INSTITUTO GALEGO DE ESTATÍSTICA

Utilización do software R nunha oficina de estatística pública

Instituto Galego de Estatística

María Martín Vila

ANO INTERNACIONAL DA ESTATÍSTICA
ORGANIZACIÓN PARTICIPANTE

XUNTA DE GALICIA

Que facemos?

O Instituto Galego de Estatística (IGE) é un organismo autónomo da Xunta de Galicia creado no ano 1988 e que se rexe basicamente pola Lei 9/1988 de Estatística de Galicia

- recompilación e difusión da documentación estatística dispoñible
- desenvolver bases de datos de interese público
- analizar as necesidades e a evolución da demanda de estatísticas e asegurar a súa difusión

INSTITUTO GALEGO DE ESTATÍSTICA

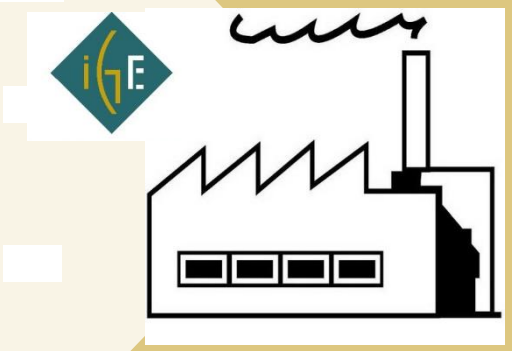
Como o facemos?

Instituto Nacional de Estadística

XUNTA DE GALICIA

eurostat

Outros: dgt, seg-social,…



Lectura de datos

Manipulación

Difusión

INSTITUTO GALEGO DE ESTATÍSTICA

Lectura de datos doutros organismos (ancho fixo):

read.fwf(file, widths, header = FALSE, sep ="\t", skip = 0, row.names, col.names, n = -1,buffersize = 2000, ...)

- Microdatos do INE: Encuesta de Estructura Salarial

```
datos<-read.fwf("EES10_WEB",widths=c(8,2,1,2,1,1,1,1,1,1,2,1,1,2,2,1
,1,2,2,2,2,2,4,2,2,3,2,1,2,1,2,9,9,9,9,9,9,6,7,2,2,1,3,1,3,1,3,1,3,10,9,9,2,12))
```

- Ficheiros txt da dgt: matriculacións mensuais

```
matriculacions<-read.fwf("MATM092013.txt",widths=c(8,1,1,1,2,5,2,8,1 ,1,8))
```

Lectura de datos de outros organismos (bases de datos):

- Instalar paquete RODBC: install.packages("RODBC")
- Crear unha conexion odbc
- Conectar: bd <- odbcConnect(“basededatos")
- Podemos extraer unha táboa completa: sqlFetch
- Ou unha consulta en sql

➢ Datos de contratos en acces:

```
ContratosSector<- sqlFetch(Contratos,"CONTRATOS_TIPOLOXIA_SECTOR")
```

➢ Datos de paro rexistrado en sql:

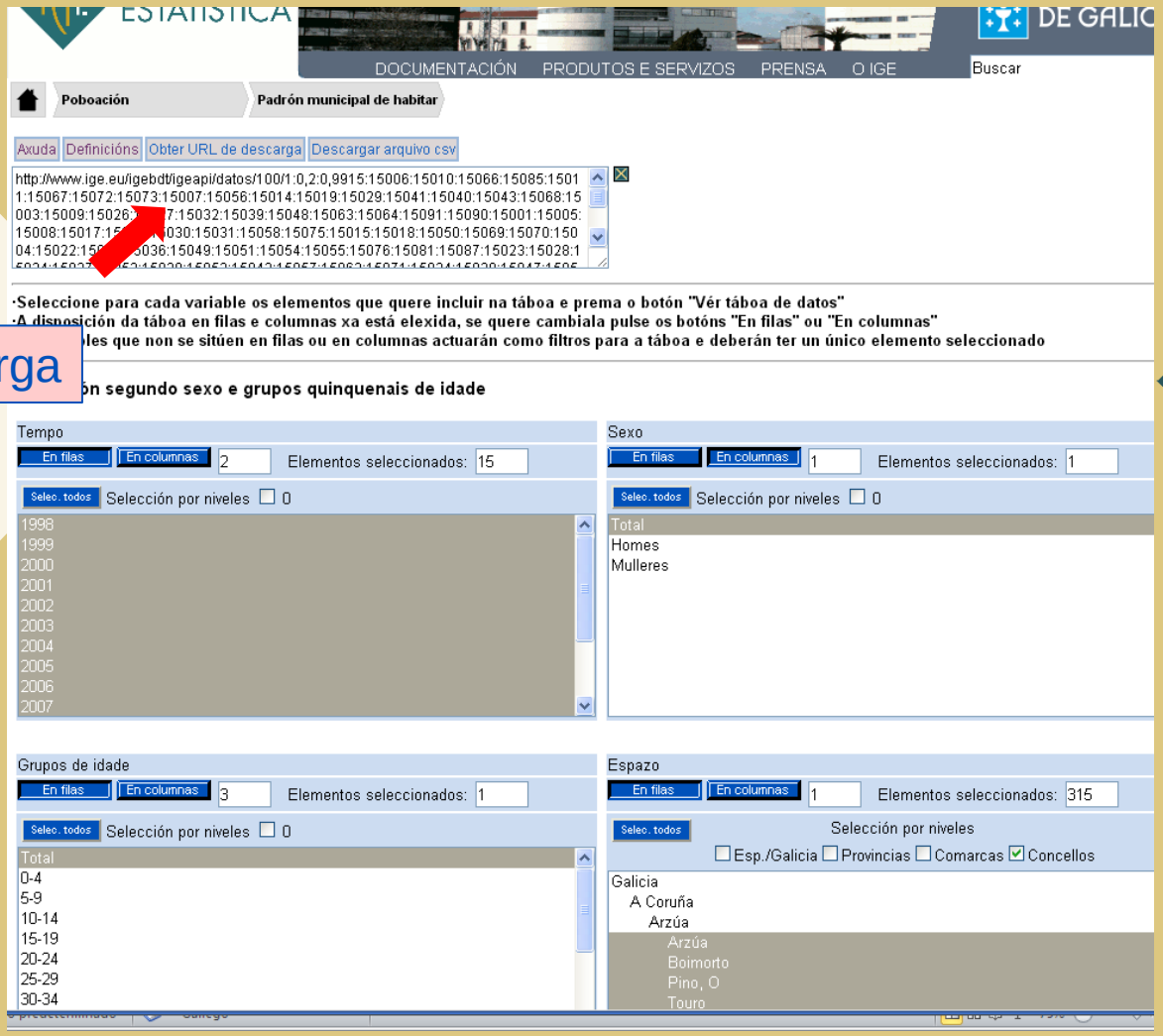
```
Paro_mes<-sqlQuery(fontes, "select ANO, MES, MUN_RES_PFIS, SEXO,
ACT_ECONOMICA, PARADOS from prx200501_ where ANO=2013 and MES=9")
```

Lectura de ficheiros web IGE:

Podemos ler datos directamente da web do IGE coa función: read.csv()

- Poboación :

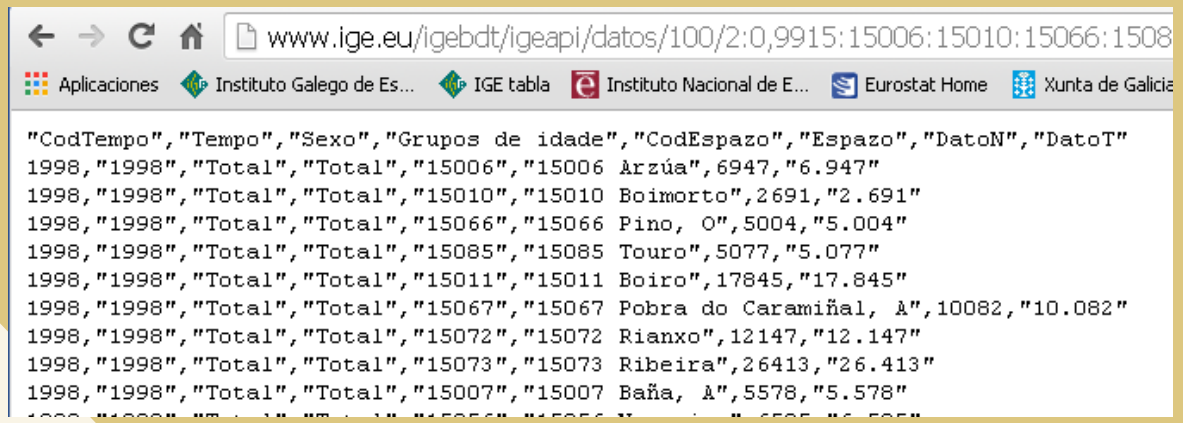
Obter URL de descarga



INSTITUTO GALEGO DE ESTATÍSTICA

Lectura de ficheiros web IGE:

Que hai nesta URL?



Lectura desde R:

poboacion<-read.csv(file="http://www.ige.eu/igebdt/igeapi/datos/100/2:0,9915:15006:15010:15066:15085:15011:15067:...")

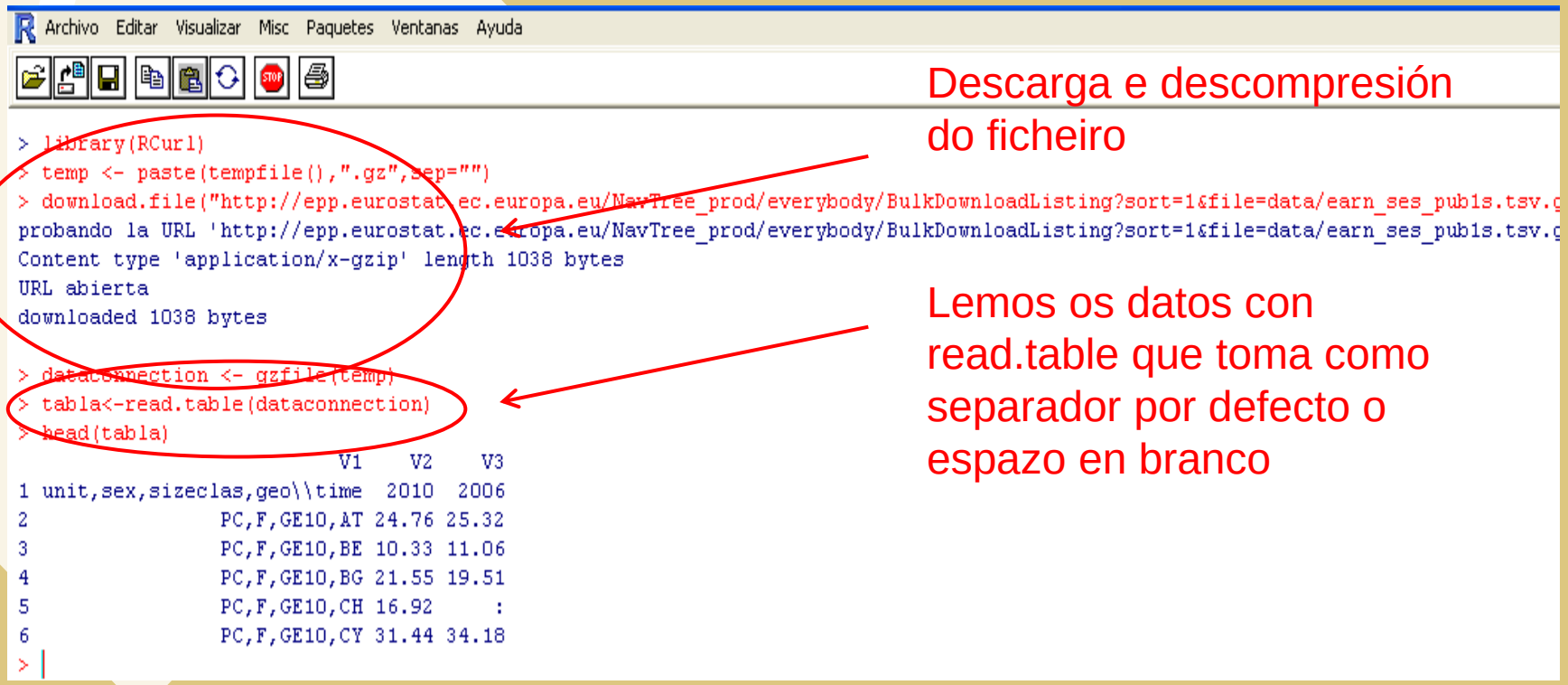
```
head(poboacion)
CodTempo Tempo  Sexo Grupos.de.idade CodEspazo                    Espazo DatoN  DatoT
    1998  1998 Total           Total     15006                15006 Arzúa  6947  6.947
    1998  1998 Total           Total     15010             15010 Boimorto  2691  2.691
    1998  1998 Total           Total     15066              15066 Pino, O  5004  5.004
    1998  1998 Total           Total     15085               15085 Touro  5077  5.077
    1998  1998 Total           Total     15011               15011 Boiro 17845 17.845
    1998  1998 Total           Total     15067 15067 Pobra do Caramiñal, A 10082 10.082
```

INSTITUTO GALEGO DE ESTATÍSTICA

Lectura de ficheiros Eurostat (.tsv comprimidos):

Proporción de asalariados con salarios baixos: earn_ses_pub1s.tsv.gz

Lectura directa:



INSTITUTO GALEGO DE ESTATÍSTICA

Lectura de ficheiros Eurostat (.tsv):

Lectura mediante a función LeeEurostat:

```
> leeEurostat<-function(file,nct){
+
baseurl="http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?sort=1&file=data/"
+ fullfilename=paste(file,".tsv.gz",sep="")
+ temp <- paste(tempfile(),".gz",sep="")
+ download.file(paste(baseurl,fullfilename,sep=""),temp)
+ dataconnection <- gzfile(temp)
+
+ pre<-""
+ for(i in 1:nct) {pre<-paste(sep="",pre,".*?\t")}
+ x<-readLines(dataconnection)
+ x<-gsub(",","\t",x,fixed=T)
+ limpia<-function(a){
+   inicio<-gsub(paste(sep="","(",pre,")",".*") , "\\1",a)
+   final<-gsub("[abcdefghijklmnopqrstuvwxyz]","",  gsub(paste(sep="",pre,"(.*)"),"\\1",a) )
+   return(paste(sep="",inicio,final))
+ }
+ x[2:length(x)]<-limpia(x[2:length(x)])
+
+ f<-tempfile()
+ write.table(x,f,quote=F,sep="\t",row.names=F,col.names=F)
+ x<-read.delim(f,na.strings = ": ")
+ for( i in (nct+1):length(x)) print(class(x[[i]])=="numeric" )
+
+ return(x)
+ }
```

Descarga e descompresión do ficheiro

Separación de variables e eliminación de caracteres intermedios

Saída de datos en formato dataframe coas variables correspondentes e os anos en columnas

INSTITUTO GALEGO DE ESTATÍSTICA

Lectura de ficheiros Eurostat (.tsv comprimidos):

R Console

```
> library(RCurl)
> source("R:/Servicios/Difusion/OPERACIÓNS ESTATISTICAS/ACTIVIDADES-PUBLICACIONES/ESTUDO SALARIOS/ESTRUCTURA S
> file="earn_ses_pub1s"
> nct=4
> datos<-leeEurostat(file,nct)
probando la URL 'http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?sort=1&file=data/
Content type 'application/x-gzip' length 1038 bytes
URL abierta
downloaded 1038 bytes

[1] TRUE
[1] TRUE
> head(datos)
  unit sex sizeclas geo.time X2010 X2006
1   PC   F      GE10       AT 24.76 25.32
2   PC   F      GE10       BE 10.33 11.06
3   PC   F      GE10       BG 21.55 19.51
4   PC   F      GE10       CH 16.92    NA
5   PC   F      GE10       CY 31.44 34.18
6   PC   F      GE10       CZ 24.53 25.08
>
```

Saída de datos en formato dataframe coas variables correspondentes e os anos en columnas

O ano non é unha variable → No proceso de manipulación transformaremos este dataframe nun coa estrutura :

```
head(datos.hip)
  Tempo  Sexo                        País  Dato
   2010 Total Unión Europea (27 países) 16,96
   2010 Total                  Bélgica  6,37
```

INSTITUTO GALEGO DE ESTATÍSTICA

Manipulación de datos:

Empresas por 1.000 habitantes segundo o grao de urbanización

Poboación por concello (Web IGE)

```
> head(poboacion)
  Tempo CodEspazo DatoN
1  1998     15006  6947
2  1998     15010  2691
3  1998     15066  5004
```

Número de empresas por concello (Web IGE)

```
> head(empresas)
  Tempo CodEspazo DatoN
1  1999     15006   406
2  1999     15010   107
3  1999     15066   198
```

Clasificación dos concellos segundo o grao de urbanización (csv a partir de excel)

```
head(clas)
CodEspazo    Espazo Grao
    15001 Abegondo  ZPP
    15002      Ames  ZIP
    15003   Aranga  ZPP
```

Único ficheiro R: datos

MERGE

pobemp<-merge(empresas,poboacion,by=c("CodEspazo","Tempo"),all.x=TRUE)

datos<-merge(pobemp,clas,by.x="CodEspazo",by.y="Código")

INSTITUTO GALEGO DE ESTATÍSTICA

Manipulación de datos:

Datos:

```
> head(datos)
  Tempo CodEspazo   Espazo Grao Poboacion Empresas
1  1999     15001 Abegondo  ZPP       263     5480
2  2000     15001 Abegondo  ZPP       269     5594
3  2001     15001 Abegondo  ZPP       286     5694
4  2002     15001 Abegondo  ZPP       287     5772
5  2003     15001 Abegondo  ZPP       297     5761
6  2004     15001 Abegondo  ZPP       319     5732
>
```

Agregamos os datos de poboación e empresas:

AGGREGATE

```
datosg<-aggregate(datos[,c(5,6)],by=list(datos$Tempo,datos$Grao),FUN="sum")
```

Creamos unha táboa base con todos os cruces das variables e a completamos cos datos:

EXPAND.GRID

```
tablabase<-expand.grid(datosg$Tempo, datosg$Grao)

tabla<-merge(tablabase,datosg, by=c("Tempo","Grao“), all.x=TRUE)
```

Difusión de datos:

O noso dataframe final ten a seguinte estrutura:

```
> head(tabla)
  Tempo  Grao Poboacion Empresas Indicador
1  1999 Total    160988  2730336  16959.87
2  1999   ZDP     61055   913016  14953.99
3  1999   ZIP     49437   810216  16388.86
4  1999   ZPP     50496  1007104  19944.23
5  2000 Total    166055  2731900  16451.78
6  2000   ZDP     62694   914364  14584.55
>
```

Para grabalo na nosa base de datos de Mysql e visualizalo na web utilizamos unha función :

grabarH(5164,tabla,12)

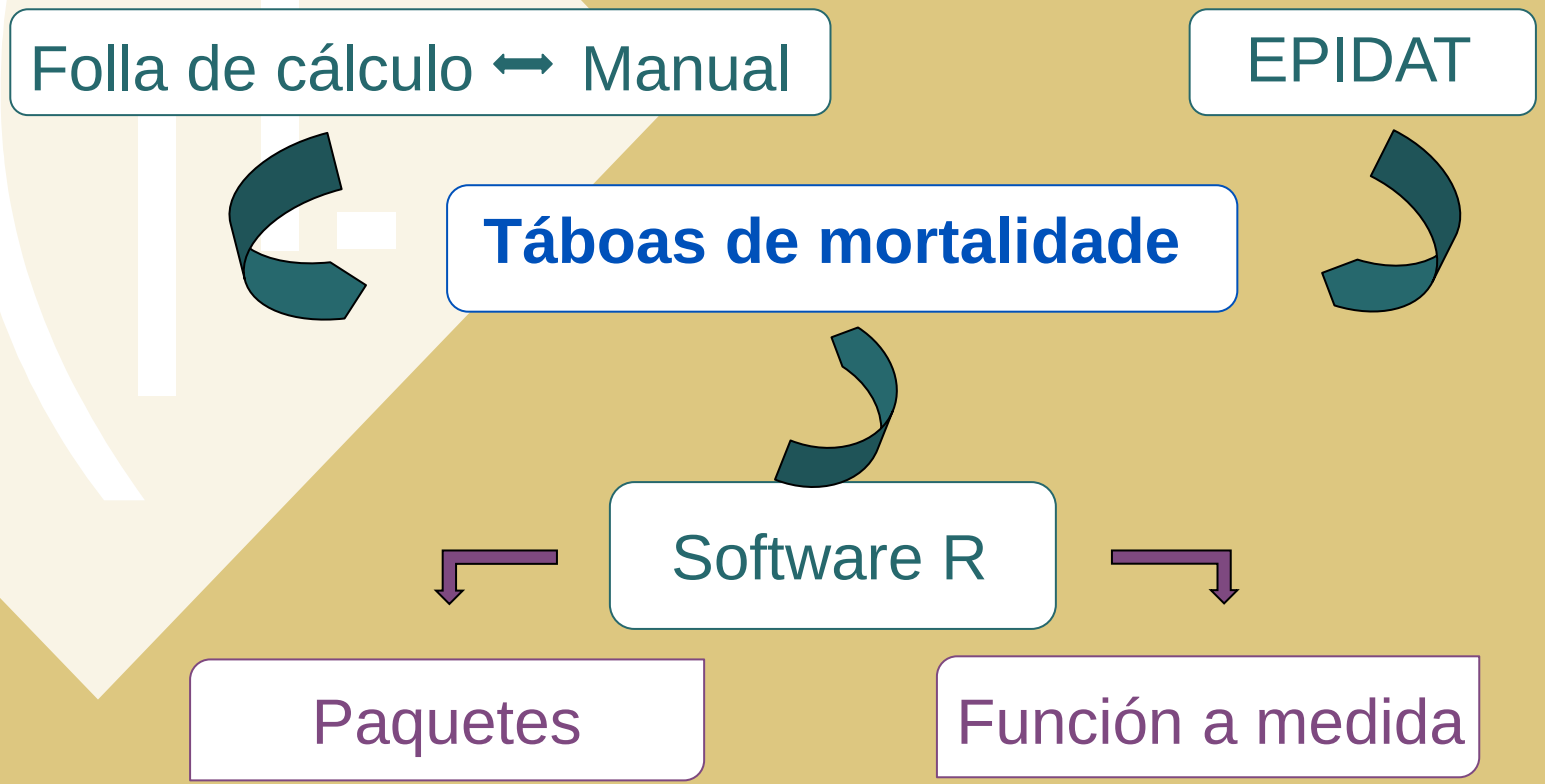
Esta función permítenos grabar unha táboa nova ou añadir datos a unha táboa xa creada.

Só temos que indicarlle o número da táboa multidimensional, o dataframe cos datos e un código identificativo do espazo.

Aplicación: Táboas de mortalidade

Necesitamos calcular táboas de mortalidade simultáneas para:

- diferentes periodos de tempo
- máis dun espacio xeográfico
- todas as variantes do sexo (homes, mulleres e total)



INSTITUTO GALEGO DE ESTATÍSTICA

Por que unha función nova?

Software	Periodos	Espazos xeográficos	Sexo	a_x	TMI
EPIDAT	X	X	✓	X	✓
R Demography	✓	X	X	X	X
R Lifetables	X	X	X	✓	X
R T_Mortalidad	✓	✓	✓	✓	✓

INSTITUTO GALEGO DE ESTATÍSTICA

Función R: T_mortalidad.R

T_mortalidad <- function (data , IM_i ,TMI_i , a_ini , int_f))

Argumentos obrigatorios:

- Defuncións e poboación por idade e sexo
- Espazo (tipo)
- Índice de masculinidade

Argumentos optativos:

- Taxa de mortalidade infantil
- Valores a_x
- Intervalo aberto final

T_mortalidad <- function (data , IM_i)

T_mortalidad <- function (data, IM_i, TMI_i, a_ini, int_f)

INSTITUTO GALEGO DE ESTATÍSTICA

T_mortalidad: Resultado

Saida formato lista → Cada elemento da lista é unha táboa de mortalidade

```
> TM<-T_mortalidad(dataq,datosIM)
> names(TM)[31]
[1] "LT Area= GA Time= 11 Sex= 0"
> TM[[31]]
       x  n          a     D      P           m           q          p        l          d        L         T         e
 [1,]  0  1 0.05000931   165  63061 2.616514e-03 0.0026100265 0.9973900 100000.00  261.00265  99752.05 8213181.8 82.131818
 [2,]  1  4 1.60613024    49 269104 1.820857e-04 0.0007280256 0.9992720  99739.00   72.61254 398782.16 8113429.7 81.346614
 [3,]  5  5 2.50000000    30 320617 9.356959e-05 0.0004677385 0.9995323  99666.38   46.61781 498215.38 7714647.6 77.404710
 [4,] 10  5 2.50000000    34 310809 1.093919e-04 0.0005468102 0.9994532  99619.77   54.47310 497962.65 7216432.2 72.439762
 [5,] 15  5 2.50000000   109 359315 3.033550e-04 0.0015156255 0.9984844  99565.29  150.90370 497449.21 6718469.5 67.478026
 [6,] 20  5 2.50000000   180 431242 4.173990e-04 0.0020848197 0.9979152  99414.39  207.26108 496553.80 6221020.3 62.576658
 [7,] 25  5 2.50000000   231 562106 4.109545e-04 0.0020526637 0.9979473  99207.13  203.63887 495526.55 5724466.5 57.702169
 [8,] 30  5 2.50000000   355 684710 5.184677e-04 0.0025889826 0.9974110  99003.49  256.31831 494376.66 5228940.0 52.815714
 [9,] 35  5 2.50000000   555 665859 8.335098e-04 0.0041588828 0.9958411  98747.17  410.67792 492709.16 4734563.3 47.946318
[10,] 40  5 2.50000000   958 641400 1.493608e-03 0.0074402566 0.9925597  98336.49  731.64875 489853.35 4241854.2 43.136113
[11,] 45  5 2.50000000  1468 608625 2.411994e-03 0.0119876857 0.9880123  97604.85 1170.05621 485099.09 3752000.8 38.440723
[12,] 50  5 2.50000000  2162 573895 3.767240e-03 0.0186604523 0.9813395  96434.79 1799.51678 477675.15 3266901.7 33.876797
[13,] 55  5 2.50000000  2723 519884 5.237707e-03 0.0258500469 0.9741500  94635.27 2446.32623 467060.55 2789226.6 29.473435
[14,] 60  5 2.50000000  3815 517531 7.371539e-03 0.0361907418 0.9638093  92188.95 3336.38634 452603.76 2322166.0 25.189202
[15,] 65  5 2.50000000  4872 451942 1.078014e-02 0.0524861997 0.9475138  88852.56 4663.53320 432603.97 1869562.3 21.041175
[16,] 70  5 2.50000000  6890 414658 1.661610e-02 0.0797669739 0.9202330  84189.03 6715.50388 404156.37 1436958.3 17.068237
[17,] 75  5 2.50000000 12120 423353 2.862859e-02 0.1335822754 0.8664177  77473.52 10349.08944 361494.89 1032801.9 13.331031
[18,] 80  5 2.50000000 15967 299871 5.324623e-02 0.2349549793 0.7650450  67124.43 15771.21981 296194.12  671307.0 10.000934
[19,] 85 NA         NA 37125 271182 1.369007e-01 1.0000000000 0.0000000  51353.21 51353.21337 375112.92  375112.9  7.304566
> 
```

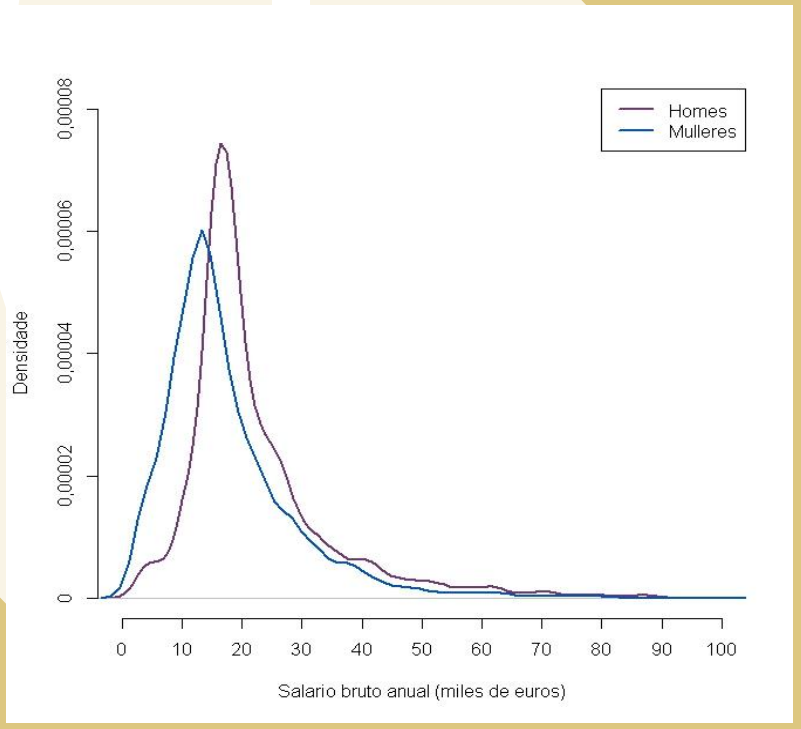
INSTITUTO GALEGO DE ESTATÍSTICA

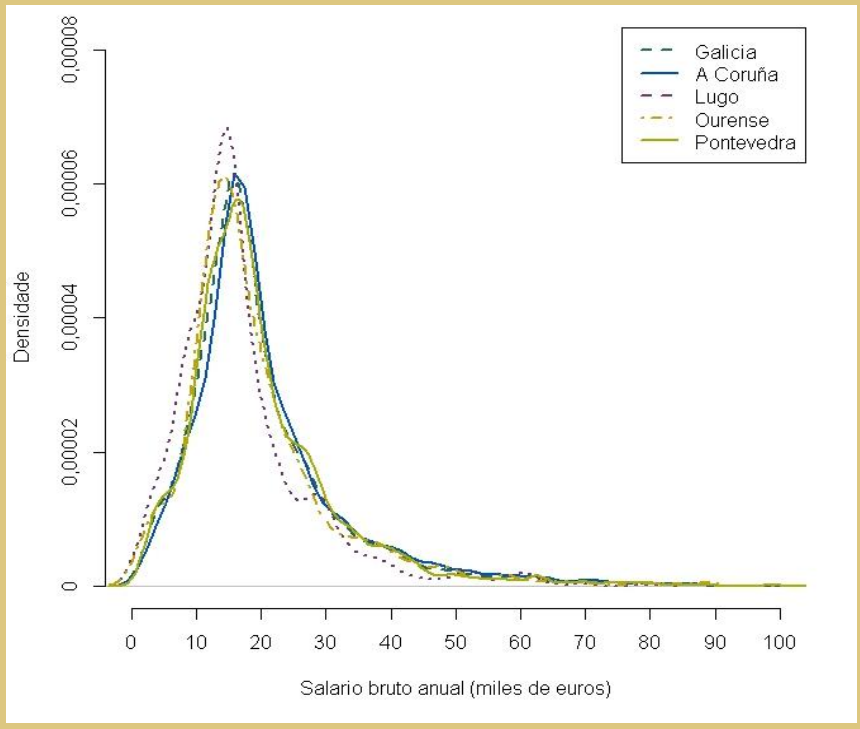
Outras aplicacións: Estudo de salarios

Microdatos da Encuesta de estructura salarial

d1<-density(datos.sexo[[1]]$salanual,

weights=datos.sexo[[1]]$frecuencia/sum(datos.sexo[[1]]$frecuencia))





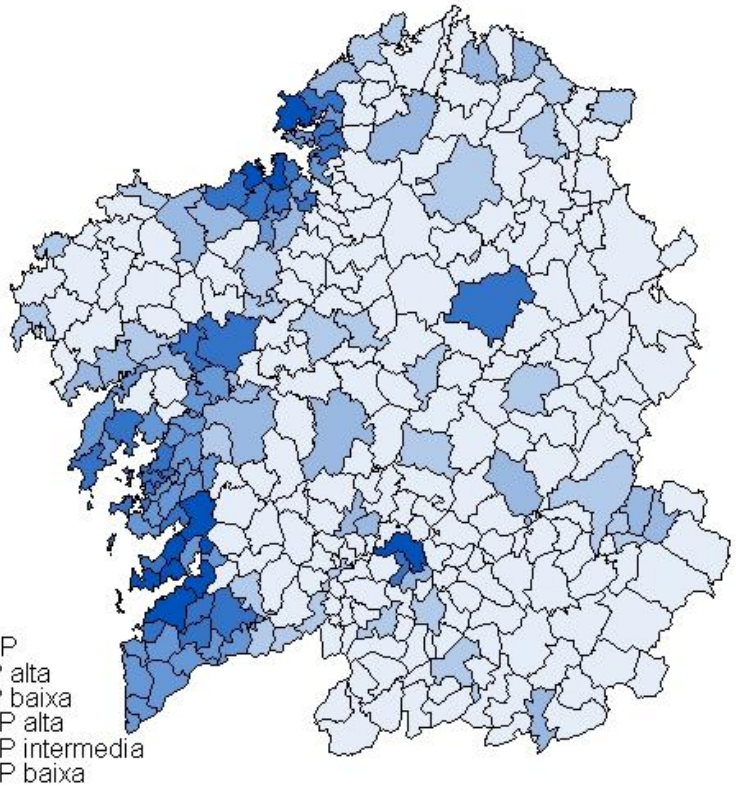
INSTITUTO GALEGO DE ESTATÍSTICA

Outras aplicacións: Mapas

Panorama rural- urbano

```
muni_shp=readShapeSpatial("R:/PANORAMA RURAL_URBANO/Información cartográfica/Municipios/CONCELLO_1000.shp")
clas=read.csv2("Clasificacion_urbanizacion.csv")
niveles=clas_ord$SubGraoUrbanización
fgs=colores [niveles]
plot(muni_shp,col=fgs)
```

Mapa dos concellos segundo o subgrao de urbanización



INSTITUTO GALEGO DE ESTATÍSTICA

Outras aplicacións: Gráficos de burbullas (gapminder)

Galicia- Norte de Portugal

Ficheiro cos datos de Galicia e Portugal no período 2000-2010

```
> datos=read.csv2("datos_gapminder.csv")
> names(datos)
[1] "Espazo"            "Ano"               "Tasa.de.natalidad"
[4] "Tasa.de.mortalidad" "Tasa.de.actividad"  "Tasa.de.paro"
[7] "PIB.ph"            "PIB.ph.PPC"        "Loc"
>
```

Sintaxe (emprégase a librería googleVis)

```
> library("googleVis")
> graf1=gvisMotionChart(datos,idvar="Espazo",timevar="Ano",
options=list(gvis.language="es",width=900,height=600))
> plot(graf1)
```

GRÁFICO

INSTITUTO GALEGO DE ESTATÍSTICA

Gracias pola súa atención

www.ige.eu

ige@ige.eu

INSTITUTO GALEGO DE ESTATÍSTICA