INSTITUTO GALEGO DE ESTATÍSTICA

Explotación dos datos da Mostra continua de vidas laborais con dplyr: claridade e rapidez

María del Pilar Romero Martínez

http://www.ige.eu

II Xornada de usuarios de R en Galicia. Santiago de Compostela, 1 de outubro de 2015

Mostra continua de vidas laborais (MCVL)

Conxunto de microdatos **anonimizados** procedentes dos rexistros administrativos:

- Bases de datos da Seguridade Social (Afiliacións, bases de cotización e pensións)
- Padrón municipal continuo do INE
- Modelo 190 do IRPF

INSTITUTO GALEGO DE ESTATÍSTICA

Selección da MCVL

A **MCVL** é unha mostra que selecciona 4 de cada 100 persoas que no ano de referencia tiveron relación coa SS:

- Afiliadas en alta laboral
- Cobrando unha prestación contributiva ou subsidio por desemprego
- Percibindo unha pensión contributiva da SS

INSTITUTO GALEGO DE ESTATÍSTICA

Obxectivos da explotación da MCVL

- Clasificar a poboación segundo a relación coa SS no ano de referencia.
- Analizar as características das relacións laborais dos traballadores: parcialidade, temporalidade, ...
- Proporcionar información sobre a estabilidade laboral e tempo traballado ao longo da vida laboral.
- Estudar os salarios e bases de cotización.

INSTITUTO GALEGO DE ESTATÍSTICA

Ficheiros que compoñen a MCVL (edición 2013)

Ficheiro	Rexistros (total)	Rexistros (Galicia)
Persoas	1.172.383	72.784
Afiliacións	20.328.833	1.234.922
Bases de cotización conta allea	21.243.178	1.162.142
Bases de cotización conta propia	4.797.548	323.886
Pensións	4.163.083	340.925
Que conviven	1.31.143	71.350
Fiscal	1.826.432	110.411

Presentes e pasados

INSTITUTO GALEGO DE ESTATÍSTICA

persoas

id_per	sexo	fnac	...
000000000120299	1	195609	...
000000000163222	1	198303	...
...	...	...	...

fiscal

id_per	id_pagador	clave	percepcion	...
000000000120299	000000000475336	A	2280080	...
000000000162864	000000000665717	C	153360	...
...	...	...	...	...

afiliacions

id_per	FecAlta	FecBaja	RegCotiza	CC_Secundaria	id_pagador	...
000000000120299	20000320	20121228	0111	0111330066174	000000000652927	...
000000000120299	20130201	20141231	0111	0111158043125	000000000475336	...
...	...	...	...	...	...	...

bases

id_per	CC_Secundaria	AnoCotiza	Base01	Base02	...	Base12
000000000120299	0111330066174	2000	0	0	...	203466
000000000120299	0111330066174	2001	203466	203466	...	203466
...	...	...	...	...	...	...

INSTITUTO GALEGO DE ESTATÍSTICA

Fases na explotación da MCVL

- Recibimos os ficheiros da SS en .txt
- Importamos os ficheiros a SQL-Server
- Depuramos a mostra con R
- Construímos ficheiros “finais” de microdatos que gardamos en SQL
- Explotamos os microdatos con R
- Gardamos as táboas resultado da explotación en Mysql mediante a función GrabarH
- Os datos preséntanse na páxina web en forma de táboas multidimensionais

INSTITUTO GALEGO DE ESTATÍSTICA

Depuración da MCVL

- Depuración das relacións de desemprego:
  - Prestacións contributivas por desemprego con duración superior aos 2 anos
  - Traballadores por conta propia ou allea con xornada completa cobrando prestacións por desemprego
- Incongruencias entre variables da mostra: idade – tempo traballado

INSTITUTO GALEGO DE ESTATÍSTICA

Depuración das relación de desemprego

1. Cruzamos os rexistros de prestacións contributivas por desemprego da táboa `afilicions` coa táboa de bases de cotizacion (`bases`) → Obtemos o último mes para o que figura unha base de cotización >0
2. Comprobamos si existen persoas que perciben varias prestacións por desemprego simultaneamente
3. Cruzamos os perceptores por desemprego cos traballadores por conta propia → Anulamos o paro nos días que hai coincidencia
4. Cruzamos os perceptores por desemprego cos traballadores por conta allea a xornada completa → Anulamos o paro nos días que hai coincidencia

INSTITUTO GALEGO DE ESTATÍSTICA

Depuración e explotación da MCVL

	Edición 2011 (só poboación xuvenil)	Edición 2012		Edición 2013
Depuración	SQL-Server	R	+ rápido	R – dplyr
Explotación e gravación de táboas multidimensionais	Excel	R	+ claro	R – dplyr

Sistema de xestión de bases de datos relacionais. Non é o procedemento idóneo para programar certas funcións, p.ex. tempo traballado

Precisábamos un método máis “automático” que nos permitise replicar as mesmas táboas en todas as edicións da mostra

INSTITUTO GALEGO DE ESTATÍSTICA

dplyr

- Creado por Hadley Wickham
- Implementa unha sintaxe en cadea polo que o código é máis doado de ler e de entender que a sintaxe aniñada **%>%**
- Traballa con data.frames pero tamén permite realizar consultas a bases de datos externas

INSTITUTO GALEGO DE ESTATÍSTICA

dplyr

- Inclúe 5 vervos para traballar con táboas:
  - filter: filtrar filas
  - select: seleccionar columnas
  - arrange: ordenar filas
  - sumarise: resumir
  - mutate: calcula novas variables
- Permite realizar operacións con dúas táboas (joins): left_join, right_join, full_join, inner_join

INSTITUTO GALEGO DE ESTATÍSTICA

Exemplo: dplyr

Partimos do ficheiro `afiliacions` que contén todas as afiliacións (presentes e pasadas) dos traballadores galegos. Este data.frame contén as seguintes variables: *id_per, FecAlta, FecBaja, AnoBaixa, RegCotiza*

Queremos obter unha táboa na que figuren para cada traballador o número de afiliacións no réxime xeral en alta laboral no ano 2013, ordenada polo identificador da persoa (*id_per*).

INSTITUTO GALEGO DE ESTATÍSTICA

Exemplo: dplyr

dplyr

```
persoas<-afiliacions %>% filter(AnoBaixa>2012,RegCotiza=="0111") %>%
group_by(id_per) %>% summarise(afiliacions=n())
```

R

```
afiliacions$contador<-1

afiliaxeral<-afiliacions[afiliacions$AnoBaixa>2012 &
afiliacions$RegCotiza=="0111",]

persoas<-
aggregate(afiliaxeral$contador,by=list(afiliaxeral$id_per),FUN="sum")

names(persoas)<-c("id_per","afiliacions")

persoas<-persoas[order(personas[,1]),]
```

INSTITUTO GALEGO DE ESTATÍSTICA

Comparación de tempos (exemplo 1)

Collemos os identificadores de persoa distintos e ordenámolos

- Con dplyr

```
> t <- proc.time()
> personas<-afiliacions %>% select(id_per) %>% distinct() %>%arrange(id_per)
>
> proc.time()-t
   user  system elapsed
   0.13    0.00    0.12
>
```

- Con R

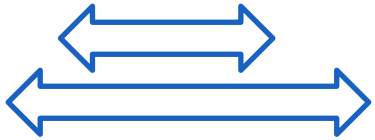
```
> t <- proc.time()
>
> afiliacions$contador<-1
> personas<-aggregate(afiliacions$contador, by=list(afiliacions$id_per), FUN="sum")
>
> names(personas)<-c("ip_per","relacions")
> personas<-personas[order(personas[,1]),]
>
> proc.time()-t
   user  system elapsed
   4.63    0.16    4.78
```

Comparación de tempos (exemplo 2)

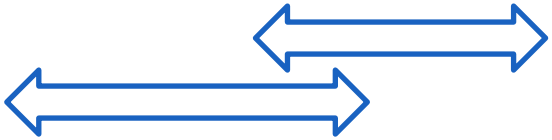
Programa que calcula o tempo traballado por unha persoa ao longo da súa vida laboral. Para cada ano desde 1960 devolve o número de días en alta laboral nese ano.

Non se pode calcular a diferenza entre a data de alta e baixa pois unha mesma persoa pode ter varias relacións laborais que se solapan no tempo:

Empregos consecutivos



Empregos totalmente superpostos



Empregos parcialmente superpostos

INSTITUTO GALEGO DE ESTATÍSTICA

Comparación de tempos (exemplo 2)

Con dplyr

```
personas<-afiliacions %>% select(id_per) %>% distinct() %>%arrange(id_per)

aux_dias<-read.csv2("R:/Servicios/Difusion/OPERACIÓNS ESTATISTICAS/ACTIVIDADES-PUBLICACIONES/MC\

acumulado<-rep(0,nrow(personas))
for (j in 1960:2012)
{if (j%%4==0 && (j%%100 != 0 || j%%400 == 0))
 {for (i in 1:366)
  {dia<-j*10000+aux_dias$sibisiesto[i]
   afiliacions<-afiliacions %>% mutate(aux=ifelse(dia<=FecBaja & dia>=FecAlta,1,0))
  t<-afiliacions %>% group_by(id_per)%>% summarise(max(aux)) %>% arrange(id_per) %>% ungroup()
  acumulado<-acumulado+t[,2]}}
else
 {for (i in 1:365)
  {dia<-j*10000+aux_dias$nonbisiesto[i]
   afiliacions<-afiliacions %>% mutate(aux=ifelse(dia<=FecBaja & dia>=FecAlta,1,0))
  t<-afiliacions %>% group_by(id_per)%>% summarise(max(aux)) %>% arrange(id_per) %>% ungroup()
  acumulado<-acumulado+t[,2]}}
personas<-cbind(personas,acumulado)
acumulado<-rep(0,nrow(personas))
}
```

Para o ano 2012 ⟶

```
   user  system elapsed
484.79   11.33  497.53
```

INSTITUTO GALEGO DE ESTATÍSTICA

Comparación de tempos (exemplo 2)

Con R

```
resultados$contador<-1
personas<-aggregate(resultados$contador, by=list(resultados$id_per), FUN="sum")
names(personas)<-c("ip_per","relacions")
personas<-personas[order(personas[,1]),]

aux_dias<-read.csv2("R:/Servicios/Difusion/OPERACIÓNS ESTATISTICAS/ACTIVIDADES-PUBLICACIONES/MCV
acumulado<-rep(0,nrow(personas))
for (j in 1960:2012)
{if (j%%4==0 && (j%%100 != 0 || j%%400 == 0))
 {for (i in 1:366)
  {dia<-j*10000+aux_dias$sibisiesto[i]
  resultados$aux<-ifelse(dia<=resultados$FecBaja & dia>=resultados$FecAlta ,1, 0)
  t<-aggregate(resultados$aux,by=list(resultados$id_per), FUN="sum")
  t[,2]<-ifelse(t[,2]>0,1,0)
  t<-t[order(t[,1]),]
  acumulado<-acumulado+t[,2]}}
else
 {for (i in 1:365)
  {dia<-j*10000+aux_dias$nonbisiesto[i]
  resultados$aux<-ifelse(dia<=resultados$FecBaja & dia>=resultados$FecAlta ,1, 0)
  t<-aggregate(resultados$aux,by=list(resultados$id_per), FUN="sum")
  t[,2]<-ifelse(t[,2]>0,1,0)
  t<-t[order(t[,1]),]
  acumulado<-acumulado+t[,2]}}
personas<-cbind(personas,acumulado)
acumulado<-rep(0,nrow(personas))
}
```

Para o ano 2012 ⟶

```
   user  system elapsed
1926.16   28.36 1972.66
```

INSTITUTO GALEGO DE ESTATÍSTICA

Difusión de resultados

- Poboación total (Galicia, provincias, 7 grandes concellos)
  - Clasificación da poboación segundo a relación coa SS.
  - Traballadores afiliados en alta laboral segundo: sexo, idade, tempo traballado, número de empresas …
  - Bases de cotización (media e mediana): sexo, idade, nacionalidade, grupos de cotización, sectores de actividade, …
  - Salarios (media e mediana): sexo, idade, tempo traballado, tipo de xornada, tipo de contrato, …
- Módulo de poboación xuvenil (Galicia e provincias)

INSTITUTO GALEGO DE ESTATÍSTICA

Difusión de resultados

Media e mediana das bases de cotización por grupos de idade. Galicia. Ano 2012

Euros/mes

	Media	Mediana
Total	1.463	1.323
De 16 a 24 anos	864	825
De 25 a 34 anos	1.297	1.216
De 35 a 44 anos	1.530	1.374
De 45 a 54 anos	1.639	1.478
De 55 a 64 anos	1.718	1.567
65 ou máis anos	1.552	1.478

Media e mediana das bases de cotización para os 7 grandes concellos. Ano 2012

Euros/mes

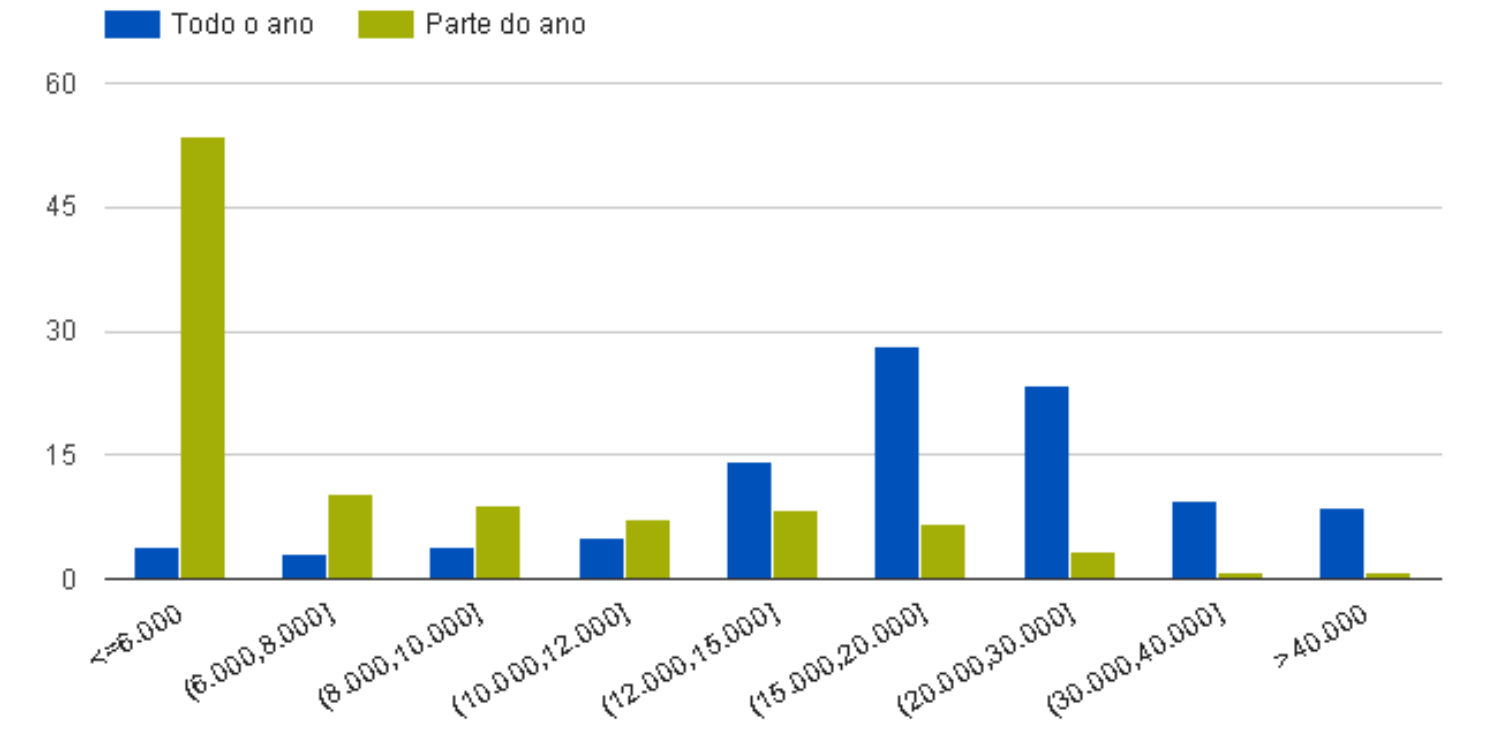
	Media	Mediana
15030 Coruña, A	1.611	1.414
15036 Ferrol	1.534	1.468
15078 Santiago de Compostela	1.654	1.495
27028 Lugo	1.492	1.330
32054 Ourense	1.478	1.292
36038 Pontevedra	1.577	1.419
36057 Vigo	1.565	1.435

INSTITUTO GALEGO DE ESTATÍSTICA

Difusión de resultados

Distribución da poboación afiliada en alta laboral por conta allea segundo o intervalo de salario bruto anual (euros). Galicia. Ano 2012

Porcentaxes



INSTITUTO GALEGO DE ESTATÍSTICA

Difusión de resultados

Media e mediana do salario bruto anual segundo o tempo traballado por conta allea no ano e o sexo. Galicia. Ano 2012

Euros/ano

	Total	Todo o ano	Parte do ano
Media			
Total	16.492	22.197	7.449
Homes	18.755	24.922	8.399
Mulleres	14.006	19.061	6.481
Mediana			
Total	14.368	18.229	5.409
Homes	16.294	19.808	6.096
Mulleres	12.201	16.031	4.822

Media do salario bruto anual segundo o tempo traballado ao longo da vida laboral. Galicia. Ano 2012

Euros/ano

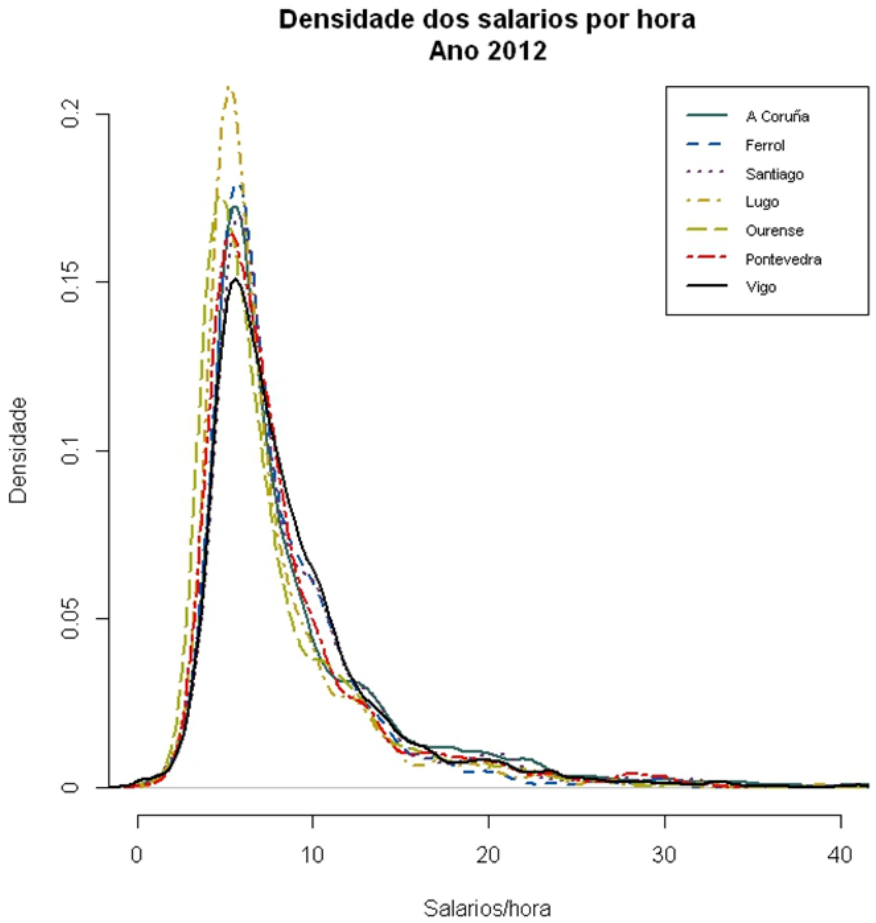
	Total	Todo o ano	Parte do ano
Total	16.492	22.197	7.449
Menos de 5 anos	7.462	15.102	5.166
De 5 anos a menos de 15 anos	15.152	19.347	8.122
De 15 anos a menos de 25 anos	21.013	24.250	9.780
De 25 anos a menos de 35 anos	25.736	28.487	11.842
35 ou máis anos	24.530	26.515	14.351

INSTITUTO GALEGO DE ESTATÍSTICA

Difusión de resultados

Densidade dos salarios por hora nos 7 grandes concellos. Ano 2012

Porcentaxes



INSTITUTO GALEGO DE ESTATÍSTICA

Grazas pola atención

www.ige.eu

INSTITUTO GALEGO DE ESTATÍSTICA