Nonparametric estimation of directional highest density regions

Paula Saavedra-Nieves* and Rosa M. Crujeiras*

*Universidade de Santiago de Compostela

VII Xornada de Usuarios de R en Galicia
15th October 2020, Santiago de Compostela

Table of contents

Main points:

- Plug-in estimation of highest density regions in the Euclidean setting.
  - Analysis of leukaemia data set.
- Plug-in estimation of highest density regions in the directional setting.
  - Analysis of sandhoppers orientation.
  - Analysis of earthquakes distribution.

A real problem: Leukaemia analysis

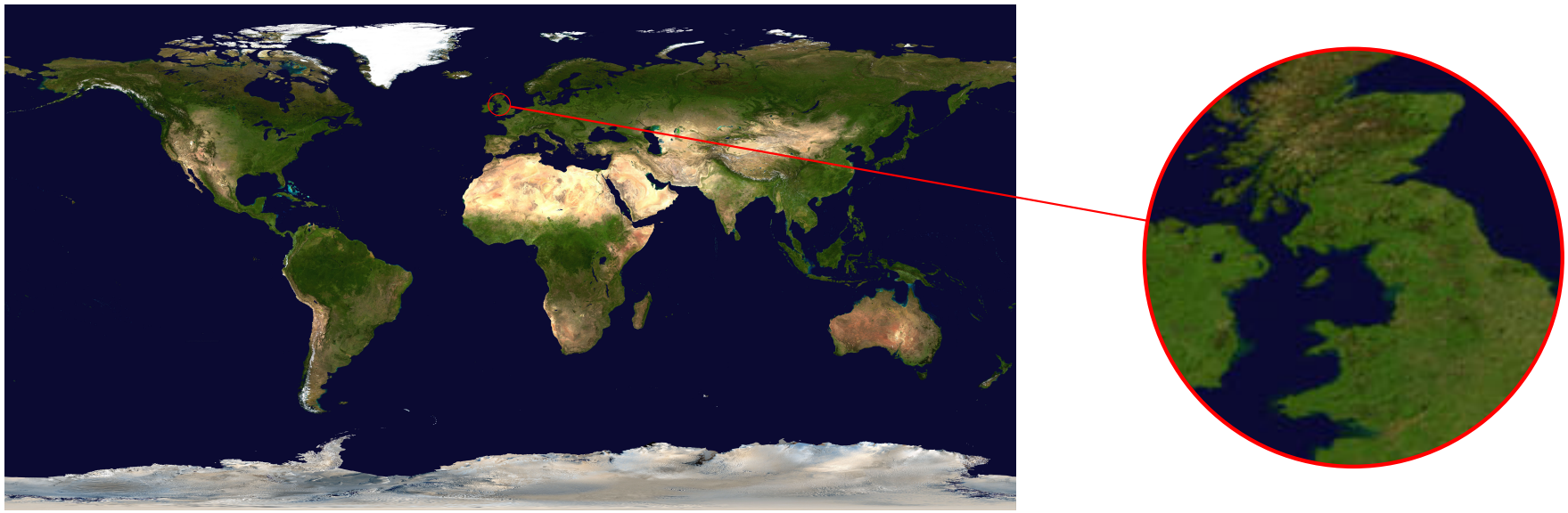

Figure: Geographical location of the North West of England.

NASA Terra satellite.

Leukaemia real example

Is there an excess of case intensity over that of population?

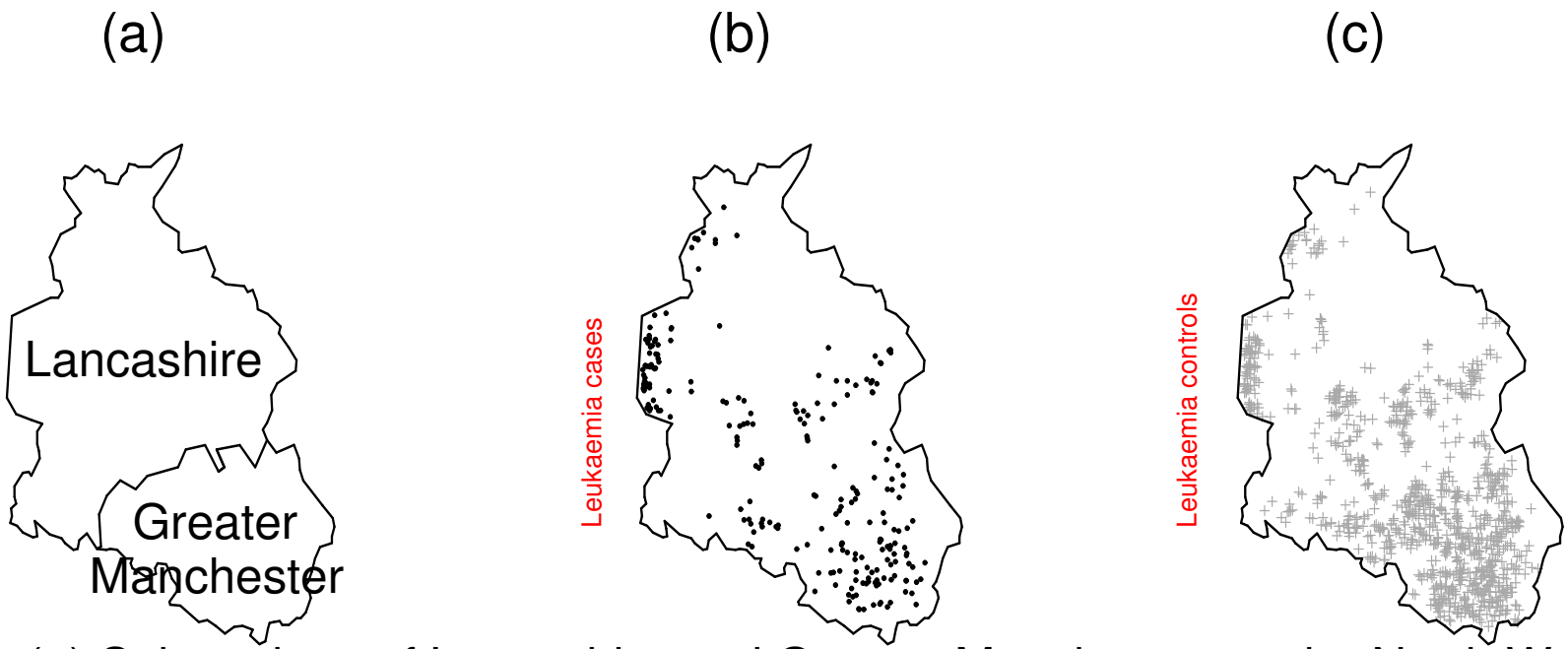

Figure: (a) Sub-regions of Lancashire and Greater Manchester on the North West of England. (b) distribution of 233 cases of diagnosed leukaemia between 1982 and 1998. (c) 988 controls on Lancashire and Greater Manchester.

Prof. Peter J. Diggle's website, Lancaster University.

Highest Density Regions estimation in $\mathbb{R}^d$

Given a random sample of points $\mathcal{X}_n = \{X_1, ..., X_n\}$ of a random vector X with values in $\mathbb{R}^d$, reconstructing the $t-$level set

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$$

where f denotes the density function of X and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geq f_\tau\}$$

where f_τ can be seen as the largest constant such that

$$\mathbb{P}(X \in L(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by f.

Highest Density Regions estimation in $\mathbb{R}^d$

Given a random sample of points $\mathcal{X}_n = \{X_1, ..., X_n\}$ of a random vector X with values in $\mathbb{R}^d$, reconstructing the $t-$level set

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$$

where f denotes the density function of X and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geq f_\tau\}$$

where f_τ can be seen as the largest constant such that

$$\mathbb{P}(X \in L(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by f.

Highest Density Regions estimation in $\mathbb{R}^d$

Given a random sample of points $\mathcal{X}_n = \{X_1, ..., X_n\}$ of a random vector X with values in $\mathbb{R}^d$, reconstructing the $t-$level set

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$$

where f denotes the density function of X and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geq f_\tau\}$$

where f_τ can be seen as the largest constant such that

$$\mathbb{P}(X \in L(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by f.

One-dimensional example

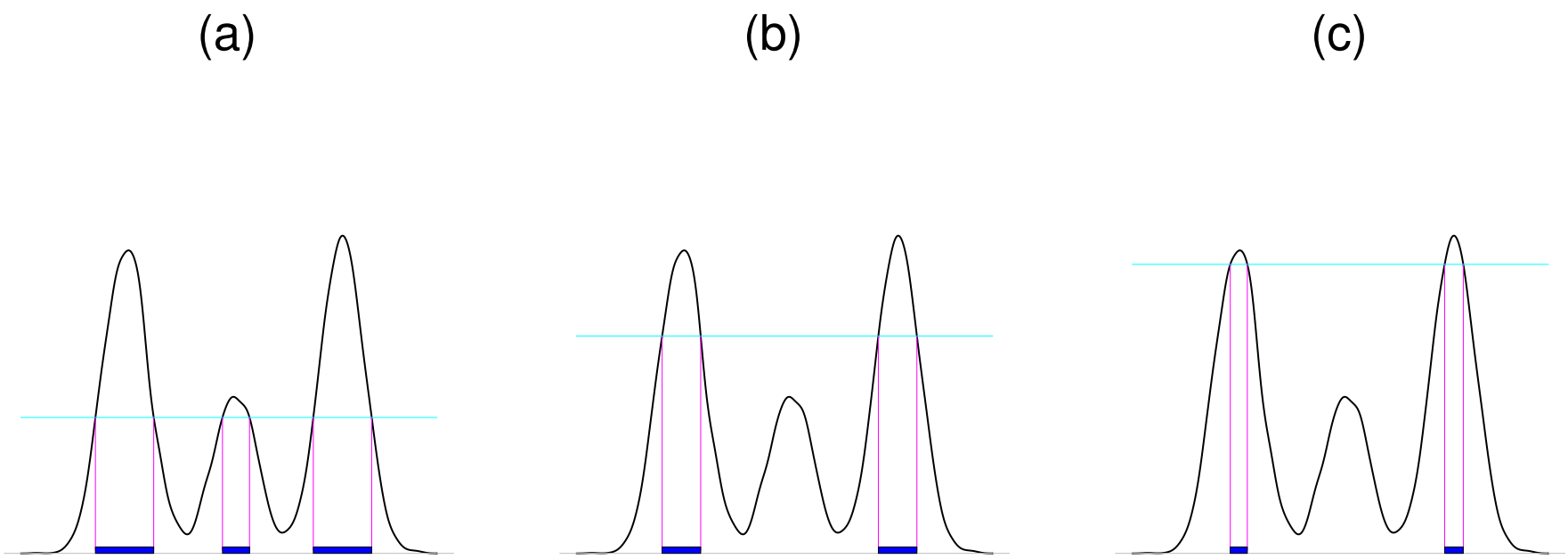

Figure: One-dimensional HDRs for a trimodal density taking τ equal to (a) 0.25, (b) 0.5 and (c) 0.75.

For low values of τ, HDR looks like the support of the distribution.

HDRs estimation

Two alternative routes for estimating a HDR:

- The only information we have comes from the sample.
  - Plug-in methodology.
- Some geometric properties of the HDR are known.
  - Excess mass methodology.
  - Hybrid methodology.

HDRs estimation

Two alternative routes for estimating a HDR:

- The only information we have comes from the sample.
  - Plug-in methodology.
- Some geometric properties of the HDR are known.
  - Excess mass methodology.
  - Hybrid methodology.

Plug-in estimation of HDRs

Plug-in methods propose

$$\hat{L}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\}$$

as an **estimator for** $L(\tau)$ where

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where K is a symmetric density function with $K_H(z) = \mid H \mid^{-1/2} K(H^{1/2}z)$, H denotes the bandwidth matrix and $\hat{f}_\tau = f_\tau(f_n)$ denotes an estimator of the threshold f_τ.

Hyndman (1996) estimated f_τ as the quantile τ of the empirical distribution of $f_n(X_1), \cdots, f_n(X_n)$.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Samworth, R.J. and Wand , M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. Annals of Statistics, 38, 1767–1792.

Plug-in estimation of HDRs

Plug-in methods propose

$$\hat{L}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\}$$

as an **estimator for** $L(\tau)$ where

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where K is a symmetric density function with $K_H(z) = \mid H \mid^{-1/2} K(H^{1/2}z)$, H denotes the bandwidth matrix and $\hat{f}_\tau = f_\tau(f_n)$ denotes an estimator of the threshold f_τ.

Hyndman (1996) estimated f_τ as the quantile τ of the empirical distribution of $f_n(X_1), \cdots, f_n(X_n)$.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Samworth, R.J. and Wand , M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. Annals of Statistics, 38, 1767–1792.

Plug-in estimation of HDRs

Plug-in methods propose

$$\hat{L}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\}$$

as an **estimator for** $L(\tau)$ where

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where K is a symmetric density function with $K_H(z) = \mid H \mid^{-1/2} K(H^{1/2}z)$, H denotes the bandwidth matrix and $\hat{f}_\tau = f_\tau(f_n)$ denotes an estimator of the threshold f_τ.

Hyndman (1996) estimated f_τ as the quantile τ of the empirical distribution of $f_n(X_1), \cdots, f_n(X_n)$.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Samworth, R.J. and Wand , M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. Annals of Statistics, 38, 1767–1792.

Analysis of leukaemia data set

Is there an excess of case intensity over that of population?

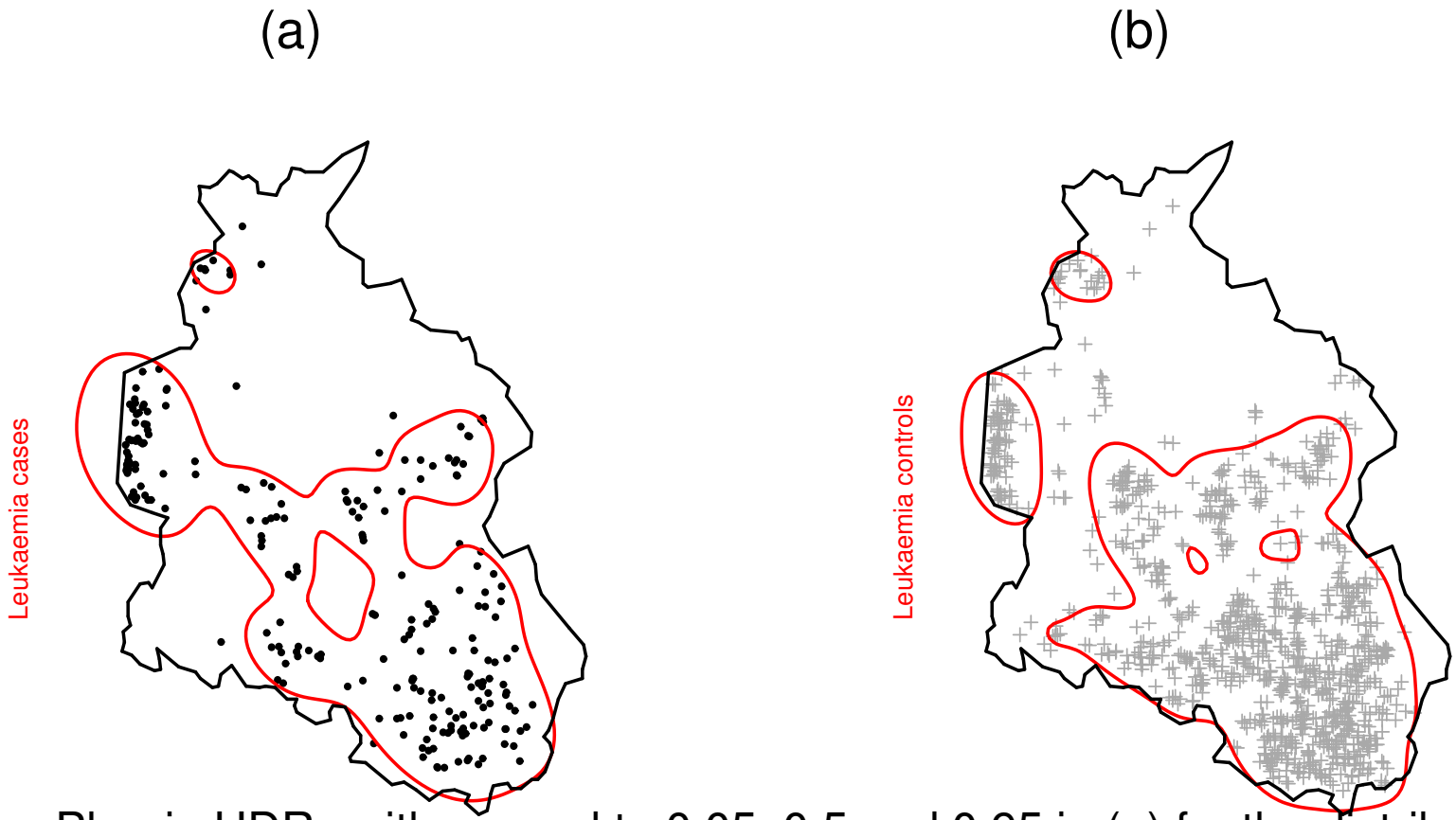

Figure: Plug-in HDRs with τ equal to 0.05, 0.5 and 0.95 in (a) for the distribution of 322 cases diagnosed of leukaemia and in (b) for the distribution of 988 controls.

Packages `ks` and `hdrcde`.

Analysis of leukaemia data set

Is there an excess of case intensity over that of population?

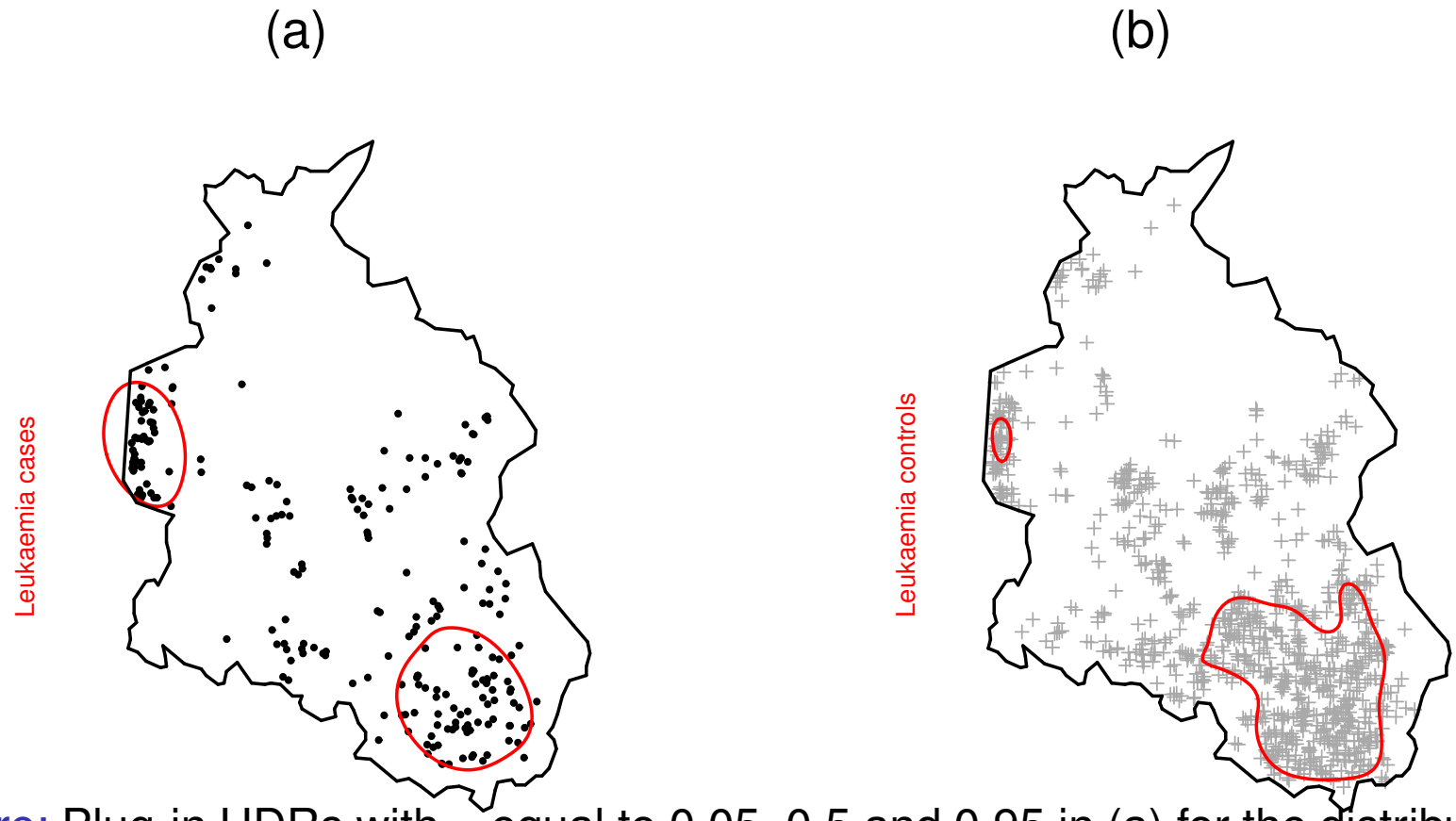

Figure: Plug-in HDRs with τ equal to 0.05, 0.5 and 0.95 in (a) for the distribution of 322 cases diagnosed of leukaemia and in (b) for the distribution of 988 controls.

Packages `ks` and `hdrcde`.

Analysis of leukaemia data set

Is there an excess of case intensity over that of population?

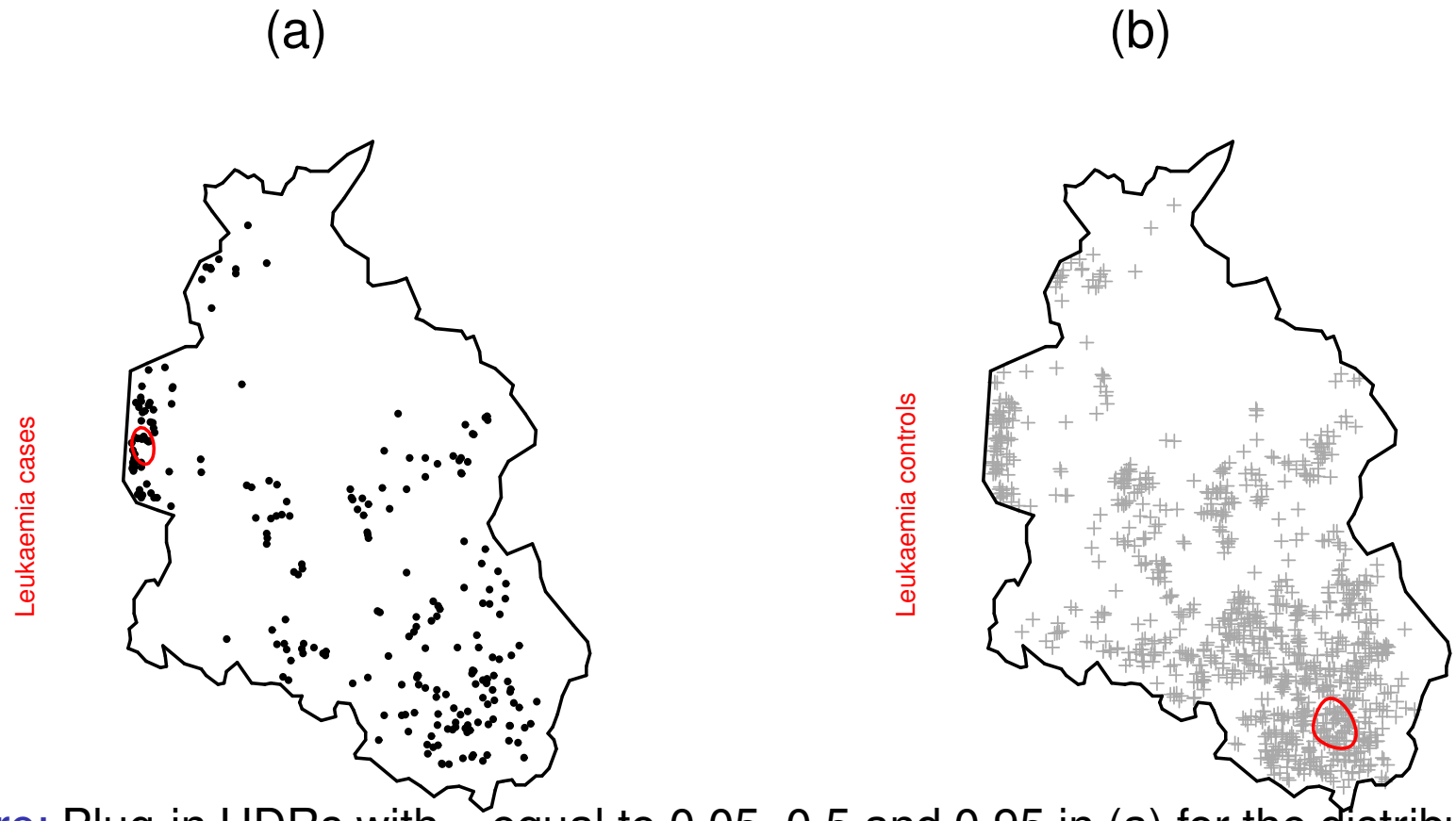

Figure: Plug-in HDRs with τ equal to 0.05, 0.5 and 0.95 in (a) for the distribution of 322 cases diagnosed of leukaemia and in (b) for the distribution of 988 controls.

Packages `ks` and `hdrcde`.

Other real problems (I): Sandhoppers orientation

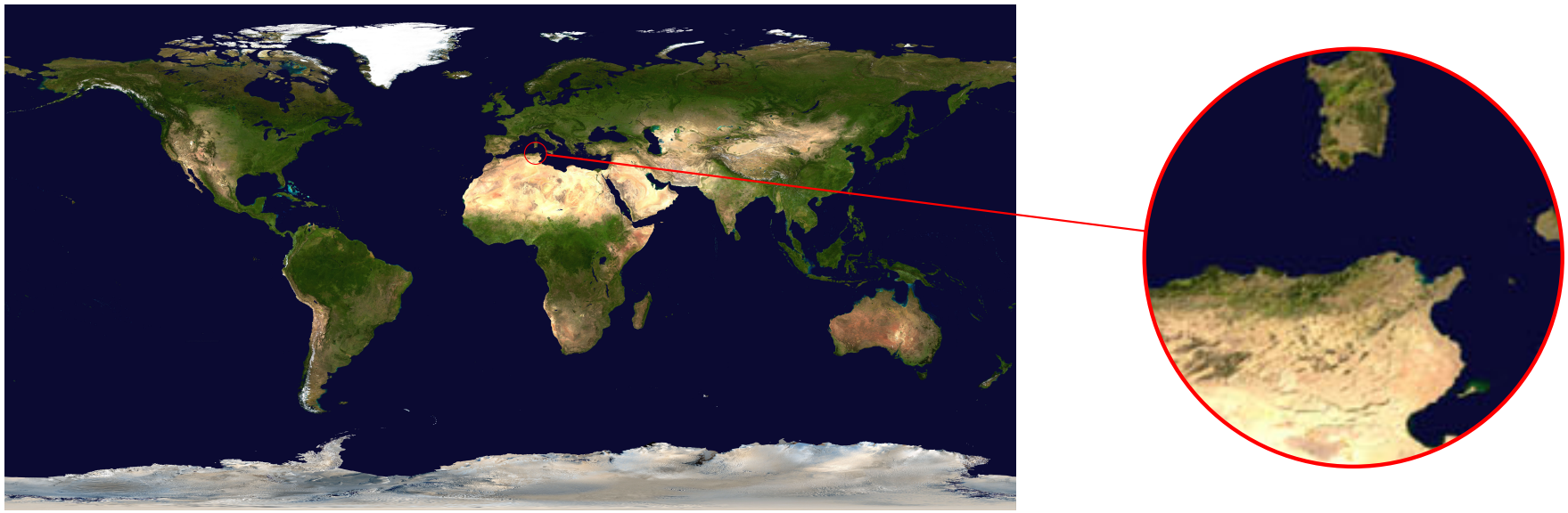

Figure: Geographical location of Zouara beach in the Tunisian northwestern coast.

NASA Terra satellite.

Other real problems (I): Sandhoppers orientation

Does the moment of the day play a significant role in sandhoppers behavior?

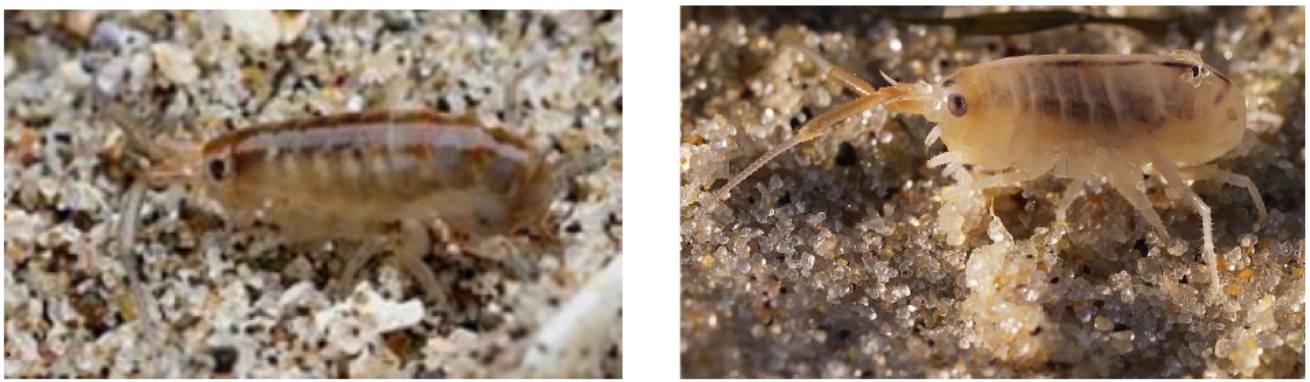

Figure: Talorchestia brito (left) and Talitrus saltator (right).

Scapini, F., Aloia, A., Bouslama, M. F., Chelazzi, L., Colombini, I., ElGtari, M., Fallaci, M. and Marchetti, G. M. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, Talitrus saltator and Talorchestia brito, from an exposed Mediterranean beach, Behav. Ecol. Sociobiol., 51(5), 403-414.

Other real problems (I): Sandhoppers orientation

Does the moment of the day play a significant role in sandhoppers behavior?

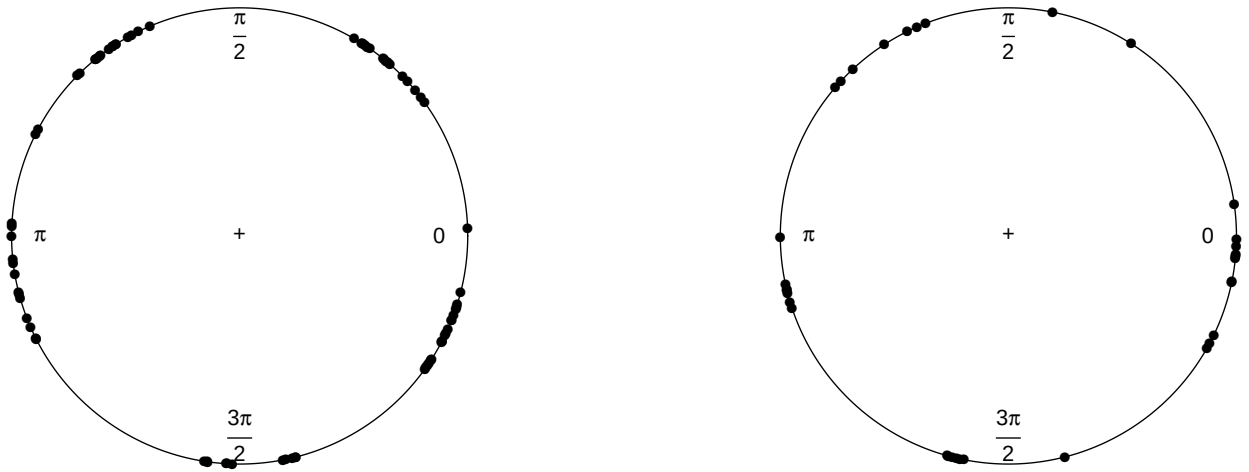

Figure: Orientation data (slightly jittered) corresponding to males of the specie Talitrus saltator registered in the morning (left) and in the afternoon (right) in April.

Scapini, F., Aloia, A., Bouslama, M. F., Chelazzi, L., Colombini, I., ElGtari, M., Fallaci, M. and Marchetti, G. M. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, Talitrus saltator and Talorchestia brito, from an exposed Mediterranean beach, Behav. Ecol. Sociobiol., 51(5), 403-414.

Other real problems (II): Earthquakes distribution

Where are earthquakes likely to happen?

Play/Pause

Figure: Distribution of earthquakes around the world between October 2004 and April 2020.

European-Mediterranean Seismological Centre: www.emsc-csem.org.

HDRs estimation in S^{d-1}

Given a random sample of points $\mathcal{Y}_n = \{Y_1, ..., Y_n\}$ of a random vector Y with values in the unit sphere S^{d-1}, reconstructing the $t-$level set

$$G_g(t) = \{y \in S^{d-1} : g(y) \geq t\}$$

where g denotes the directional density function of Y and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L_g(\tau) = \{y \in S^{d-1} : g(y) \geq g_\tau\}$$

where g_τ can be seen as the largest constant such that

$$\mathbb{P}(Y \in L_g(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by g.

HDRs estimation in S^{d-1}

Given a random sample of points $\mathcal{Y}_n = \{Y_1, ..., Y_n\}$ of a random vector Y with values in the unit sphere S^{d-1}, reconstructing the $t-$level set

$$G_g(t) = \{y \in S^{d-1} : g(y) \geq t\}$$

where g denotes the directional density function of Y and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L_g(\tau) = \{y \in S^{d-1} : g(y) \geq g_\tau\}$$

where g_τ can be seen as the largest constant such that

$$\mathbb{P}(Y \in L_g(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by g.

HDRs estimation in S^{d-1}

Given a random sample of points $\mathcal{Y}_n = \{Y_1, ..., Y_n\}$ of a random vector Y with values in the unit sphere S^{d-1}, reconstructing the $t-$level set

$$G_g(t) = \{y \in S^{d-1} : g(y) \geq t\}$$

where g denotes the directional density function of Y and $t > 0$. Or, if the practitioner fixes a value $\tau \in (0, 1)$, estimating the Highest Density Region (HDR) with probability content $1 - \tau$

$$L_g(\tau) = \{y \in S^{d-1} : g(y) \geq g_\tau\}$$

where g_τ can be seen as the largest constant such that

$$\mathbb{P}(Y \in L_g(\tau)) \geq 1 - \tau$$

with respect to the distribution induced by g.

Circular examples

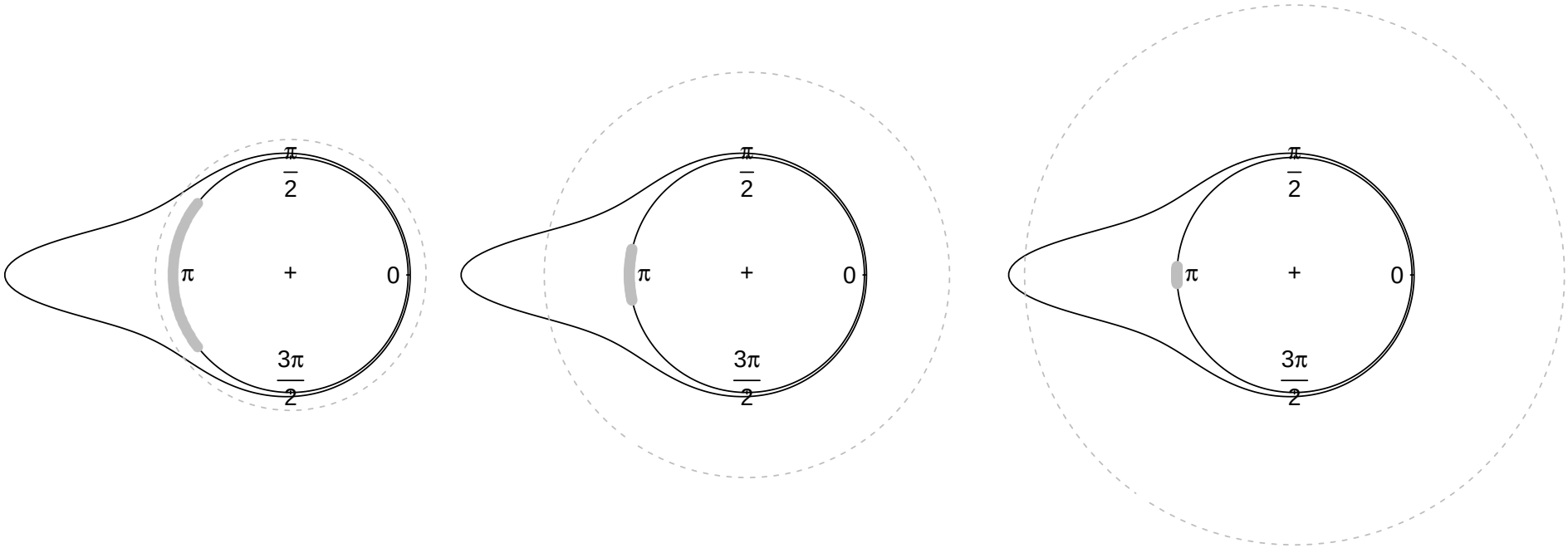

Figure: HDRs for $\tau = 0.2$ (first column), $\tau = 0.5$ (second column) and $\tau = 0.8$ (third column).

These densities are in package `NPCirc`.

Circular examples

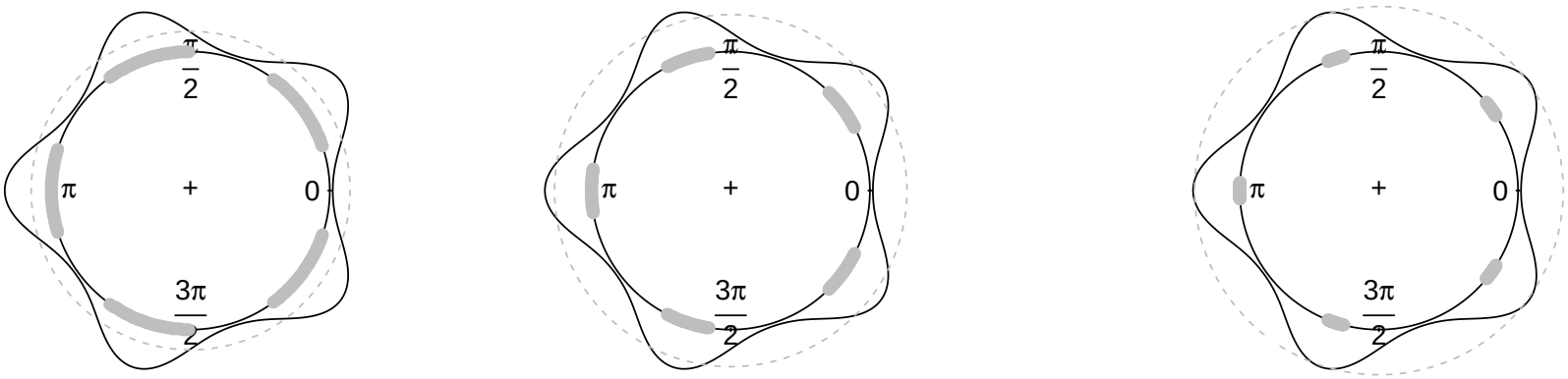

Figure: HDRs for $\tau = 0.2$ (first column), $\tau = 0.5$ (second column) and $\tau = 0.8$ (third column).

 These densities are in package `NPCirc`.

Circular examples

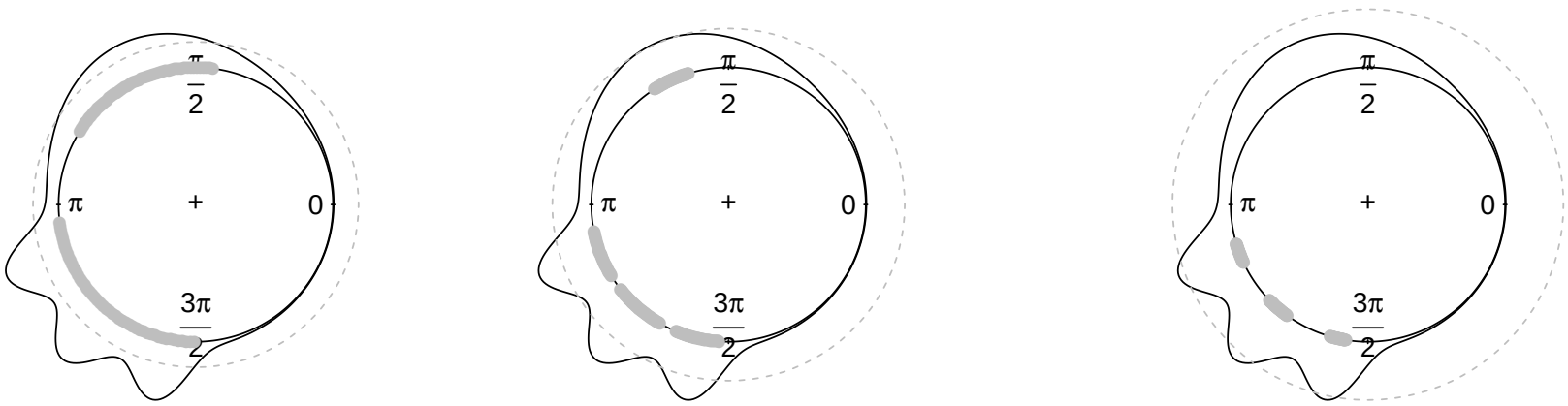

Figure: HDRs for $\tau = 0.2$ (first column), $\tau = 0.5$ (second column) and $\tau = 0.8$ (third column).

These densities are in package `NPCirc`.

Spherical examples

Figure: Finite mixtures of von Mises-Fisher spherical models for simulations. HDRs are represented for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$.

Self-programmed densities using packages `movMF` and `Directional`.

Plug-in estimation of directional HDRs

Plug-in methods propose

$$\hat{L}_g(\tau) = \{y \in S^{d-1} : g_n(y) \geq \hat{g}_\tau\}$$

as an **estimator for** $L_g(\tau)$ where

$$g_n(y) = \frac{1}{n}\sum_{i=1}^{n} K_{vM}(y; Y_i; 1/h^2),$$

where $1/h^2 > 0$ is concentration parameter and K_{vM} denotes the von Mises-Fisher kernel density.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Saavedra-Nieves, P. and Crujeiras, R. M. (2020). Nonparametric estimation of directional highest density regions. arXiv preprint arXiv:2009.08915.

Plug-in estimation of directional HDRs

Plug-in methods propose

$$\hat{L}_g(\tau) = \{y \in S^{d-1} : g_n(y) \geq \hat{g}_\tau\}$$

as an **estimator for** $L_g(\tau)$ where

$$g_n(y) = \frac{1}{n}\sum_{i=1}^{n} K_{vM}(y; Y_i; , 1/h^2),$$

where $1/h^2 > 0$ is concentration parameter and K_{vM} denotes the von Mises-Fisher kernel density.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Saavedra-Nieves, P. and Crujeiras, R. M. (2020). Nonparametric estimation of directional highest density regions. arXiv preprint arXiv:2009.08915.

Plug-in estimation of directional HDRs

Plug-in methods propose

$$\hat{L}_g(\tau) = \{y \in S^{d-1} : g_n(y) \geq \hat{g}_\tau\}$$

as an **estimator for** $L_g(\tau)$ where

$$g_n(y) = \frac{1}{n}\sum_{i=1}^{n} K_{vM}(y; Y_i; , 1/h^2),$$

where $1/h^2 > 0$ is concentration parameter and K_{vM} denotes the von Mises-Fisher kernel density.


Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.


Saavedra-Nieves, P. and Crujeiras, R. M. (2020). Nonparametric estimation of directional highest density regions. arXiv preprint arXiv:2009.08915.

Circular examples

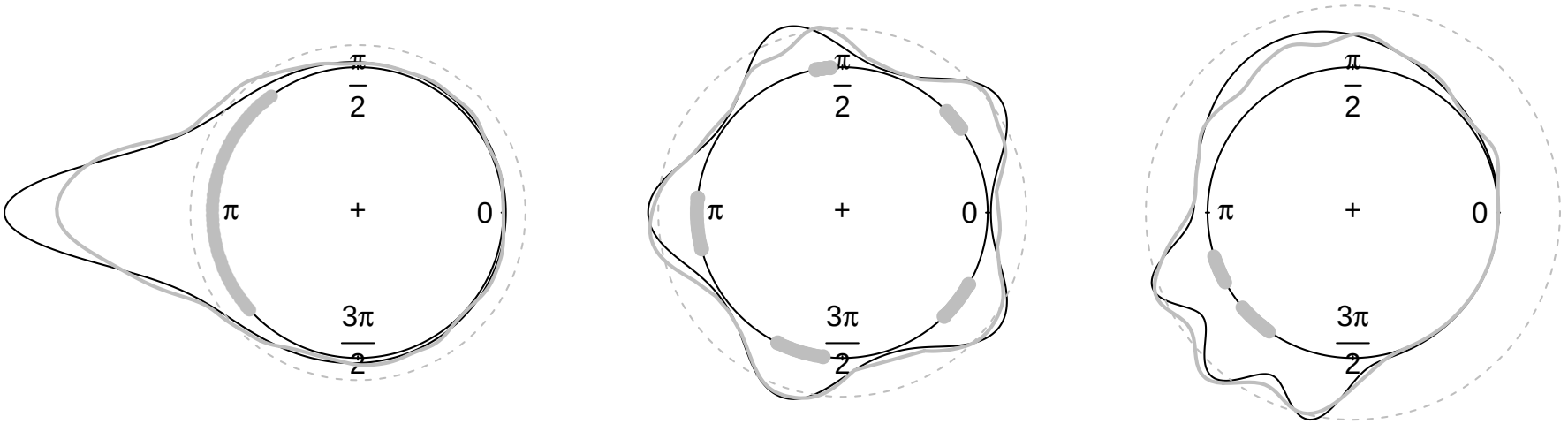

Figure: Plug-in HDRs from $\mathcal{Y}_{250}$ for three different circular densities with τ_1 (first column), τ_2 (second column) and τ_3 (third column) verifying $0 < \tau_1 < \tau_2 < \tau_3$.

Package `NPCirc`.

Spherical examples

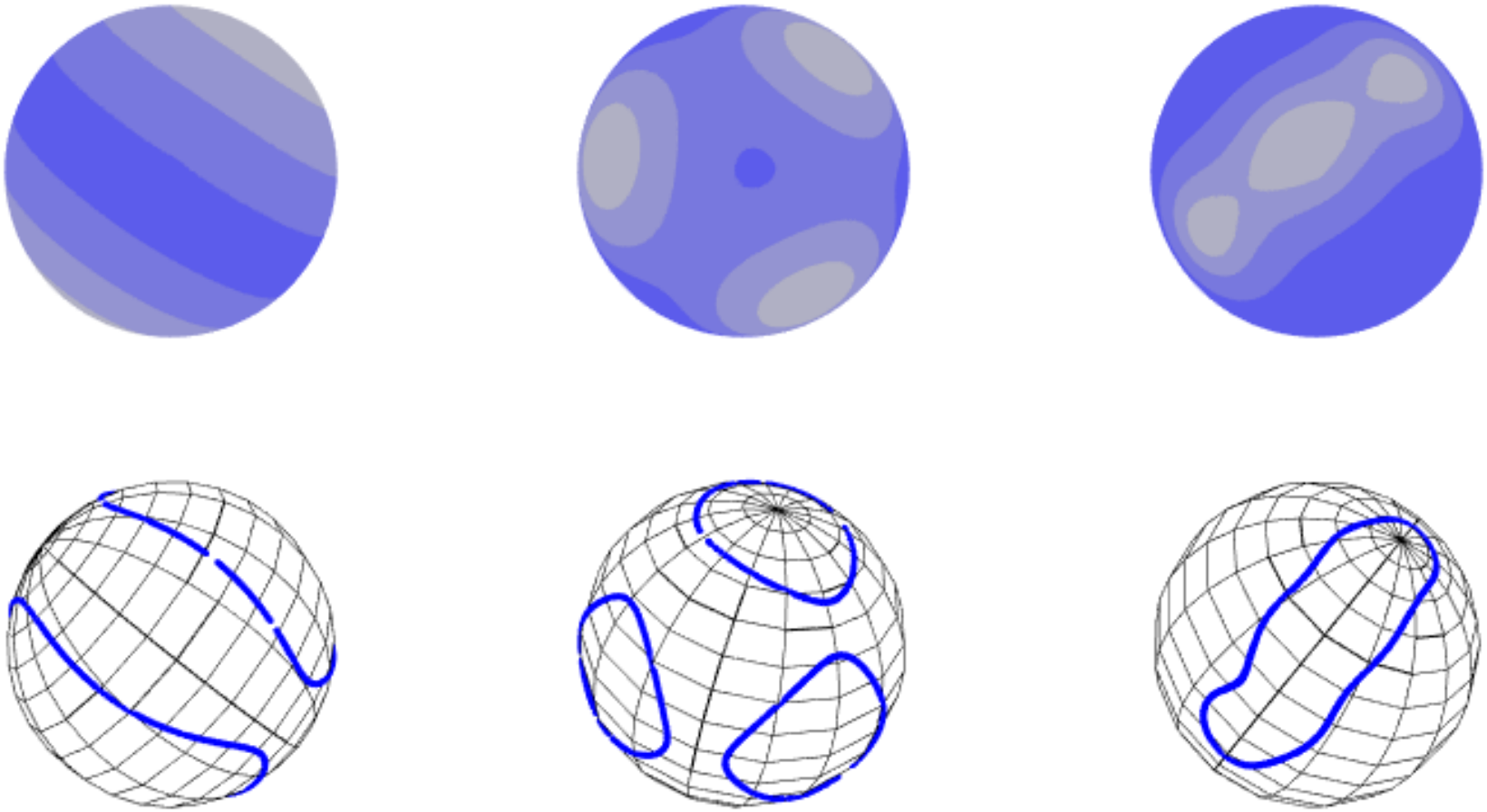

Figure: Plug-in HDRs from $\mathcal{Y}_{2500}$ for three different spherical densities with $\tau = 0.5$.

Package `Directional`.

Analysis of sandhoppers orientation

Does the month of the year play a significant role in sandhoppers behavior?

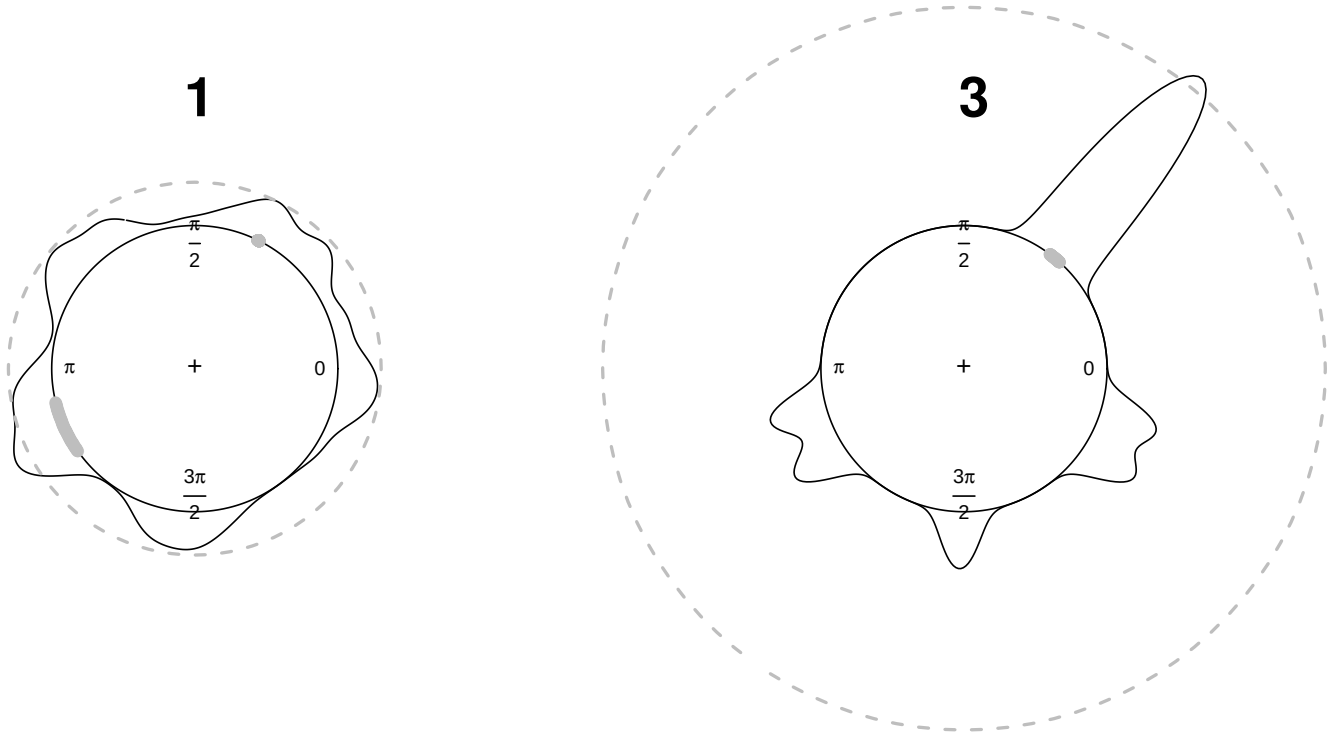

Figure: Plug-in HDRs when $\tau = 0.8$. The largest modes of these distributions can be observed.

Regions 1 and 3 is correspond to the orientation for males of the specie Talorchestia Brito when the orientation is measure in morning during October and April, respectively.

Analysis of sandhoppers orientation

Does the moment of the day play a significant role in sandhoppers behavior?

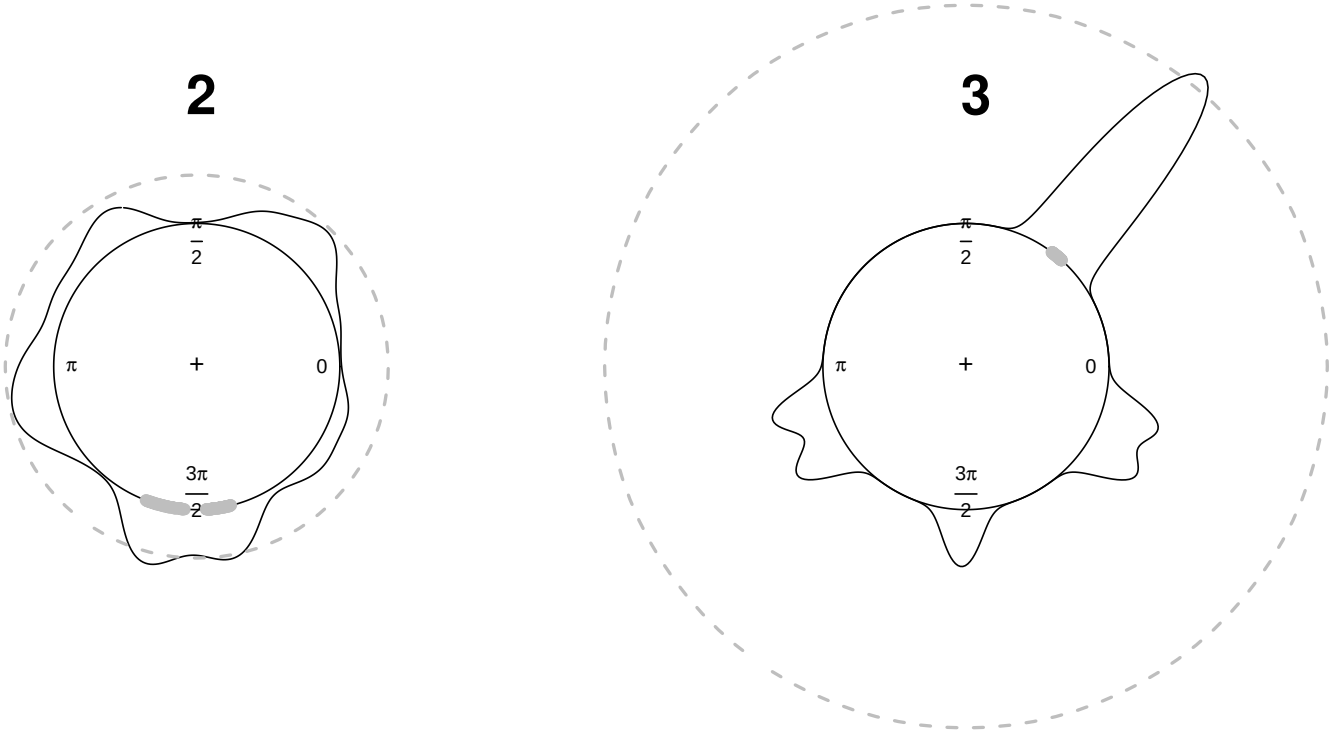

Figure: Plug-in HDRs when $\tau = 0.8$. The largest modes of these distributions can be observed.

Regions 2 and 3 correspond to the orientation for males of the specie Talorchestia Brito when the orientation is measure in noon and morning during April, respectively.

Analysis of earthquakes distribution

Play/Pause

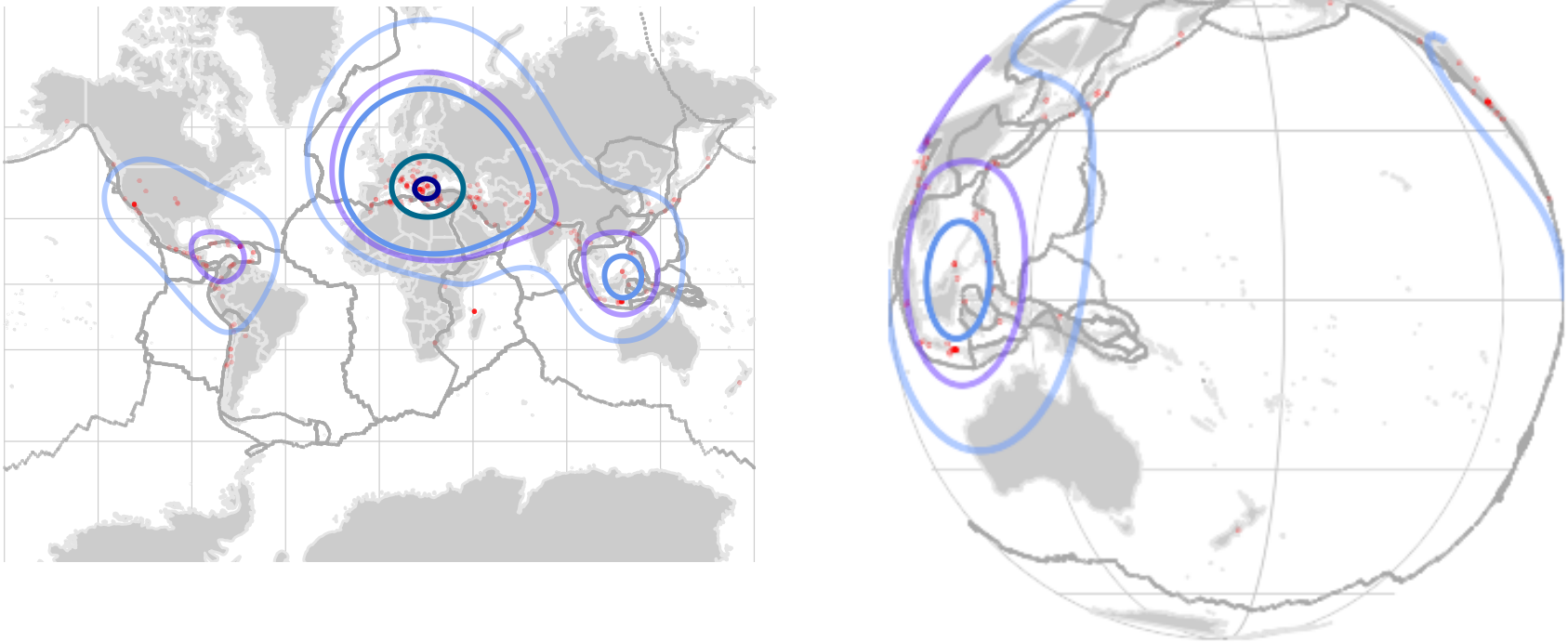

Figure: Distribution of earthquakes around the world between October 2004 and April 2020 (red color). Contours of plug-in HDRs for $\tau_1 = 0.1$, $\tau_2 = 0.3$, $\tau_3 = 0.5$, $\tau_4 = 0.7$ and $\tau_5 = 0.9$ (bluish colors).

Forthcoming ...

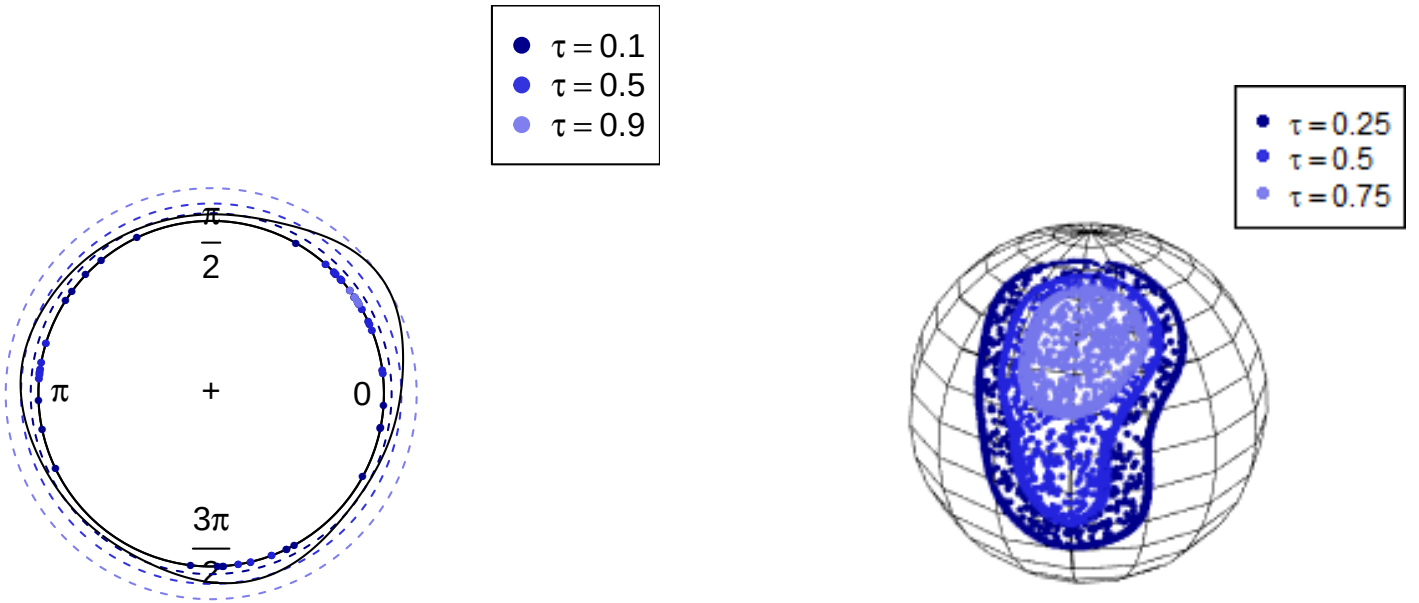

Figure: Circular and spherical scatterplots from $\mathcal{Y}_{50}$ and $\mathcal{Y}_{1500}$, respectively.

We are developing a new package called `HDiR` including these new tools.

Thanks!

Nonparametric estimation of directional highest density regions

Paula Saavedra-Nieves* and Rosa M. Crujeiras*

*Universidade de Santiago de Compostela

VII Xornada de Usuarios de R en Galicia
15th October 2020, Santiago de Compostela