBound**  
  a numeric value that specify the size of the grid. Default is -1.

**Details**

Entries of the stratification matrix indicates how the units are close from each others. Hence a large value \( w_{kl} \) means that the unit \( k \) is close to the unit \( l \). This function considers that if unit \( k \) were selected in the sample drawn from the population then \( k \) would represent \( 1/\pi_k \) units in the population and, as a consequence, it would be natural to consider that \( k \) has \( n_k = (1/\pi_k - 1) \) neighbours in the population. The \( n_k \) neighbours are the nearest neighbours of \( k \) according to distances. The weights are so calculated as follows:

- \( w_{kl} = 1 \) if unit \( l \in N_{[n_k]} \)
- \( w_{kl} = n_k - [n_k] \) if unit \( l \) is the \( [n_k] \) nearest neighbour of \( k \).
- \( w_{kl} = 0 \) otherwise.
\[ \lfloor nk \rfloor \text{ and } \lceil nk \rceil \text{ are the inferior and the superior integers of } nk. \]

The option shift add a small normally distributed perturbation \( \text{rnorm}(0, 0.01) \) to the coordinates of the centroid of the stratum considered. This could be useful if there are many unit that have the same distances. Indeed, if two units have the same distance and are the last unit before that the bound is reached, then the weights of the both units is updated. If a shift perturbation is used then all the distances are differents and only one unit weight is update such that the bound is reached.

The shift perturbation is generated at the beginning of the procedure such that each stratum is shifted by the same perturbation.

Value

A sparse matrix representing the spatial weights.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References


See Also

wpik, distUnitk, wave.

Examples

```r
N <- 25
n <- 5
X <- as.matrix(cbind(runif(N), runif(N)))
pik <- sampling::inclusionprobabilities(runif(N), n)
W <- wpikInv(X, pik)
```
Index

BalancedSampling::sb, 6

dist, 3
distUnitk, 2, 8, 10, 12

IB, 3

sb, 5
sb_vk, 5
sparseMatrix, 3

varHAJ, 6

wave, 3, 7, 10, 12
wpik, 3, 4, 8, 9, 12
wpikInv, 10, 11