A Mann-Whitney spatial scan statistic for continuous data

Lionel Cucala, Christophe Dematteï

August 24th 2010

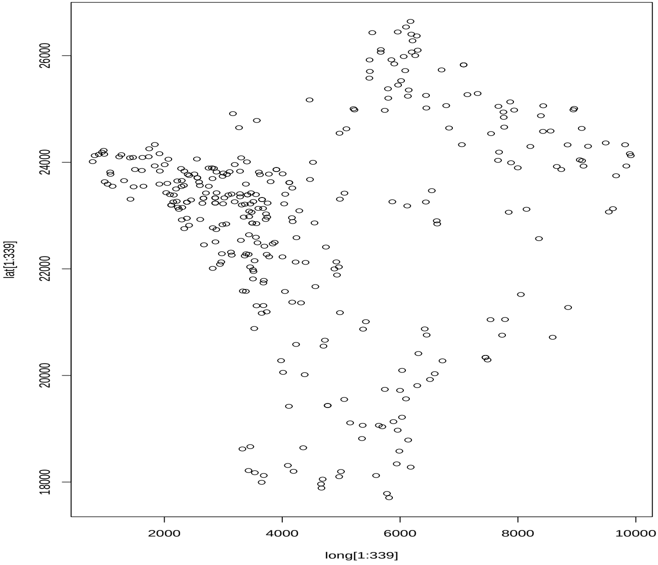
UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Outline

Outline

A data set to analyze

A data set to analyze



A data set to analyze

A data set to analyze

- 339 dairy farms located in France.

A data set to analyze

- 339 dairy farms located in France.
- Somatic individual score (indicator for a disease called mastitis).

A data set to analyze

- 339 dairy farms located in France.

- Somatic individual score (indicator for a disease called mastitis).

- Marked point process : $(X_i, C_i), i = 1, \cdots, n$ where $X_i \in A \subset \mathbb{R}^d$ and $C_i \in \mathbb{R}$.

Definition :

Cluster= geographical area where the continuous variable is higher than outside.

Definition :

Cluster= geographical area where the continuous variable is higher than outside.

Question :

Are there one or more clusters ? Where ?

The null hypothesis

The null hypothesis

- We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

The null hypothesis

- We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

- Goals : detecting cluster(s) **and** testing the significance according to H_0.

The null hypothesis

- We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

- Goals : detecting cluster(s) **and** testing the significance according to H_0.

  Two steps :

The null hypothesis

➡ We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

➡ Goals : detecting cluster(s) **and** testing the significance according to H_0.

Two steps :

- defining the set of potential clusters.

The null hypothesis

➡ We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

➡ Goals : detecting cluster(s) **and** testing the significance according to H_0.

Two steps :

- defining the set of potential clusters.
- choosing a concentration index.

The null hypothesis

➡ We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

➡ Goals : detecting cluster(s) **and** testing the significance according to H_0.

Two steps :

- defining the set of potential clusters.
- choosing a concentration index.

Statistic : maximal concentration among potential clusters.

The null hypothesis

- We attempt to reject $H_0 : (C_1, \cdots, C_N)$ independent and identically distributed.

- Goals : detecting cluster(s) **and** testing the significance according to H_0.

Two steps :

- defining the set of potential clusters.
- choosing a concentration index.

Statistic : maximal concentration among potential clusters.

Significance estimated by a Monte-Carlo procedure.

Outline

Outline

Fixed-shape potential clusters

Fixed-shape potential clusters

➡ Circles centered on one event, another event on the circumference.

Fixed-shape potential clusters

- Circles centered on one event, another event on the circumference.
- Elliptic clusters, with given orientation and shape (2D only).

Data-based potential clusters

Data-based potential clusters

Graphs $\mathcal{G}(\delta)$ associated to the point process :

Data-based potential clusters

Graphs $\mathcal{G}(\delta)$ associated to the point process :

- Vertices : $\{1, \cdots, n\}$.

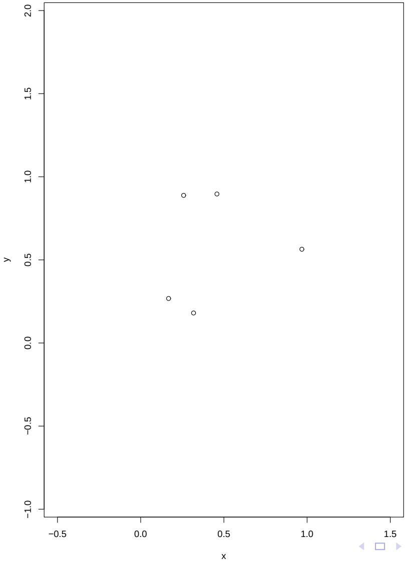
Data-based potential clusters

Graphs $\mathcal{G}(\delta)$ associated to the point process :

- Vertices : $\{1, \cdots, n\}$.
- Edges : $\{(i,j) : d(x_i, x_j) \leq \delta, 1 \leq i < n, i < j \leq n\}$.

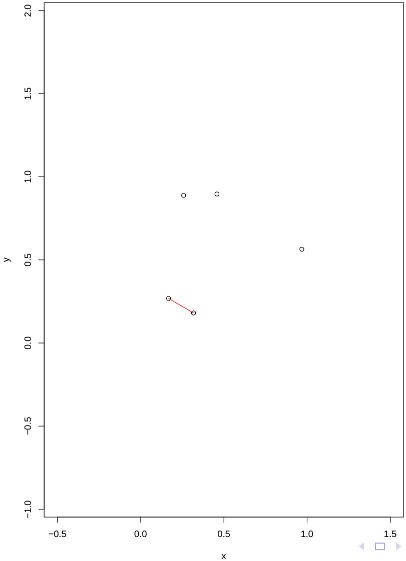
Data-based potential clusters

Data-based potential clusters



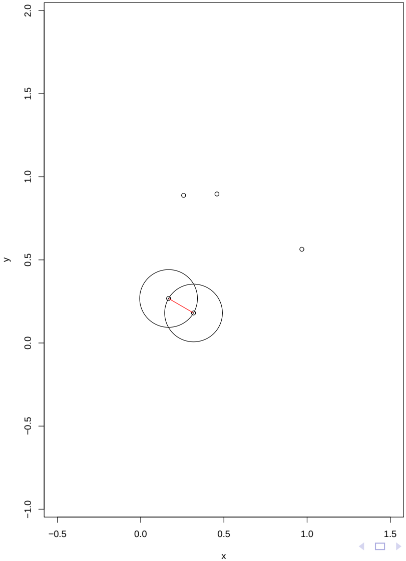
Data-based potential clusters

Data-based potential clusters



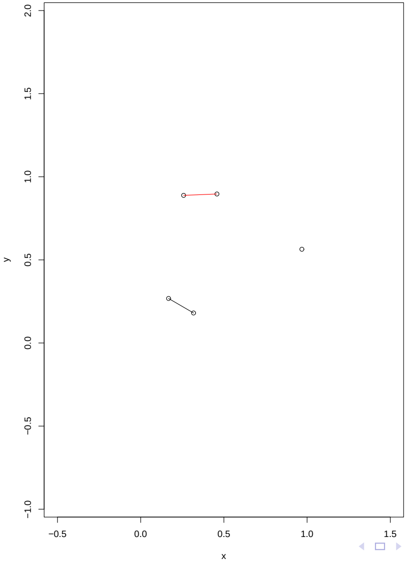
Data-based potential clusters

Data-based potential clusters



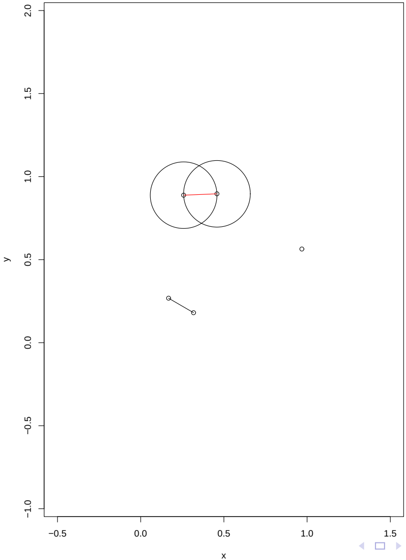
Data-based potential clusters

Data-based potential clusters



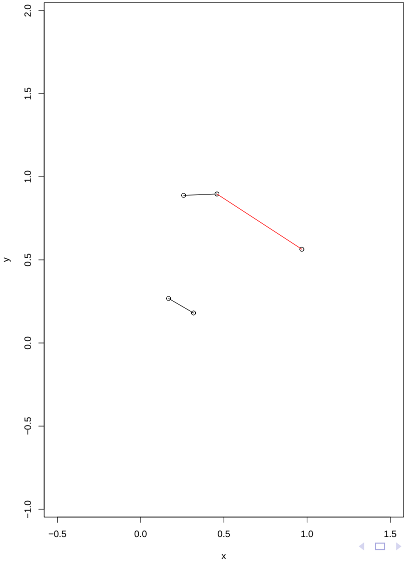
Data-based potential clusters

Data-based potential clusters



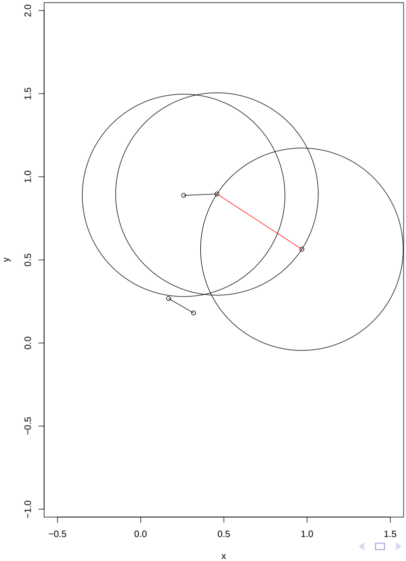
Data-based potential clusters

Data-based potential clusters



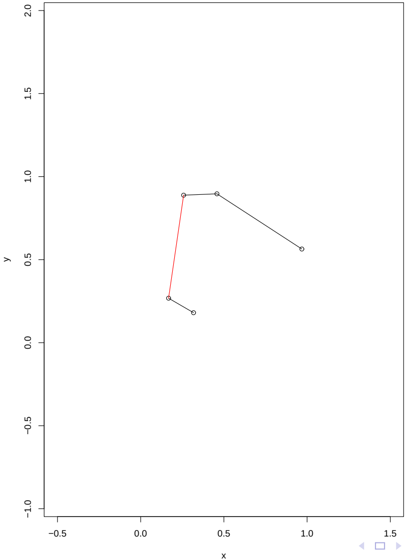
Data-based potential clusters

Data-based potential clusters



Data-based potential clusters

Data-based potential clusters



Data-based potential clusters

Data-based potential clusters

- Connected component of the edge i in $\mathcal{G}(\delta)$: $\mathcal{N}_i(\delta)$.

Data-based potential clusters

- Connected component of the edge i in $\mathcal{G}(\delta)$: $\mathcal{N}_i(\delta)$.
- $A_i(\delta) = \{x \in A : \exists j \in \mathcal{N}_i(\delta),\ d(x, x_j) \leq \delta\}$.

Data-based potential clusters

- Connected component of the edge i in $\mathcal{G}(\delta)$: $\mathcal{N}_i(\delta)$.
- $A_i(\delta) = \{x \in A : \exists j \in \mathcal{N}_i(\delta),\, d(x, x_j) \leq \delta\}$.
- Potential clusters :

  C$= \{A_i(\delta) : 1 \leq i \leq n,\, \delta \in \mathbb{R}^+\}$.

Data-based potential clusters

- Connected component of the edge i in $\mathcal{G}(\delta)$: $\mathcal{N}_i(\delta)$.
- $A_i(\delta) = \{x \in A : \exists j \in \mathcal{N}_i(\delta),\ d(x, x_j) \leq \delta\}$.
- Potential clusters :

  C$= \{A_i(\delta) : 1 \leq i \leq n,\ \delta \in \mathbb{R}^+\}$.
- Only $n - 1$ areas, arbitrarily shaped.

Outline

Outline

Concentration indices

Concentration indices

We evaluate the concentration in $Z \subset A$.

Concentration indices

We evaluate the concentration in $Z \subset A$.

- $n(Z) = \sharp\{i : X_i \in Z\}$.

Concentration indices

We evaluate the concentration in $Z \subset A$.

- $n(Z) = \sharp\{i : X_i \in Z\}$.
- $\mu(Z) = \frac{1}{n(Z)} \sum_{i:X_i \in Z} C_i$.

Concentration indices

We evaluate the concentration in $Z \subset A$.

- $n(Z) = \sharp\{i : X_i \in Z\}$.
- $\mu(Z) = \frac{1}{n(Z)} \sum_{i:X_i \in Z} C_i$.
- $\sigma(Z)^2 = \frac{1}{n(Z)} \sum_{i:X_i \in Z} \big(C_i - \mu(Z)\big)^2$.

Likelihood-based indices

Likelihood-based indices

We test the presence of a cluster in $Z \subset A$.

Likelihood-based indices

We test the presence of a cluster in $Z \subset A$.

- $H_0 : C_i \sim f_0$.

Likelihood-based indices

We test the presence of a cluster in $Z \subset A$.

- $H_0 : C_i \sim f_0$.
- $H_{1,Z} : C_i | X_i \sim f_Z \quad \mathbb{1}(X_i \in Z) + f_{\bar{Z}} \quad \mathbb{1}(X_i \in \bar{Z})$.

Likelihood-based indices

We test the presence of a cluster in $Z \subset A$.

- $H_0 : C_i \sim f_0$.
- $H_{1,Z} : C_i | X_i \sim f_Z \quad \mathbb{1}(X_i \in Z) + f_{\bar{Z}} \quad \mathbb{1}(X_i \in \bar{Z})$.

Likelihood ratio :

$$I(Z) = \frac{L_{1,Z}(X_1, \cdots, X_n, C_1, \cdots, C_n)}{L_0(X_1, \cdots, X_n, C_1, \cdots, C_n)} \quad \mathbb{1}\left(\mu(Z) > \mu(\bar{Z})\right).$$

Likelihood-based indices

Likelihood-based indices

The Exponential model

Likelihood-based indices

The Exponential model

➡ $H_0 : C_i \sim \mathcal{E}\big(1/\mu(A)\big)$.

Likelihood-based indices

The Exponential model

- $H_0 : C_i \sim \mathcal{E}\big(1/\mu(A)\big)$.
- $H_{1,Z} : C_i|X_i \sim \mathcal{E}(1/\mu(Z)\big)\ \mathbb{1}(X_i \in Z) + \mathcal{E}\big(1/\mu(\bar{Z})\big)\ \mathbb{1}(X_i \in \bar{Z})$.

Likelihood-based indices

The Exponential model

- $H_0 : C_i \sim \mathcal{E}\big(1/\mu(A)\big)$.
- $H_{1,Z} : C_i|X_i \sim \mathcal{E}\big(1/\mu(Z)\big)\ \mathbb{1}(X_i \in Z) + \mathcal{E}\big(1/\mu(\bar{Z})\big)\ \mathbb{1}(X_i \in \bar{Z})$.

Likelihood ratio :

$$l_{exp}(Z) = -n(Z)\log\big(\mu(Z)\big) - n(\bar{Z})\log\big(\mu(\bar{Z})\big).$$

Likelihood-based indices

Likelihood-based indices

The homoscedastic Gaussian model

Likelihood-based indices

The homoscedastic Gaussian model

➡ $H_0 : C_i \sim \mathcal{N}\left(\mu(A), \sigma(A)^2\right)$.

Likelihood-based indices

The homoscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\big(\mu(A), \sigma(A)^2\big)$.
- $H_{1,Z} : C_i | X_i \sim \mathcal{N}\big(\mu(Z), \sigma^2_{1,Z}\big) \ \mathbb{1}(X_i \in Z)$
  $+\mathcal{N}\big(\mu(\bar{Z}), \sigma^2_{1,Z}\big) \ \mathbb{1}(X_i \in \bar{Z})$

Likelihood-based indices

The homoscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\big(\mu(A), \sigma(A)^2\big)$.
- $H_{1,Z} : C_i | X_i \sim \mathcal{N}\big(\mu(Z), \sigma_{1,Z}^2\big)\ \mathbb{1}(X_i \in Z)$
  $+\mathcal{N}\big(\mu(\bar{Z}), \sigma_{1,Z}^2\big)\ \mathbb{1}(X_i \in \bar{Z})$

  where $\sigma_{1,Z}^2 = \frac{n(Z)\sigma(Z)^2 + n(\bar{Z})\sigma(\bar{Z})^2}{n}$.

Likelihood-based indices

The homoscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\left(\mu(A), \sigma(A)^2\right)$.
- $H_{1,Z} : C_i | X_i \sim \mathcal{N}\left(\mu(Z), \sigma^2_{1,Z}\right) \ \mathbb{1}(X_i \in Z)$
  $+\mathcal{N}\left(\mu(\bar{Z}), \sigma^2_{1,Z}\right) \ \mathbb{1}(X_i \in \bar{Z})$

  where $\sigma^2_{1,Z} = \frac{n(Z)\sigma(Z)^2 + n(\bar{Z})\sigma(\bar{Z})^2}{n}$.

Likelihood ratio :

$$I_{homgau}(Z) = \frac{1}{\sigma^2_{1,Z}}.$$

Likelihood-based indices

Likelihood-based indices

The heteroscedastic Gaussian model

Likelihood-based indices

The heteroscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\big(\mu(A), \sigma(A)^2\big)$.

Likelihood-based indices

The heteroscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\left(\mu(A), \sigma(A)^2\right)$.
- $H_{1,Z} : C_i | X_i \sim \mathcal{N}\left(\mu(Z), \sigma(Z)^2\right) \ \mathbb{1}(X_i \in Z)$
  $+\mathcal{N}\left(\mu(\bar{Z}), \sigma(\bar{Z})^2\right) \ \mathbb{1}(X_i \in \bar{Z})$.

Likelihood-based indices

The heteroscedastic Gaussian model

- $H_0 : C_i \sim \mathcal{N}\left(\mu(A), \sigma(A)^2\right)$.
- $H_{1,Z} : C_i | X_i \sim \mathcal{N}\left(\mu(Z), \sigma(Z)^2\right)\ \mathbb{1}(X_i \in Z)$
  $+\mathcal{N}\left(\mu(\bar{Z}), \sigma(\bar{Z})^2\right)\ \mathbb{1}(X_i \in \bar{Z})$.

Likelihood ratio :

$$I_{hetgau}(Z) = -n(Z)\log\left(\sigma(Z)^2\right) - n(\bar{Z})\log\left(\sigma(\bar{Z})^2\right).$$

No distribution assumption : Mann-Whitney test

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.
- $RS(Z) = \sum_{i:X_i \in Z} R_i$.

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.
- $RS(Z) = \sum_{i:X_i \in Z} R_i$.
- Under H_0, $\mathbb{E}\big(RS(Z)\big) = M(Z) = \frac{n(Z)(n+1)}{2}$.

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.
- $RS(Z) = \sum_{i:X_i \in Z} R_i$.
- Under H_0, $\mathbb{E}\big(RS(Z)\big) = M(Z) = \frac{n(Z)(n+1)}{2}$.
- Under H_0, $Var\big(RS(Z)\big) = V(Z) = \frac{n(Z)n(\bar{Z})(n+1)}{12}$

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.
- $RS(Z) = \sum_{i: X_i \in Z} R_i$.
- Under H_0, $\mathbb{E}\big(RS(Z)\big) = M(Z) = \frac{n(Z)(n+1)}{2}$.
- Under H_0, $Var\big(RS(Z)\big) = V(Z) = \frac{n(Z)n(\bar{Z})(n+1)}{12}$
- Under H_0, $\frac{RS(Z)-M(Z)}{\sqrt{V(Z)}} \underset{d}{\longrightarrow} \mathcal{N}(0,1)$.

No distribution assumption : Mann-Whitney test

- R_j is the rank of C_j among the $C_i, 1 \leq i \leq n$.
- $RS(Z) = \sum_{i:X_i \in Z} R_i$.
- Under H_0, $\mathbb{E}\big(RS(Z)\big) = M(Z) = \frac{n(Z)(n+1)}{2}$.
- Under H_0, $Var\big(RS(Z)\big) = V(Z) = \frac{n(Z)n(\bar{Z})(n+1)}{12}$
- Under H_0, $\frac{RS(Z)-M(Z)}{\sqrt{V(Z)}} \underset{d}{\longrightarrow} \mathcal{N}(0,1)$.

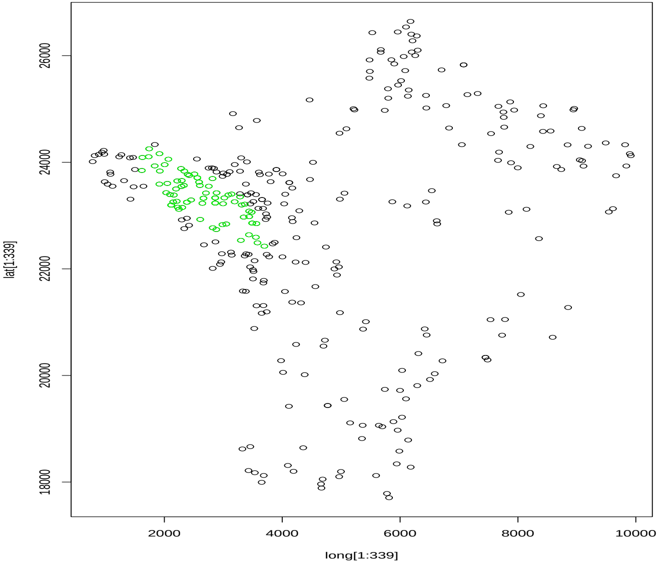
$$I_{rank}(Z) = \frac{RS(Z) - M(Z)}{\sqrt{V(Z)}}.$$

Outline

Outline

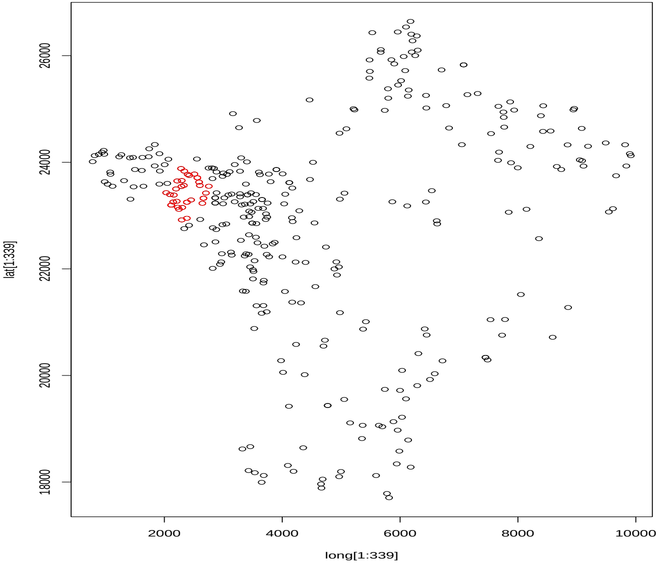
The farms data set

The farms data set



The farms data set

The farms data set



Astronomical data

Astronomical data

Observation of a cubic part of the Universe : $(20 \text{ Mpc})^3$.

Astronomical data

Observation of a cubic part of the Universe : $(20 \text{ Mpc})^3$.
$1 \text{ Mpc} = 3 \times 10^{22}$ metres .

Astronomical data

Observation of a cubic part of the Universe : $(20 \text{ Mpc})^3$.
1 Mpc $= 3 \times 10^{22}$ metres .

➡ Locations of the galaxies : $X_1, \cdots, X_n$.

Astronomical data

Observation of a cubic part of the Universe : $(20\ \text{Mpc})^3$.
$1\ \text{Mpc} = 3 \times 10^{22}$ metres .

- Locations of the galaxies : $X_1, \cdots, X_n$.
- Light intensities : $C_1, \cdots, C_n$.

Astronomical data

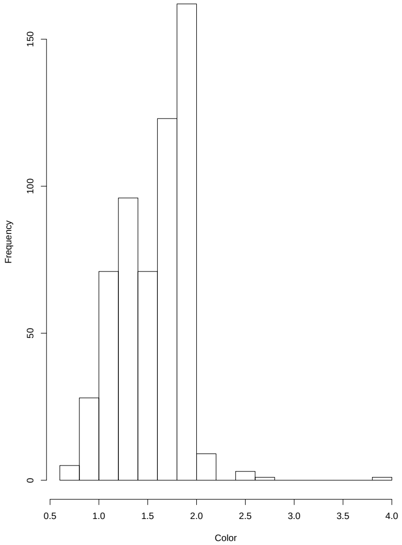
Observation of a cubic part of the Universe : $(20\ \text{Mpc})^3$.
1 Mpc $= 3 \times 10^{22}$ metres .

- Locations of the galaxies : $X_1, \cdots, X_n$.
- Light intensities : $C_1, \cdots, C_n$.

Goal : detecting areas where galaxies are "redder".

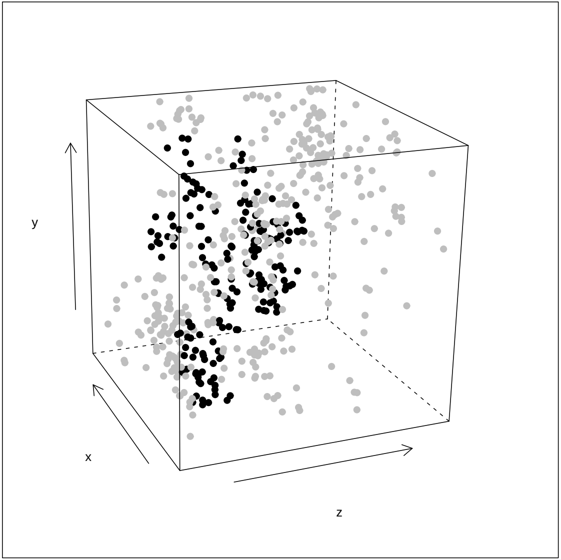
Light intensities

Light intensities



Astronomical data

Astronomical data



Conclusion

Conclusion

- ➡ Graph-based possible clusters, useful in 3D or for multidimensional data.

Conclusion

- Graph-based possible clusters, useful in 3D or for multidimensional data.
- The MW concentration index is hypothesis-free.

Conclusion

- Graph-based possible clusters, useful in 3D or for multidimensional data.
- The MW concentration index is hypothesis-free.
- Simulation study to compare concentration indices.