Algunos ejemplos de la utilidad de R para el análisis estadístico en estudios de epidemiología ambiental desarrollados en el CREAL

Jose Barrera-Gómez[a]

jbarrera@creal.cat

[a]Centre for Research in Environmental Epidemiology (CREAL)

26 de septiembre de 2011

Índice

Parc de Recerca Biomèdica de Barcelona (PRBB)

Parc de Recerca Biomèdica de Barcelona (PRBB)

El Parc de Recerca Biomèdica de Barcelona

- Iniciativa de la Generalitat de Catalunya, el Ayuntamiento de Barcelona y la Universitat Pompeu Fabra (UPF),
- Conexión física con el Hospital del Mar de Barcelona,
- Uno de los principales núcleos de investigación biomédica a nivel internacional,
- 1200 trabajadores,
- 30 % personal científico extranjero. Más de 50 paises. Europa, América, Asia,
- Perfil joven: 65 % menores de 35 años,
- Perfil significativamente[1] femenino: 60 % mujeres,
- Inversión en I+D $\approx$ 70 mill. €/anuales,
- Colaboración docente en estudios de grado, máster y/o doctorado: UPF, UAB , UB, UPC,...

[1] p–valor $< 10^{-11}$.

El Parc de Recerca Biomèdica de Barcelona

7 centros públicos de investigación ($\approx$ 100 grupos) coordinados entre sí:

- Informática biomédica, epigenética, biología celular, farmacología, genética humana,...
- ... y Epidemiología y salud pública $\Longrightarrow$ Epidemiología ambiental $\Longrightarrow$ CREAL

El CREAL: objetivo

Centro de Investigación en Epidemiología Ambiental (CREAL)

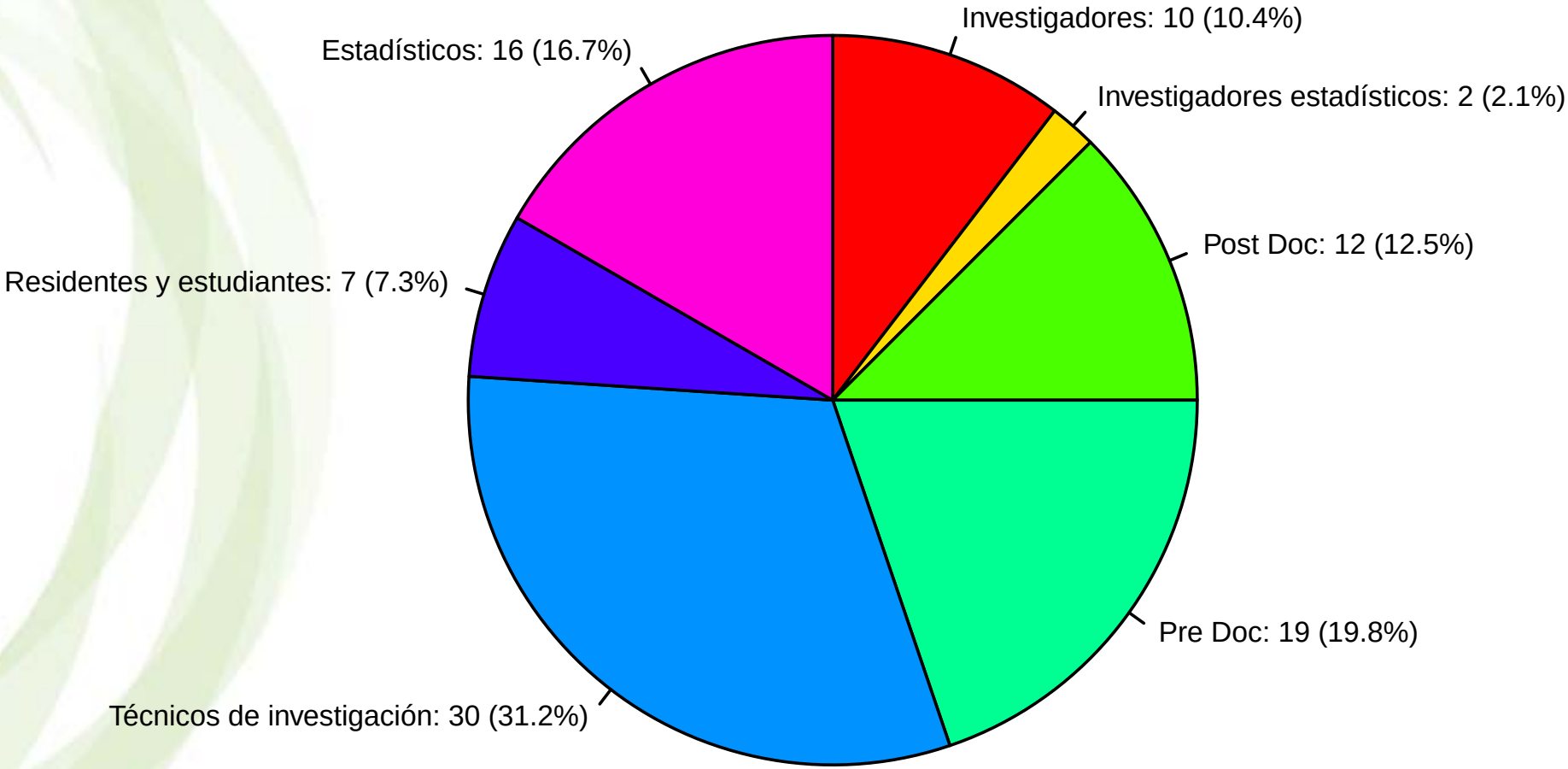
"Identificamos los determinantes ambientales de la salud y promovemos su prevención y control."

El CREAL: 6 programas de investigación

Programa	¿Cómo afecta?	a...
Respiratorio	Factores ambientales, laborales y genéticos	Enfermedades respiratorias (asma y EPOC[1])
Cáncer	Factores ambientales, y genéticos	Vejiga urinaria, mama, colon, leucemia, linfomas,...
Salud infantil	Partículas finas, organoclorados, mercurio,...	Crecimiento (intraut., postnat., inf.) Reproducción, sist. neuroconductual
Contaminación atmosférica	Tráfico,  y otros	Sist. cardiorrespiratorio y del neurodesarrollo
Contaminación del agua	Productos desinfectantes de agua potable y piscinas	Cáncer, enferm. respiratorias, trastornos reproductivos,...
Radiaciones	Radiaciones ionizantes () y no ionizantes ()	Tumores y otros

[1] EPOC: Enfermedad Pulmonar Obstructiva Crónica.

Distribución del personal del CREAL ($n = 96$)



Errores en la valoración de la exposición facial a radiación ultravioleta

Dadvand P, Basagaña X, Barrera-Gómez J, Diffey B, Nieuwenhuijsen M.
Measurement errors in the assessment of exposure to solar ultraviolet radiation and its impact on risk estimates in epidemiological studies.
Photochemical & Photobiological Sciences. 2011, 10, 1161-1168.

Objetivos

- **Simular la exposición facial total anual a rayos ultravioletas** para trabajadores de interior en 6 ciudades europeas, Atenas (37^{o}N 23^{o}E), Grenoble (45^{o}N 5^{o}E), Milán (45^{o}N 9^{o}E), Praga (50^{o}N 14^{o}E), Oxford (52^{o}N 1^{o}W) y Helsinki (60^{o}N 24^{o}E), durante 1997.

- **Evaluar el error** cometido en el impacto sobre riesgos para la salud, **al aproximar la exposición personal por la ambiental**.

Modelo para la simulación

$$E_{id} = UVR_d \cdot EF_{id} \cdot \left[1 - \left(1 - \frac{h_{id}}{H_d}\right)^2\right]$$

donde

- E_{id} (**resultado de la simulación)**:
  exposición facial a *UVR* en el individuo i durante el día d,
- UVR_d (medidas disponibles para el año y ciudades de interés):
  nivel medio ambiental de UVR para el día d,
- EF_{id} (**a simular)**:
  fracción facial de exposición (fracción de UVR ambiental recibida por el individuo i durante el día d).
- h_{id} (**a simular a partir de una muestra aleatoria en cada ciudad en el año de interés)**:
  tiempo bajo exposición para el individuo i durante el día d,
- H_d (medidas disponibles para el año y ciudades de interés):
  tiempo con luz natural en el punto medio del mes de interés y en la latitud de interés.

Simulación de EF_{id} (fracción facial de exposición)

$$E_{id} = UVR_d \cdot EF_{id} \cdot \left[1 - \left(1 - \frac{h_{id}}{H_d}\right)^2\right]$$

- Se asumió $EF_{id} \sim \mathcal{U}(EF_{\text{min}}, EF_{\text{max}})$ (`runif`),
- $EF_{\text{min}} = 0{,}05$,
- EF_{max}:
  - 0,25 (laborable),
  - 0,30 (fin de semana de invierno),
  - 0,40 (fin de semana de verano),
  - 0.50 (vacaciones de verano).

Simulación de h_{id} (tiempo bajo exposición)

$$E_{id} = UVR_d \cdot EF_{id} \cdot \left[1 - \left(1 - \frac{h_{id}}{H_d}\right)^2\right]$$

- Información de partida: muestra aleatoria de la variable para días laborables. Tamaños muestrales desde $N = 79$ (Praga) hasta $N = 418$ (Helsinki).
- Muestras modeladas paramétricamente con diversas distribuciones: LogNormal (`plnorm`), Gamma (`pgamma`), Weibull (`pweibull`), χ^2 (`pchisq`) y LogLogistic (`pfisk` $\in$ `VGAM`), estimando los parámetros por MV (`fitdistr` $\in$ `MASS`) y decidiendo la mejor distribución según su bondad de ajuste mediante el test de Kolmogorov–Smirnov (`ks.test` $\in$ `truncgof`).
- Obtención de los parámetros para la misma distribución con igual varianza pero con moda multiplicada por el factor *WF* (*weekend factor*: 2 (invierno) o 4 (verano)) (`uniroot`). Valoración gráfica de la transformación mediante simulación (`density`).

Simulación de h_{id} (tiempo bajo exposición)

$$E_{id} = UVR_d \cdot EF_{id} \cdot \left[1 - \left(1 - \frac{h_{id}}{H_d}\right)^2\right]$$

- 15 días de vacaciones en un único periodo distribuido uniformemente en julio (Praga, Oxford, Helsinki) o agosto (Atenas, Grenoble, Milán).
- Durante las vacaciones, $h_{id} \sim \mathcal{N}(5, 1)$ (`rnorm`).
- El 10 % no viaja durante las vacaciones (`rbern` $\in$ `Rlab`).
- El 90 % lo hace a determinadas ciudades europeas según una distribución multinomial de parámetros conocidos (`rmultinom`).

Resultados de la simulación de la exposición personal

- Con las condiciones anteriores creamos una función para simular la exposición de un individuo durante todo el año:

```
OneSimulation <- function(city="Oxford",
                          wf=c(4, 2),
                          percHoli=90,
                          Holidays=15)
```

- Salida: matriz con tiempo de exposición y exposición a UVR facial totales anuales, desglosados por tipo de día.

Resultados de la simulación de la exposición personal

- Para cada ciudad, se simularon 10.000 individuos durante cada uno de los días del año:

```
replicate(n=10000, OneSimulation(city=citynames[i], ...)
```

- y se calculó un resumen descriptivo de la exposición total anual desglosada según el tipo de día:

DATOS: Muestra de *h* en laborables + UVR ambiental				**SIMULACIÓN:** Exposición UVR facial anual (mediana y porcentaje de contribución)					
								Fin de semana	
City	***N***	$\bar{h}$ **(sd)**	**UVR (anual)**	**Total**	**Vacaciones**	**Laborables**	**Total**	**Verano**	**Invierno**
Athens	98	1.68 (1.28)	10113	532	89 (16 %)	176 (42 %)	225 (42 %)	184 (35 %)	33 (7 %)
Grenoble	101	1.53 (1.32)	7446	339	70 (19 %)	108 (42 %)	129 (39 %)	99 (31 %)	22 (8 %)
Milan	291	1.24 (0.93)	6941	297	66 (21 %)	91 (38 %)	120 (41 %)	96 (34 %)	20 (7 %)
Prague	79	1.52 (1.15)	5238	254	54 (22 %)	76 (39 %)	98 (39 %)	82 (33 %)	13 (6 %)
Oxford	104	1.67 (1.08)	5003	299	74 (24 %)	87 (38 %)	111 (38 %)	92 (32 %)	15 (6 %)
Helsinki	418	1.58 (1.17)	3673	211	56 (27 %)	58 (35 %)	78 (38 %)	71 (35 %)	6 (3 %)

Relación estadística entre la exposición personal y la ambiental y el tiempo de exposición

- Modelo lineal:

```
lm(logUVPersAnual ~ logTiempoAnual + logUVambientalAnual)
```

- $R^2 = 0{,}40$.

Variabilidad intra-ciudad e inter-ciudad

- Estimación de $R_{0,95} = \frac{pct_{97,5}}{pct_{2,5}}$:

```
require (lme4) # lmer
# Modelo con efecto aleatorio de la ciudad:
mod <- lmer(logUV ~ (1|city), family=gaussian, data=dat)
# Variabilidad intra e inter ciudad:
sdBetweenWithin <- as.numeric(summary(mod)@REmat[, "Std.Dev."])
R95City <- exp(2*1.96*sdBetweenWithin)
names(R95City) <- c("Between", "Within")
R95City
```

- La variabilidad dentro de la ciudad domina a la variabilidad entre ciudades.

$$\frac{R_{0,95} \text{ intra-ciudad}}{R_{0,95} \text{ inter-ciudad}} \approx 3$$

- **Parece no resultar una buena aproximación caracterizar la exposición del individuo por la exposición ambiental de su ciudad.**

Pérdida de potencia al aproximar la exposición personal por la ambiental mediante la simulación de un efecto

- Simulamos una respuesta binaria Y asociada a la exposición personal a partir de un modelo logístico bajo las condiciones siguientes:

```
# Modelo de simulación:
#   logit(Y=1) = beta0 + beta1*log(uv)
# input:
#   p = P(Y=1|UV=median(UV)
#   OR = OR(Q1(UV) -> Q3(UV))
#   logUV = vector anual de log(UV personal simulado)
simulateY <- function(p=0.1, OR=1.5, logUV)
 {
  b1 <- log(OR)/log(Qr)          # Qr = Q3(UV)/Q1(UV)
  b0 <- log(p/(1-p)) - b1*log(UVmedian)
  logitY <- b0 + b1*logUV
  pr <- 1/(1+exp(-logitY))
  Y <- sapply(pr, FUN=function(pr) rbern(n=1, prob=pr))
  Y
 }
```

Pérdida de potencia al aproximar la exposición personal por la ambiental mediante la simulación de un efecto

- Fijamos OR $= 1{,}5$ cuando la exposición pasa del primer al tercer cuartil.
- Fijamos prevalencias en la mediana de la exposición $p = 0{,}1$ y $p = 0{,}001$.
- En cada caso, seleccionamos el número de simulaciones tal que la potencia en el modelo de referencia (usando la exposición simulada como regresora) fuese aproximadamente del 80 %.
- Calculamos el ARE (asymptotic relative efficiency) como medida comparativa de eficiencia:

$$\text{ARE} = \frac{\text{Tamaño muestral en el modelo alternativo}|\text{Potencia}_0}{\text{Tamaño muestral en el modelo de referencia}|\text{Potencia}_0} = \left(\frac{\frac{\beta_{\text{alt}}}{\text{sd}(\beta_{\text{alt}})}}{\frac{\beta_{\text{ref}}}{\text{sd}(\beta_{\text{ref}})}} \right)^2$$

Prevalencia de *Y* en UV mediana	***N* por ciudad**	**Valores del ARE según regresora**			
		UV simul.	**UVR amb.**	**Latitud**	**Tiempo exp.**
1/10	180	1	5.8	6.4	4.6
1/1000	33000	1	5.2	5.5	4.0

- **Bajo las condiciones anteriores, necesitaríamos multiplicar el tamaño muestral por un factor entre 4 y 6 para conservar una potencia del 80 %.**

Estimación del riesgo atribuible bajo diseño de estudio *case-crossover*

Basagaña X, Sartini C, Barrera-Gómez J, Dadvand P, Cunillera J, Ostro B, Sunyer J, Medina-Ramón M.
Heat waves and cause-specific mortality at all-ages
Epidemiology. 2011, 22(6), 765-772.

Objetivos

- **Estimar el efecto del calor extremo sobre la mortalidad** en Catalunya durante la temporada cálida (16 mayo - 15 octubre) desde 1983 hasta 2006 (503.389 muertes).

- Explorar el efecto anterior sobre **adultos y niños** y estratificando por **66 y 8 causas de muerte respectivamente**.

El diseño *case-crossover*

- Puede utilizarse para valorar la asociación entre una exposición y una respuesta aguda y temporalmente muy cercana a la exposición.
- Diseño similar al caso-control.
- Cada caso se emplea también como control.
- Para cada caso, se suelen tomar varios controles (control de posibles tendencias).
- La cercanía temporal entre el caso y sus controles, y la ubicación temporal de estos, controlan posibles confusoras.

El diseño *case-crossover*

		Invariante intra-individuo	Temp., hum. y Lags	Estación meteorológica	Invariantes intra-individuo
Id	**Caso**	**dow–m–y**	$X_1, \ldots, X_p$	**Base**	**Edad, Sexo,..., Causa**
1	0	$d_1 - m_1 - y_1$	$x_{111}, \ldots, x_{p11}$	B_j	$e_1, s_1, \ldots, c_1$
1	1	$d_1 - m_1 - y_1$	$x_{112}, \ldots, x_{p12}$	B_j	$e_1, s_1, \ldots, c_1$
1	0	$d_1 - m_1 - y_1$	$x_{113}, \ldots, x_{p13}$	B_j	$e_1, s_1, \ldots, c_1$
1	0	$d_1 - m_1 - y_1$	$x_{114}, \ldots, x_{p14}$	B_j	$e_1, s_1, \ldots, c_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	$d_n - m_n - y_n$	$x_{1n1}, \ldots, x_{pn1}$	B_k	$e_n, s_n, \ldots, c_n$
n	0	$d_n - m_n - y_n$	$x_{1n2}, \ldots, x_{pn2}$	B_k	$e_n, s_n, \ldots, c_n$
n	0	$d_n - m_n - y_n$	$x_{1n3}, \ldots, x_{pn3}$	B_k	$e_n, s_n, \ldots, c_n$
n	1	$d_n - m_n - y_n$	$x_{1n4}, \ldots, x_{pn4}$	B_k	$e_n, s_n, \ldots, c_n$

El diseño *case-crossover*

- Se definen y calculan los valores de las variables indicadoras de *Hot Day* (HD) y/o de *Heat Wave* (HW).
- Existen diversos criterios. Por ejemplo, uno podría ser:
  Un día se considera HD si su temperatura máxima supera el percentil 95 de la serie histórica de temperaturas máximas en la base meteorológica asociada.

```
require(doBy) # summaryBy
myFun <- function(x) quantile(x, probs=0.95, na.rm=TRUE)
pct95byBase <- summaryBy(Temp ~ Base, data=myData, FUN=myFun)
myData <- merge(myData, pct95byBase, by="Base")
myData$HD <- myData$Temp >= myData$Temp.95%
```

- Se puede hacer intervenir la humedad considerando la temperatura aparente.
- Un día se considera dentro de una HW si es HD y también, por ejemplo, lo son los dos días previos.

El diseño *case-crossover*

Indicadoras (0/1) de calor relativo extremo

Id	Caso	$H_0, \ldots, H_q$
1	0	$h_{011}, \ldots, h_{q11}$
1	1	$h_{012}, \ldots, h_{q12}$
1	0	$h_{013}, \ldots, h_{q13}$
1	0	$h_{014}, \ldots, h_{q14}$
$\vdots$	$\vdots$	$\vdots$
n	0	$h_{0n1}, \ldots, h_{qn1}$
n	0	$h_{0n2}, \ldots, h_{qn2}$
n	0	$h_{0n3}, \ldots, h_{qn3}$
n	1	$h_{0n4}, \ldots, h_{qn4}$

- Por ejemplo, $H_k = \text{Lag}_k(\text{HD}), \quad k = 0, \ldots, q.$

El diseño *case-crossover*

- Modelo de *Conditional Logistic Regression*

```
require(survival)
clogit(Caso ~ H0 + ... + Hq) + strata(Id), data=myData)
```

- Matemáticamente es equivalente a *Conditional Poisson Regression* con variables indicadoras del estrato.
  $\Longrightarrow \boldsymbol{\beta = \log RR}$.

- y podemos interpretar

$$\log \mathrm{RR}_i = \log\left(\frac{O_i}{E_i}\right) = \boldsymbol{H}_i\,\boldsymbol{\beta}^T, \quad i = 1, \ldots, N$$

  donde
  - $\boldsymbol{H}_i = (H_{0i}, \ldots, H_{qi})$ es el patrón de exposición asociado al día i,
  - $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)$ es el vector de parámetros asociados a $\boldsymbol{H}$ en el modelo,
  - O_i es el número de muertes observadas en el día i,
  - E_i es el número de muertes esperadas en el día i, si éste está asociado a exposición nula ($\boldsymbol{H}_i = \mathbf{0}$).
  - N es el número de días con muertes en la serie temporal.

Mortalidad atribuible al calor extremo

- La mortalidad atribuible al calor extremo es,

$$
\begin{aligned}
\text{MA} &= \sum_{i=1}^{\#\text{días}} (O_i - E_i) \\
&= \sum_{j=1}^{\#\text{patrones } E} (O_j - E_j) \\
&= \sum_{j=1}^{\#\text{patrones } E} O_j \left(1 - \frac{1}{\text{RR}_j}\right) \\
&= M \sum_{j=1}^{\#\text{patrones } E} P(\text{caso}|E_j) \left(1 - \frac{1}{\text{RR}_j}\right)
\end{aligned}
$$

donde M es la mortalidad total observada.

Identificación de los patrones de exposición

- 2^{q+1} patrones posibles de exposición:
  - $E_0 = \{H_0 = H_1 = \cdots = H_{q-1} = H_q = 0\}$ (exposición nula),
  - $E_1 = \{H_0 = H_1 = \cdots = H_{q-1} = 0, H_q = 1\}$,
  - $\cdots$,
  - $E_{2^{q+1}} = \{H_0 = H_1 = \cdots = H_{q-1} = H_q = 1\}$.

- Posible etiquetado de los patrones:

```
E <- as.matrix(H) %*% (10^(q:0))
```

```
  H0 H1 H2 H3 H4 H5   E
1  0  0  0  0  0  0   0
2  0  0  0  0  0  1   1
3  0  0  0  0  1  0  10
4  0  0  0  1  0  0 100
```

Prevalencia de casos y de patrones de exposición

- Estimamos la prevalencia de cada escenario de exposición, $P(E_j)$:

```
pE <- summary(as.factor(E))
pE <- pE/sum(pE)
```

- y la prevalencia de casos observada en cada escenario

```
require(doBy)
casosByE <- summaryBy(caso ~ E, FUN=sum, ...)
```

- y la tabla de prevalencias de E:

```
# Patrones de E existentes:
duplicated(...)
# Fusión de datos:
merge(data1, data2, by="E")
# Orden por prevalencia:
myData <- myData[order(myData$pE, decreasing=TRUE), ]
```

```
     E   H0 H1 H2 H3 H4 H5     pE     pCaso
1      0  0  0  0  0  0  0 0.7294 0.7410
2    100  0  0  0  1  0  0 0.0403 0.0392
3 100000  1  0  0  0  0  0 0.0402 0.0404
```

Riesgo relativo y mortalidad atribuible

```
Hm <- as.matrix(H)
L <- as.vector(Hm %*% betas)
varL <- diag(Hm %*% covBetas %*% t(Hm))
signError <- matrix(-1:1, nrow=nrow(Hm), ncol=3, byrow=TRUE)
RR <- exp(L + signError*1.96*sqrt(varL))
MAbyE <- M*pCaso*(1 - 1/RR[, 2])
MA <- sum(MAbyE)
```

```
      E H0 H1 H2 H3 H4 H5     pE  casos RRlo95    RR  RRup95    MA
1     0  0  0  0  0  0  0 0.7293 1843  1.000 1.000  1.000     0
2 10000  0  1  0  0  0  0 0.0409  106  1.025 1.054  1.084 5.449
3  1000  0  0  1  0  0  0 0.0407  104  1.012 1.042  1.074 4.320
```

Algunos resultados

- 3 días consecutivos de calor extremo incrementan la mortalidad total diaria en un 19 %.
- 1,69 % de muertes atribuídas al calor (333 muertes anuales en temporada cálida).
- $\approx$ 40 % de esas muertes no ocurrió durante una ola de calor.
- RR más elevados: enfermedades cardiovasculares y respiratorias, desórdenes mentales y del aparato nervioso, algunas infecciosas, aparato digestivo, diabetes, algunas causas externas incluyendo el suicidio.
- En infantes, el efecto se observó en el mismo día y sólo para condiciones originadas en el período perinatal (RR 1,53 (1,16 - 2,02)).

Imputación múltiple en análisis de conglomeración

Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. **Multiple imputation in cluster analysis** (in preparation).

Objetivo

- Proponer un **procedimiento** para trasladar la **incertidumbre debido a datos faltantes** a los resultados de un **análisis de conglomeración**.

Imputación múltiple

- Multivariate Imputations by Chained Equations (MICE):

```
# Para M imputaciones:
MI <- mice(data, m = M)
# Lista con las imputaciones:
MI$imp
```

Datos centrados (y estandarizados)

- En cada imputación, se centran todas las variables en su media:

```
CenterDataBase <- function(DataBase)
 {
  n <- nrow(DataBase)
  m <- apply(DataBase, 2, mean)
  CenteredDataBase <- DataBase - rep(1, n)%*%t(m)
  CenteredDataBase
 }
```

Datos centrados (y estandarizados)

- ... y se estandarizan por su desviación estándar las contínuas:

```
StandardizeDataBaseContinuous <- function(DataBase)
 {
  Xs <- as.matrix(CenterDataBase(DataBase))
  p <- ncol(DataBase)
  for (i in 1:p)
   {
    x <- Xs[, i]
    # is x binary?
    if (length(levels(factor(x))) > 2)
      Xs[, i] <- x/sd(x, na.rm=TRUE)
   }
  Xs <- as.data.frame(Xs)
  names(Xs) <- names(DataBase)
  Xs
 }
```

Algoritmo de conglomeración

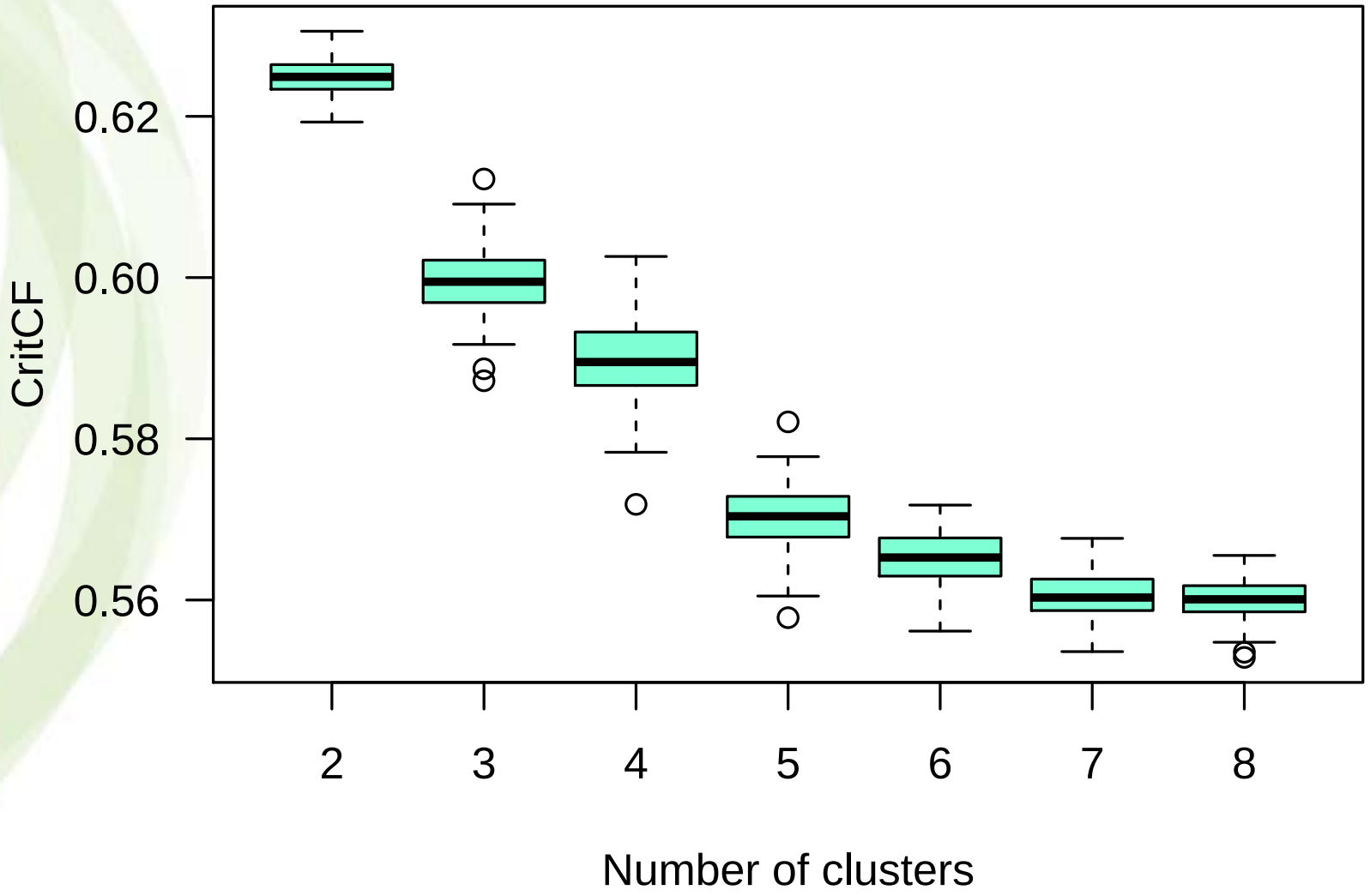
- Se prefijan los valores posibles para el número de clusters, $k \in (2, 3, \ldots, k_{\max})$.
- Para cada imputación y para cada k posible:
  - Se realiza una conglomeración *k-means* (`kcca` $\in$ `flexclust`) partiendo con todas las variables y usando como centroides iniciales los obtenidos por un cluster jerárquico (`hclust` $\in$ `flexclust`), cortado a k grupos (`cutree`),
  - Se elimina aquella variable que minimiza el valor de

$$CritCF = \left[\left(1 + \frac{1}{2m}\right)\left(1 + \frac{W}{B}\right)\right]^{-\frac{1+\log_2(k+1)}{1+\log_2(m+1)}}$$

    donde m es el número de variables y W y B son las inercias intra-cluster e inter-cluster respectivamente.
  - Se continua el procedimiento de eliminación de variables una a una hasta que la eliminación de una variable no mejora el valor de *CritCF*.
  - Se fija, para esa imputación, el valor de k y el conjunto de variables seleccionadas asociados al valor màximo de *CritCF*.
- Ahora tenemos, para cada una de las M imputaciones, un valor óptimo de k y un conjunto de variables conglomeradoras.

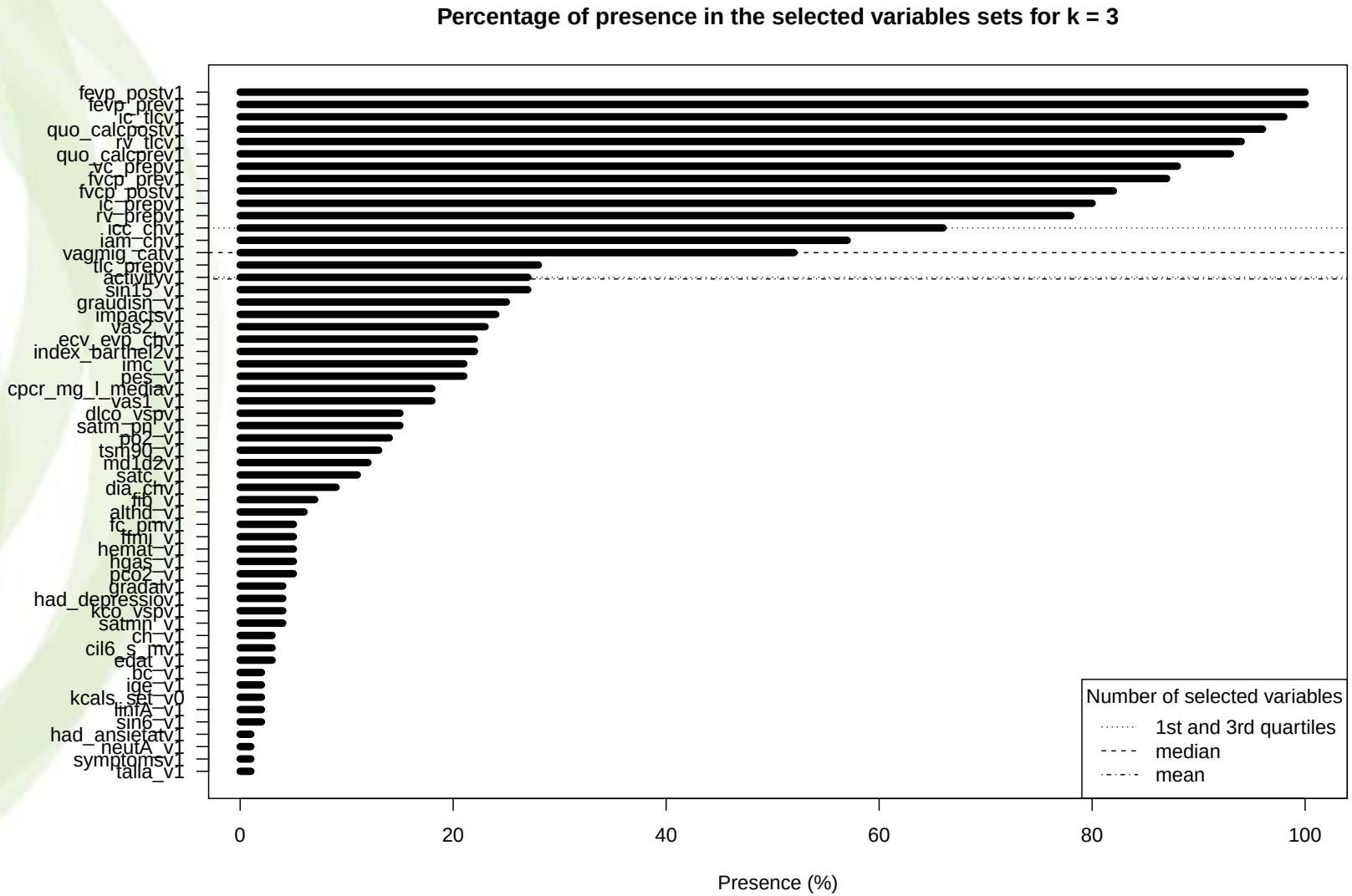
Integración de la imputación múltiple en el análisis de conglomeración

- Podemos decidir como valor óptimo de *k* el asociado al valor máximo de *CritCF* o bien aquél que queda seleccionado en más imputaciones.



Integración de la imputación múltiple en el análisis de conglomeración

- Una vez decidido el valor de k, realizamos un análisis descriptivo de la presencia de variables seleccionadas:



Integración de la imputación múltiple en el análisis de conglomeración

- Reetiquetamos los clusters:

```
RelabelClusters <- function(RefCenters, Centers, Cluster)
 {
  res <- NULL
  k <- dim(RefCenters)[1]
  n <- length(Cluster)
  permut <- permn(1:k)
  nPermut <- length(permut)
  distances <- rep(NA, nPermut)
  for (i in 1:nPermut)
   {
    id <- permut[[i]]
    auxCenters <- Centers[id, ]
    distances[i] <- sum((auxCenters - RefCenters)^2)
   }
  id <- which(distances == min(distances))[1]
  order <- permut[[id]]
  newCenters <- Centers[order, ]
  res$newCenters <- newCenters
  newCluster <- rep(NA, n)
  for (i in 1:k)
   newCluster[Cluster == i] <- order[i]
  res$newCluster <- newCluster
  res
 }
```

Integración de la imputación múltiple en el análisis de conglomeración

- Estimamos las probabilidades de asignación a cada cluster:

Id	*P*(Cluster 1)	*P*(Cluster 2)
1	0,92	0,08
2	0,93	0,07
3	0,02	0,98
⋮	⋮	⋮

Integración de la imputación múltiple en el análisis de conglomeración

- Y resumimos la distribución de estas probabilidades:

	Mínimo	Cuartil 1	Mediana	Cuartil 3	Máximo
Cluster 1	0,64	1	1	1	1
Cluster 2	0,57	1	1	1	1

Cómo lucrarse con R: Concurso del "Caganer" en el CREAL

Objetivo

- **Crear un "Caganer" ecológico** (con material reciclado y/o reutilizable) para competir **por una cesta navideña**.

Funciones de R usadas

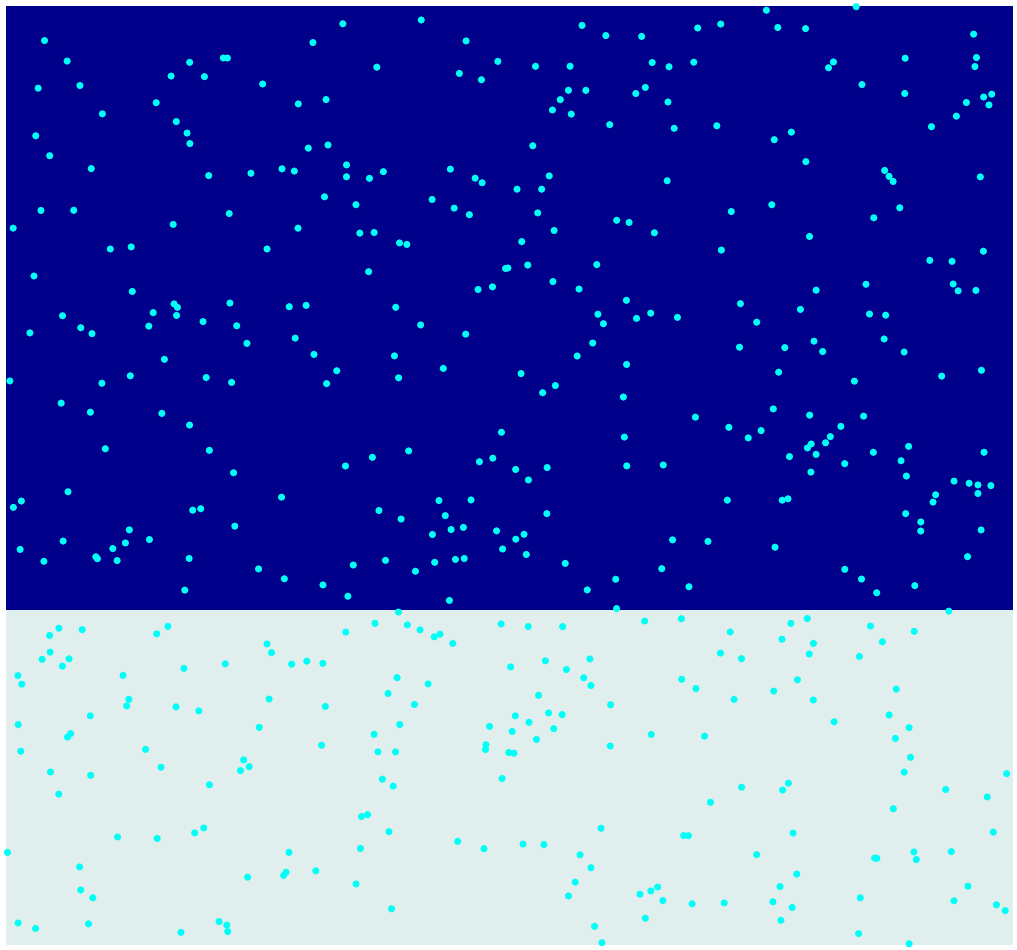
- Aleatorias: `runif` y `rnorm`
- Matemáticas: `abs`, `sin` y `cos`
- Para dibujar: `plot`, `lines`, `polygon` y `rainbow`

CREAL (CaganeR Ecològic i ALeatoritzat)

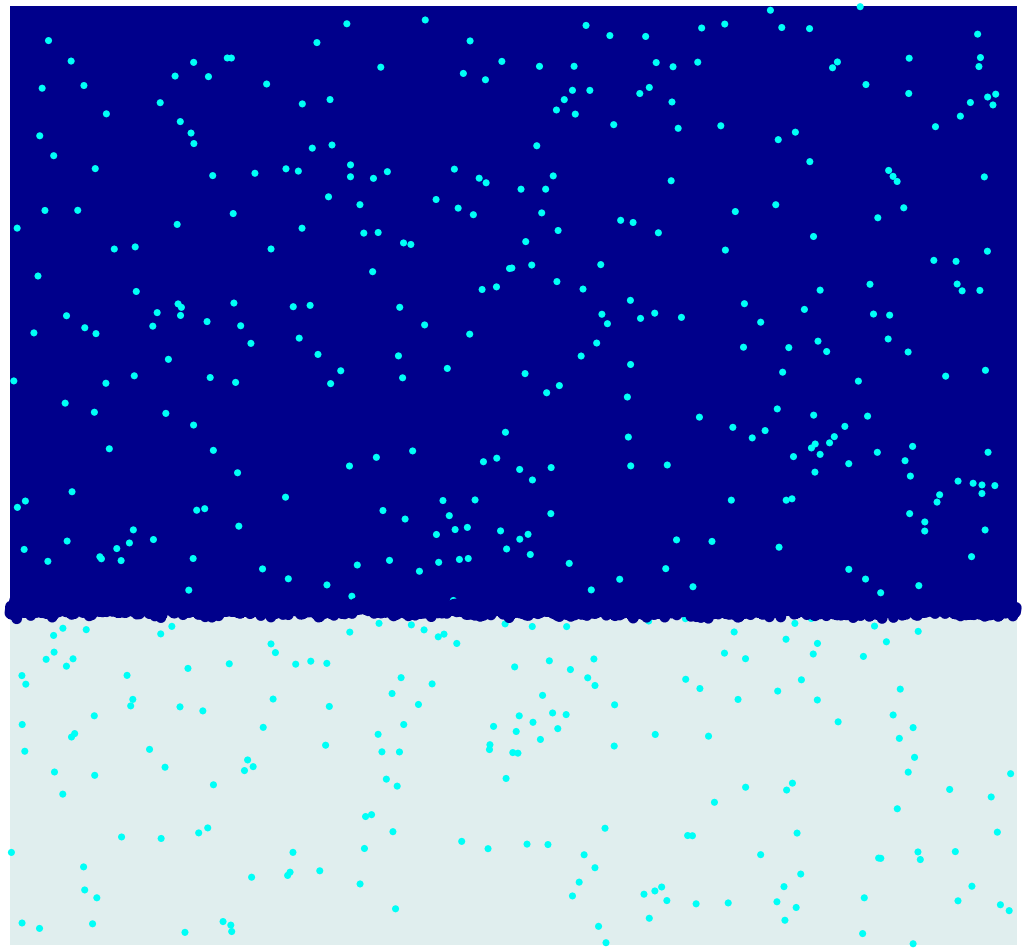
CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

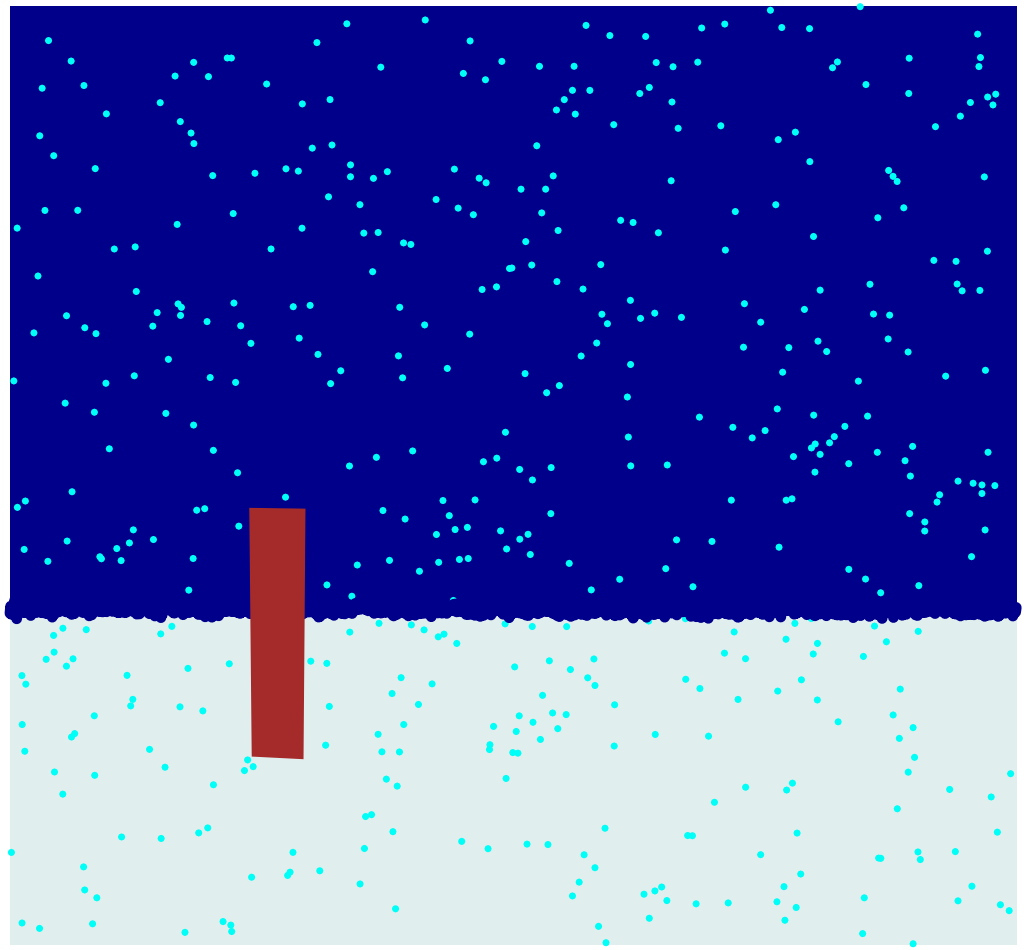
CREAL (CaganeR Ecològic i ALeatoritzat)



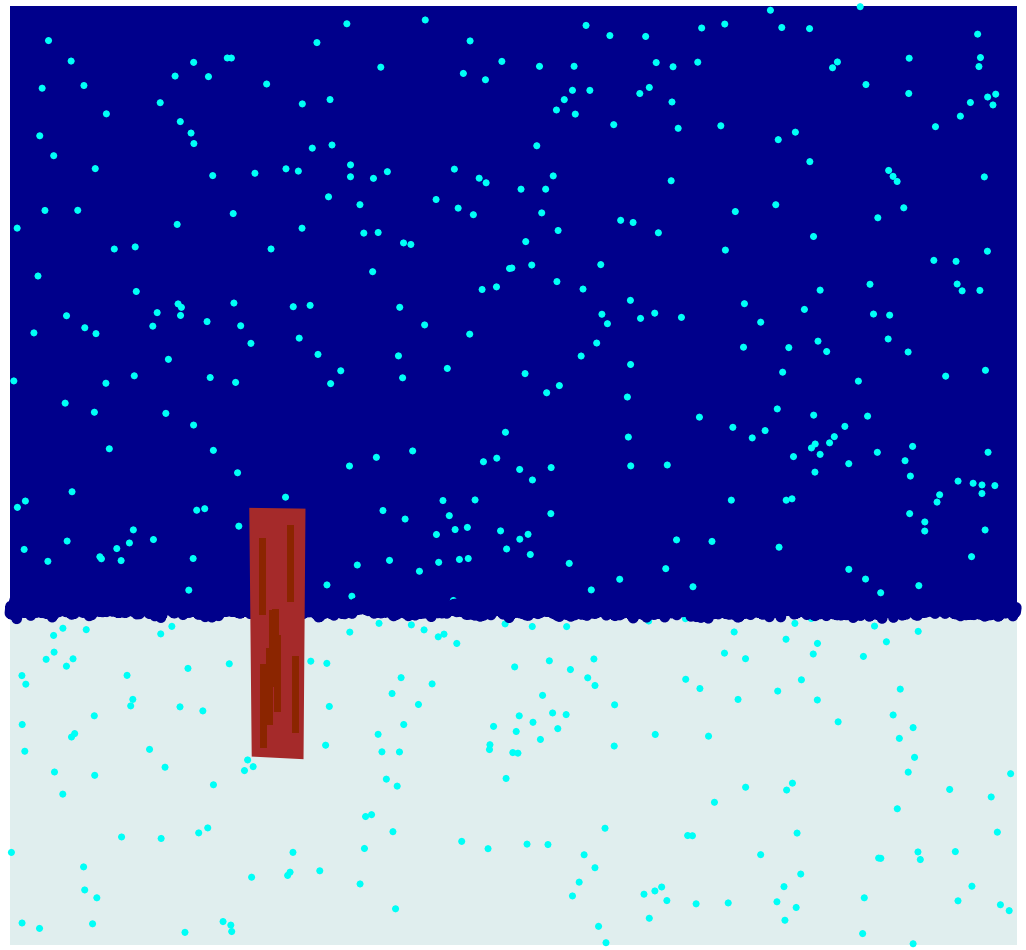
CREAL (CaganeR Ecològic i ALeatoritzat)



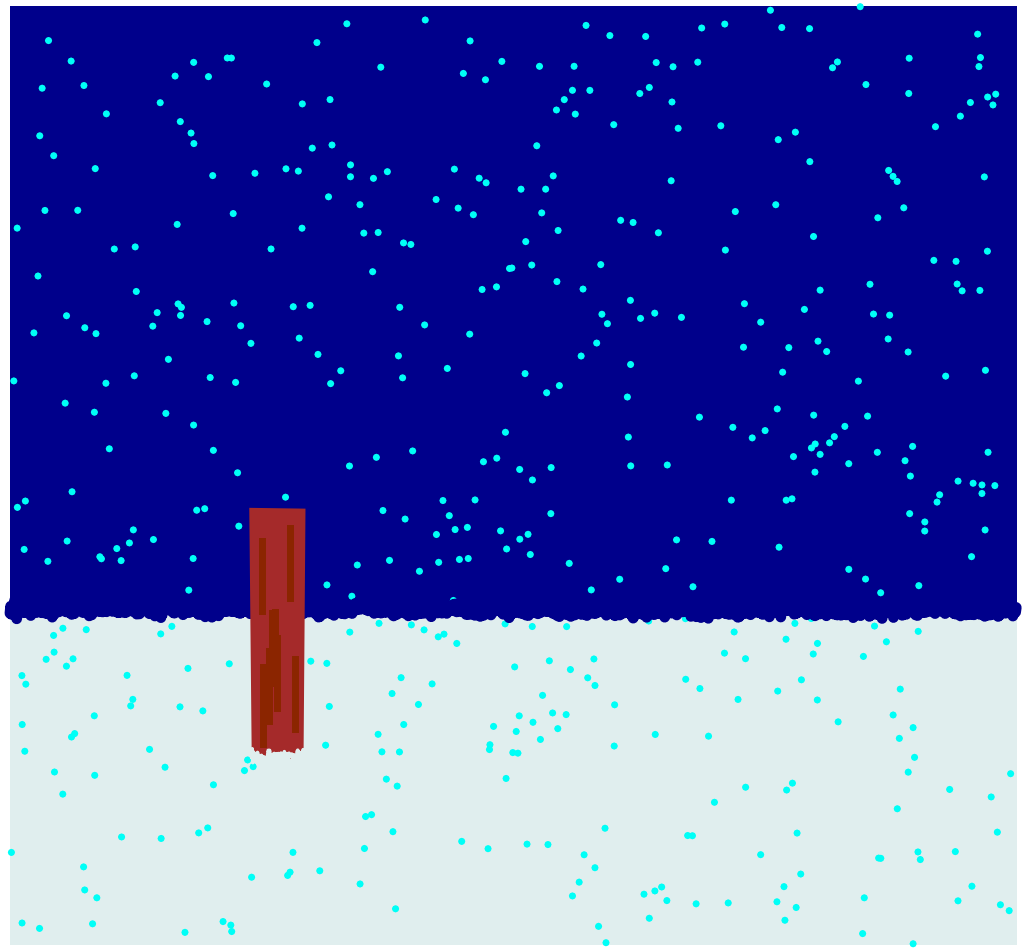
CREAL (CaganeR Ecològic i ALeatoritzat)



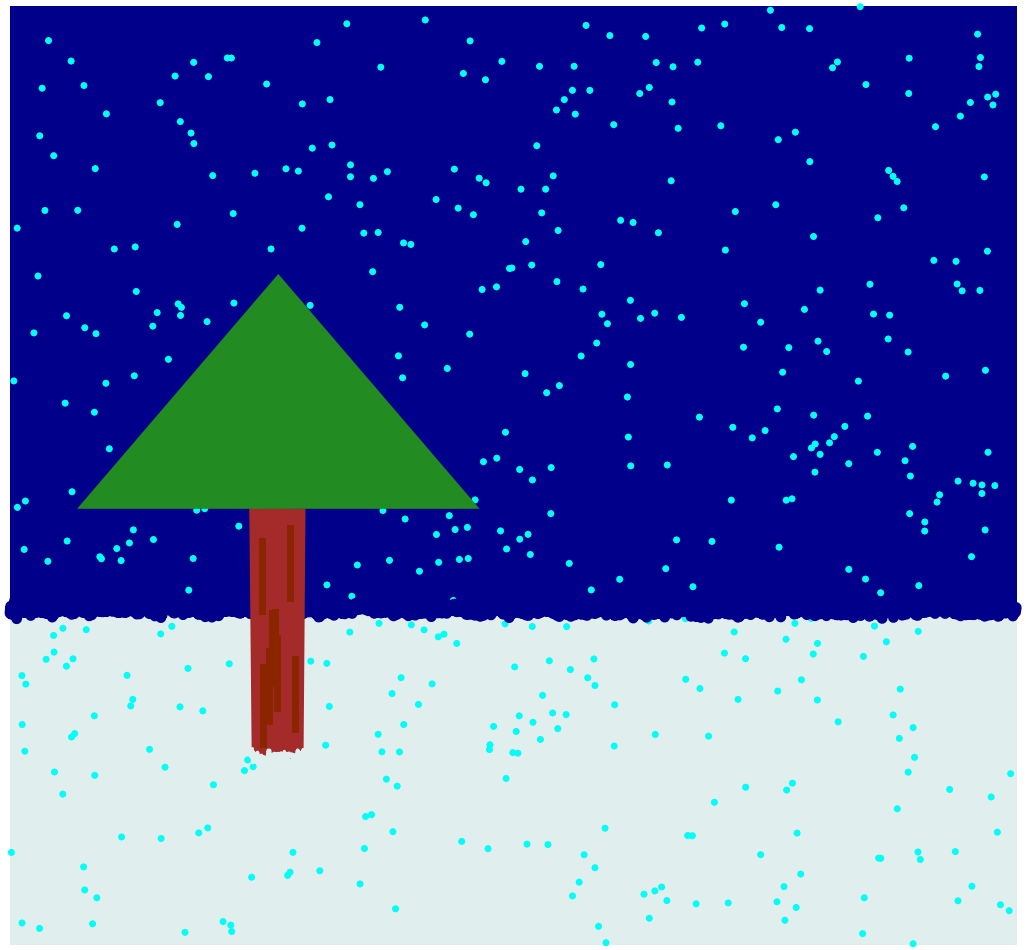
CREAL (CaganeR Ecològic i ALeatoritzat)



CREAL (CaganeR Ecològic i ALeatoritzat)



CREAL (CaganeR Ecològic i ALeatoritzat)



CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

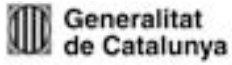
CREAL (CaganeR Ecològic i ALeatoritzat)

CREAL (CaganeR Ecològic i ALeatoritzat)

“L’estadística és sentit comú en un 80 %”
Llorenç Badiella

Gràcies i Felicitats, SEA!

Centre for Research
in Environmental
Epidemiology



Parc de Recerca Biomèdica de Barcelona
Doctor Aiguader, 88
08003 Barcelona (Spain)
Tel. (+34) 93 214 70 00
Fax (+34) 93 214 73 02

info@creal.cat
www.creal.cat