Formación de talento humano en el Centro de Estudios Avanzados del Instituto Venezolano de Investigaciones Científicas

JR Ferrer-Paris
Laboratorio de Ecología Espacial
Centro de Estudios Botánicos y Agroforestales
Instituto Venezolano de Investigaciones Científicas

LEE.CEA.2017.1
DOI:10.6084/m9.figshare.3394789
Versión de 8 de mayo de 2017

A quien pueda interesar

Este es un documento de carácter didáctico generado por el laboratorio de Ecología Espacial del Centro de Estudios Botánicos y Agroforestales del Instituto Venezolano de Investigaciones Científicas, con el objetivo de demostrar la aplicación de herramientas de análisis estadísticos.

Este documento es generado utilizando las funciones de `Sweave` desde una sesión de *R* (R Development Core Team, 2015), por tanto todas las tablas y figuras se generan y actualizan automáticamente a partir de los datos suministrados. Se suministra un enlace al código fuente en *R* y los archivos de datos. Dentro de *R* utilizamos los paquetes `plotrix` (Lemon, 2006); `rvest` (Wickham, 2015a); `xml2` (Wickham, 2015b).

1. Preambulo

En este documento demuestro el uso de funciones de los paquetes `xml2` y `rvest` para leer datos de una tabla procedente de una página web y transformarlos en un marco de datos de *R* sobre el cuál se pueden operar varias funciones para visualizar proporciones y tendencias.

2. Caso de estudio: Los egresados del Centro de Estudios Avanzados del IVIC entre 1973 y 2017

El Centro de Estudios Avanzados (CEA) del Instituto Venezolano de Investigaciones Científicas (IVIC) es un organismo coordinador de actividades académicas de cuarto nivel en Venezuela. Fue creado a principios de 1970 y tiene su sede en Altos de Pipe, estado Miranda. Con el objetivo de documentar la actividad del CEA para la página correspondiente en la Wikipedia, quise descargar los datos detallados de los egresados de esta institución entre 1973 y 2017. Éstos datos están disponibles en `http://cea.ivic.gob.ve/?accion=egresados` (fecha de consulta: 8 de mayo de 2017).

Los datos están separados por año, y para cada año se detalla la información de Nº de cédula, nombre completo del estudiante, grado académico otorgado, área de estudio, sexo y nacionalidad.

La información está disponible en un portal público y sólo es necesario descargar cada una de las 43 páginas, copiar el contenido de cada página y juntarlo en una tabla o marco de datos. La automatización de este procedimiento es conocida como "web scraping" o "web harvesting".

2.1. Web scraping

Afortunadamente las páginas a consultar están estructurados homogéneamente. El URL `http://cea.ivic.gob.ve/?accion=consultaregresados&param=1973` contiene el parámetro `param` seguido del valor del año a consultar.

Para descargar los datos podemos utilizar la función `download.file`. Para el año 1973 usamos:

```
> k <- 1973
> if (!file.exists(sprintf("EgresadosAño%s.html",k)))
+     download.file(sprintf("http://cea.ivic.gob.ve/?accion=consultaregresados&param=
```

Con un "bucle" (*for loop*) es muy sencillo descargar todos los archivos para k años:

```
> for (k in 1973:2017) {
+     dst.file <- sprintf("EgresadosAño%s.html",k)
+     if (!file.exists(dst.file))
+         download.file(sprintf("http://cea.ivic.gob.ve/?accion=consultaregresados&pa
+
+ }
>
```

Al final debemos tener 44 archivos:

```
> dir(pattern="EgresadosAño")

 [1] "Documento1_PostgradoCEA-EgresadosAño.pdf"
 [2] "EgresadosAño1973.html"
 [3] "EgresadosAño1974.html"
 [4] "EgresadosAño1975.html"
 [5] "EgresadosAño1976.html"
 [6] "EgresadosAño1977.html"
 [7] "EgresadosAño1978.html"
 [8] "EgresadosAño1979.html"
 [9] "EgresadosAño1980.html"
[10] "EgresadosAño1981.html"
[11] "EgresadosAño1982.html"
[12] "EgresadosAño1983.html"
[13] "EgresadosAño1984.html"
[14] "EgresadosAño1985.html"
[15] "EgresadosAño1986.html"
[16] "EgresadosAño1987.html"
[17] "EgresadosAño1988.html"
[18] "EgresadosAño1989.html"
[19] "EgresadosAño1990.html"
[20] "EgresadosAño1991.html"
[21] "EgresadosAño1992.html"
[22] "EgresadosAño1993.html"
[23] "EgresadosAño1994.html"
[24] "EgresadosAño1995.html"
[25] "EgresadosAño1996.html"
[26] "EgresadosAño1997.html"
[27] "EgresadosAño1998.html"
[28] "EgresadosAño1999.html"
[29] "EgresadosAño2000.html"
[30] "EgresadosAño2001.html"
[31] "EgresadosAño2002.html"
[32] "EgresadosAño2003.html"
[33] "EgresadosAño2004.html"
[34] "EgresadosAño2005.html"
[35] "EgresadosAño2006.html"
[36] "EgresadosAño2007.html"
[37] "EgresadosAño2008.html"
[38] "EgresadosAño2009.html"
[39] "EgresadosAño2010.html"
[40] "EgresadosAño2011.html"
```

```
[41] "EgresadosAño2012.html"
[42] "EgresadosAño2013.html"
[43] "EgresadosAño2014.html"
[44] "EgresadosAño2015.html"
[45] "EgresadosAño2016.html"
[46] "EgresadosAño2017.html"

>
```

Podemos leer un archivo html en R usando la función `read_html` del paquete `rvest`.

```
> require(rvest)
>    pg <- read_html(sprintf("EgresadosAño%s.html",k))
>
```

Nos interesa ubicar dentro de la página la tabla con la información de los egresados. Con `html_nodes` podemos verificar cuantas tablas hay:

```
>  html_nodes(pg,"table")

{xml_nodeset (6)}
[1] <table width="500" border="0" align="center" cellpadding="0" cellspacing= ...
[2] <table width="500" border="0" align="left" cellpadding="0" cellspacing="0 ...
[3] <table width="500" border="0" cellpadding="0" cellspacing="0">\n<tr>\n<td ...
[4] <table border="0"><tr>\n<td align="right"><b>            Año de Graduació ...
[5] <table border="1" width="900" align="center" rules="rows">\n<tr bgcolor=" ...
[6] <table width="910" border="0" align="center" cellpadding="0" cellspacing= ...
```

Y luego podemos ver el inicio de cada tabla con:

```
> str(lapply(html_nodes(pg,"table"),xml_text))

List of 6
 $ : chr " \r\n\r\n        \r\n\t\t\r\n      \r\n       \r\n      \r\n\r\n\r\n    \r\n\
 $ : chr "\r\n        \r\n\t\t\r\n      \r\n       \r\n      \r\n\r\n\r\n    \r\n\r\ndi
 $ : chr "\r\n\t \t\t            Inicio  /  Egresados  /  Consultar  Egresados del Año:
 $ : chr "            Año de Graduación:\r\n\t2017\r\n"
 $ : chr "Cédula\r\n\tEstudiante\r\n\tGrado Académico\r\n\tArea de Estudio\r\n\tSexo\r\
 $ : chr "\r\n           Este sitio está desarrollado cumpliendo con los lineamientos d
```

Por lo visto la tabla número cinco es la que contiene la información requerida, probamos importarla con `html_table`:

```
>   rr <- html_table(html_nodes(pg,"table")[5],header=T,fill=T)[[1]]
>   head(rr)
```

```
    Cédula                          Estudiante    Grado Académico
1 14750554    ECHEVARRIA DIAZ GABRIELA ELENA DOCTOR EN CIENCIAS
2 17340558                 RUIZ RAMONI DAMIAN DOCTOR EN CIENCIAS
3 14250555     RAMIREZ LASSO LILIA MARGARITA DOCTOR EN CIENCIAS
4 14091642       ROSALES ANZOLA SERGIO DAVID DOCTOR EN CIENCIAS
5 14377139 MELENDEZ GALLARDO JOSE GUMERCINDO DOCTOR EN CIENCIAS
6 12404080    ARRIA BOHORQUEZ MELISSA LORENA DOCTOR EN CIENCIAS
                  Area de Estudio      Sexo Nacionalidad
1                        ECOLOGIA  Femenino   VENEZOLANA
2                        ECOLOGIA Masculino   VENEZOLANA
3 ESTUDIOS SOCIALES DE LA CIENCIA  Femenino   VENEZOLANA
4                          FISICA Masculino   VENEZOLANA
5          FISIOLOGIA Y BIOFISICA Masculino   VENEZOLANA
6                 GENETICA HUMANA  Femenino   VENEZOLANA

>
```

Ahora es cuestión de repetir esta misma operación para todos los años. Nuevamente usamos un bucle para los k años:

```
> rm(rs)
> for (k in 1973:2017) {
+     pg <- read_html(sprintf("EgresadosAño%s.html",k))
+     rr <- html_table(html_nodes(pg,"table")[5],header=T,fill=T)[[1]]
+     if (nrow(rr)>0) {
+         rr$Año <- k
+         if (!exists("rs")) {
+             rs <- rr
+         } else {
+             rs <- rbind(rs[,colnames(rr)],rr)
+         }
+     }
+ }
>
```

A cada tabla le agregamos una columna con el año para poder separar los registros por esta variable, luego unimos las tablas fila tras fila (`rbind`).

Finalmente el objeto `rs` contiene el resultado de unir todas las tablas:

```
> str(rs)

'data.frame':        1364 obs. of  7 variables:
 $ Cédula         : chr  "E-987543" "1670128" "5134007" "1741694" ...
 $ Estudiante     : chr  "BARROS PITA JOSE CARLOS" "DIAZ DE EWALD MARIA" "FRAGACHAN MIR
 $ Grado Académico: chr  "MAGISTER SCIENTIARUM" "MAGISTER SCIENTIARUM" "MAGISTER SCIENT
```

```
 $ Area de Estudio: chr  "ARTICULO 9" "ARTICULO 9" "ARTICULO 9" "ARTICULO 9" ...
 $ Sexo           : chr  "MASCULINO" "FEMENINO" "MASCULINO" "MASCULINO" ...
 $ Nacionalidad   : chr  NA NA NA NA ...
 $ Año            : int  1973 1973 1973 1973 1973 1973 1973 1973 1973 1973 ...
```

2.2. Normalización de los datos

Es muy común encontrar información heterogénea o formatos incosistentes en bases de datos. El uso no normalizado de mayúsculas y minúsculas, errores de tipeo o cambios de formato son los problemas más frecuentes. Verificamos las columnas que usaremos en el análisis. El grado académico aparece con tres valores posibles:

```
> table(rs$"Grado Académico")

     DOCTOR EN CIENCIAS    MAGISTER SCIENTIARUM PHILOSOPHUS SCIENTIARUM
                    235                     986                     143
```

Esto se debe a un cambio formal en el CEA, a partir del 2002 se sustituyó el título de "Philosophus scientiarum" por el de "Doctor en Ciencias" (referido como *Philosophae doctor* o *PhD* en el contexto internacional).

La columna Sexo muestra un valor vacío y uso inconsistente de mayúsculas:

```
> table(rs$Sexo,useNA="always")

          Femenino  FEMENINO Masculino MASCULINO      <NA>
        1       175       594       103       491         0
```

El valor vacío se refiere a un persona con un nombre neutral y no podemos verificar su verdadero valor, pero al ser un caso aislado no es muy importante. Los demás valores los pasamos a minúsculas:

```
> rs$Sexo <- tolower(rs$Sexo)
> table(rs$Sexo,useNA="always")

          femenino masculino      <NA>
        1       769       594         0

>
```

La columna de nacionalidad tiene más de 700 casos con valores vacíos, y no fue llenada de manera consistente:

```
> table(rs$Nacionalidad,useNA="always")
```

```
                      CHILENA         COLOMBIANA         COLOMBIANO
          492               1                 30                  2
COLOMBO-VENEZOLANA        CUBANA            CUBANO         ECUATORIANA
            1               2                  1                  3
       Ecuatoriano  ECUATORIANO          GEORGIANA           ITALIANA
            1               2                  2                  1
    NORTEAMERICANA   Venezolana         VENEZOLANA         VENEZOLANO
            1               1                583                  2
           <NA>
          239
```

Sin embargo la mayoría de los egresados con información de nacionalidad fueron de nacionalidad venezolana, y suponemos que un patrón similar se observará en los que tienen valores vacíos.

La columna del área de estudio es la que muestra más inconsistencia en el uso de acentos, errores de tipeo y cambios de nombre que dan un total de 48 valores únicos:

```
> table(rs$"Area de Estudio")

                          ANTOPOLOGIA
                                    1
                         ANTROPOLOGIA
                                   48
                         ARTICULO 10
                                    4
                          ARTICULO 9
                                   26
              BACTERIOLOGIA Y MICOLOGIA
                                    3
                            BIOLOGIA
                                   12
       BIOLOGIA DE A REPRODUCCION HUMANA
                                    1
      BIOLOGIA DE LA REPRODUCCION HUMANA
                                   30
      BIOLOGIA DE LA REPRODUCCIÒN HUMANA
                                    4
      BIOLOGÌA DE LA REPRODUCCIÒN HUMANA
                                    1
                          BIOQUIMICA
                                  133
                          BIOQUÌMICA
                                    1
                            ECOLOGIA
                                  116
                                    7
```

ESTUDIO SOCIALES DE LA CIENCIA	1
ESTUDIOS DE LA CIENCIAS	1
ESTUDIOS SOCIALES DE LA CIENCIA	20
FISICA	105
FISICA MEDICA	52
FISICOLOGIA Y BIOFISICA	3
FISIOLOGIA Y BIOFISICA	88
FISIOLOGIA Y BIOFISÌCA	2
FISIOLOGIA Y BIOFÌSICA	1
FISIOLOGÍA Y BIOFÍSICA	1
FISOLOGIA Y BIOFISICA	1
GENETICA HUMANA	41
GENTICA HUMANA	1
IMNUNOLOGIA	2
INGENIERIA AMBIENTAL	11
INGENIERIA ELECTRICA	11
INMUMOLOGIA	1
INMUNOLOGIA	164
INMUNOLOGÌA	10
MATEMATICA	11
MATEMATICAS	21
METALURGIA Y CIENCIA DE LOS MATERIALES	1

```
    METALURGIA Y CIENCIA DE LOS MATERIALES
                                         1
   METALURGICA Y CIENCIA DE LOS MATERIALES
                                         7
  METALURGICA Y CIENCIAS DE LOS MATERIALES
                                         2
METALURGICA Y TECNOLOGIA DE LOS MATERIALES
                                         1
   METALURIGIA Y CIENCIA DE LOS MATERIALES
                                         1
                            MICROBIOLOGIA
                                       130
                            MICROBIOLOGÌA
                                         2
                         MODELOS ALEATORIOS
                                        11
                                  QUIMICA
                                       262
                                  QUÌMICA
                                         4
                          QUIMICA AMBIENTAL
                                         5
                        REPRODUCCION HUMANA
                                         3
                                 VIROLOGIA
                                         6
```

La función `adist` calcula la distancia generalizada de Levenshtein entre un par de palabras. Al aplicarla a la columna de área de estudio genera una matriz de disimilitud con la cual podemos hacer un análisis de clasificación jerárquica con `hclust`:

```
> d0 <- adist(rs$"Area de Estudio")
> h0 <- hclust(as.dist(d0))
```

Usando una tolerancia de tres diferencias entre dos nombres, el número de grupos se reduce a 23. Un ejemplo del resultado se muestra a continuación:

```
> table(rs$"Area de Estudio",cutree(h0,h=3))[18:30,c(4:15,23)]
```

```
                          4  5  6  7  8  9 10 11 12 13 14 15 23
  FISICA MEDICA           0  0  0  0  0  0  0  0  0  0  0  0 52
  FISICOLOGIA Y BIOFISICA 3  0  0  0  0  0  0  0  0  0  0  0  0
  FISIOLOGIA Y BIOFISICA 88  0  0  0  0  0  0  0  0  0  0  0  0
  FISIOLOGIA Y BIOFISÌCA  2  0  0  0  0  0  0  0  0  0  0  0  0
  FISIOLOGIA Y BIOFÌSICA  1  0  0  0  0  0  0  0  0  0  0  0  0
```

```
  FISIOLOGÍA Y BIOFÍSICA   1  0  0  0  0  0  0  0  0  0  0  0  0
  FISOLOGIA Y BIOFISICA    1  0  0  0  0  0  0  0  0  0  0  0  0
  GENETICA HUMANA          0  0  0  0  0  0 41  0  0  0  0  0  0
  GENTICA HUMANA           0  0  0  0  0  0  1  0  0  0  0  0  0
  IMNUNOLOGIA              0  0  0  2  0  0  0  0  0  0  0  0  0
  INGENIERIA AMBIENTAL     0  0  0  0  0  0  0  0  0  0  0 11  0
  INGENIERIA ELECTRICA     0  0  0  0  0  0  0  0  0 11  0  0  0
  INMUMOLOGIA              0  0  0  1  0  0  0  0  0  0  0  0  0

>
```

Las seis versiones de Fisiología y Biofísica son clasificadas dentro del grupo 4, las dos versioness de Genética Humana están en el grupo 10, etc.

Para simplificar, genero una columna con el área corregida:

```
> c0 <- cutree(h0,h=3)
> for (j in unique(c0)) {
+     rs$area.corregida[c0 %in% j] <- names(  sort(table(rs[c0 %in% j,"Area de Estudi
+
+ }
```

La tabla final todavía tiene un par de errores menos obvios:

```
> table(rs$area.corregida)

                              ANTROPOLOGIA
                                        49
                                ARTICULO 9
                                        30
                BACTERIOLOGIA Y MICOLOGIA
                                         3
      BIOLOGIA DE LA REPRODUCCION HUMANA
                                        36
                              BIOQUIMICA
                                       134
                                ECOLOGIA
                                       134
                ESTUDIOS DE LA CIENCIAS
                                         1
         ESTUDIOS SOCIALES DE LA CIENCIA
                                        21
                                FISICA
                                       105
                           FISICA MEDICA
                                        52
```

```
                      FISIOLOGIA Y BIOFISICA
                                          96
                            GENETICA HUMANA
                                          42
                       INGENIERIA AMBIENTAL
                                          11
                       INGENIERIA ELECTRICA
                                          11
                                 INMUNOLOGIA
                                         177
                                MATEMATICAS
                                          32
   METALURGICA Y CIENCIA DE LOS MATERIALES
                                          12
METALURGICA Y TECNOLOGIA DE LOS MATERIALES
                                           1
                              MICROBIOLOGIA
                                         132
                         MODELOS ALEATORIOS
                                          11
                                    QUIMICA
                                         266
                          QUIMICA AMBIENTAL
                                           5
                        REPRODUCCION HUMANA
                                           3

>
```

Corregimos pequeñas diferencias en las áreas de "Metalurgica" y "Estudios de la Ciencia" y separa "Biología" de "Ecología":

```
> rs[grep("METALURGICA",rs$area.corregida),"area.corregida"] <- "METALURGICA Y CIENCI
> rs[rs$"Area de Estudio" %in% "BIOLOGIA","area.corregida"] <- "BIOLOGIA"
> rs[rs$area.corregida %in% c("ESTUDIOS SOCIALES DE LA CIENCIA","ESTUDIOS DE LA CIENC
```

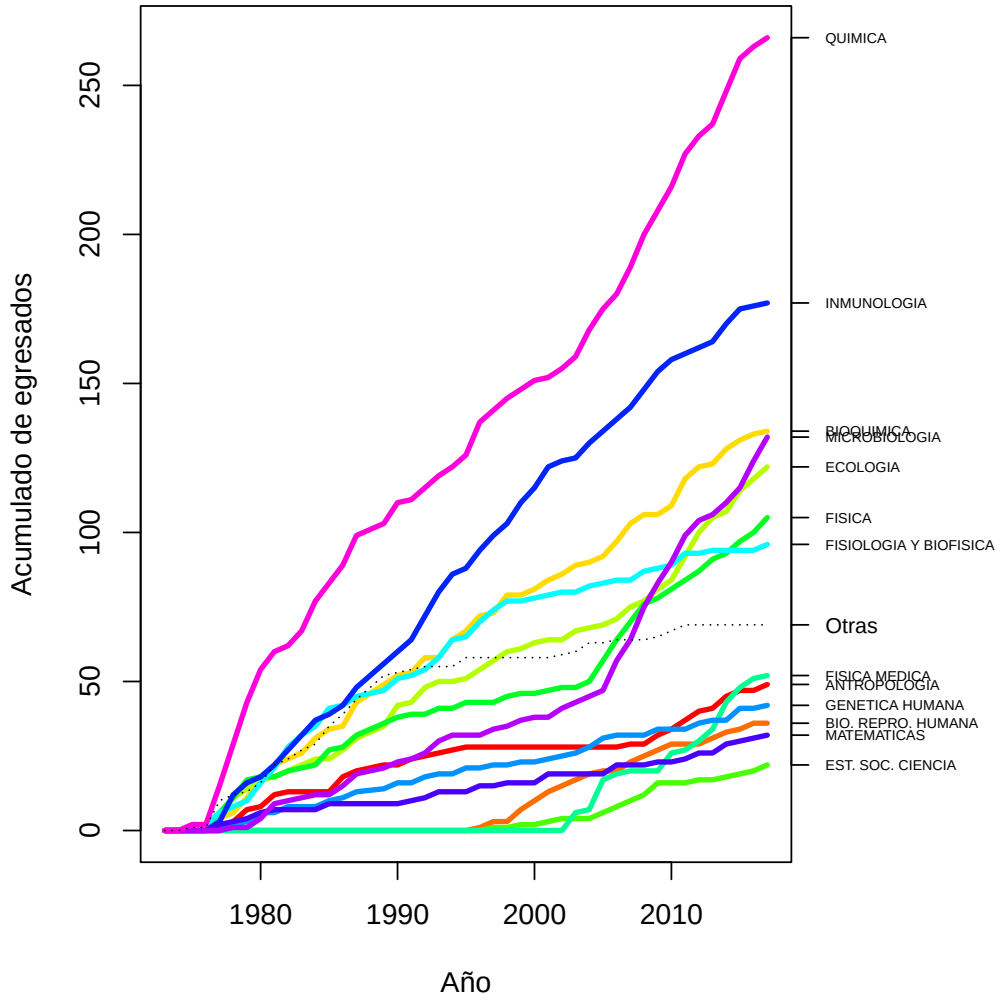
Ahora podemos empezar a analizar los datos.

2.3. Análisis

Primero vamos a visualizar el número acumulado de egresados por año en cada una de las áreas de estudio. Este gráfico permite observar el desarrollo de cada área a través del tiempo, e identificar cambios puntuales en las tendencias generales.

```
> dts <- table(rs$Año,rs$area.corregida)
> ss <- colSums(dts)>19
```

```
> par(mar=c(5,4,0,7))
> matplot(as.numeric(rownames(dts)),
+         apply(dts,2,cumsum)[,ss & !(colnames(dts) %in% "ARTICULO 9")],
+                             type="l",lty=1,lwd=3,col=rainbow(14),
+                             xlab="Año",ylab="Acumulado de egresados")
> axis(4,at=colSums(dts)[ss & !(colnames(dts) %in% "ARTICULO 9")],
+      lab=sub("BIOLOGIA DE LA REPRODUCCION","BIO. REPRO.",
+          sub("ESTUDIOS SOCIALES DE LA","EST. SOC.",
+              colnames(dts)[ss & !(colnames(dts) %in% "ARTICULO 9")])),
+      cex.axis=.5,las=2)
> lines(as.numeric(rownames(dts)),cumsum(rowSums(dts[,!ss])),lty=3)
> axis(4,at=sum(rowSums(dts[,!ss])),"Otras",cex.axis=.75,las=2)
```

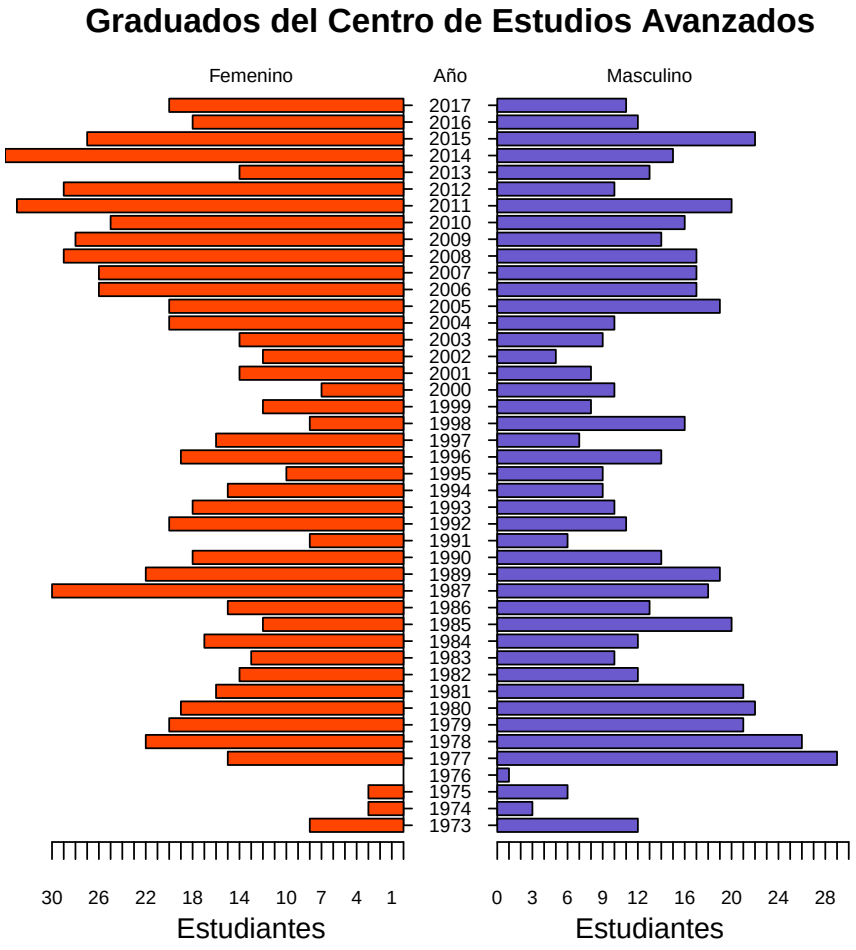


Química, Inmunología y Bioguímica son las áreas que se han mantenido en los primeros lugares con un crecimiento sostenido en el tiempo. Microbiología y Física presentan un claro repunte a partir de 2005 y en la última década hubo más graduados en estas áreas que en los 30 años anteriores. El repunte en Ecología es menos marcado, pero igualmente notable. Hasta 2005

el número acumulado de egresados estaba por debajo de los del área de Fisiología y Biofísica, pero un incremento en los últimos siete años permitieron superar ésta área y alcanzar más de 120 egresados.

Usamos un gráfico de pirámide para ver la proporción de sexos de los egresados por año. Se observa que en la década de 1990 hubo una reducción notable en el número de egresados, y que a partir del año 2001 hay un cambio en la proporción de sexos, con una clara mayoría de mujeres egresadas en los últimos 15 años.

```
> require(plotrix)
> piramide <- with(rs,tapply(Estudiante,list(Año,tolower(Sexo)),length))
> pyramid.plot(piramide[,"femenino"],piramide[,"masculino"],
+              labels=rownames(piramide),gap=4,##xlim=c(-35,-1,1,35),
+              lxcol="orangered",rxcol="slateblue",
+              top.labels=c("Femenino","Año","Masculino"),unit="Estudiantes",
+              main="Graduados del Centro de Estudios Avanzados",labelcex=.65)
[1] 5.1 4.1 4.1 2.1
>
```



El balance total es positivo para el sexo femenino, y se refleja en los dos tipos de títulos otorgados:

```
> table(rs$"Grado Académico",rs$Sexo)
```

```
                            femenino masculino
  DOCTOR EN CIENCIAS        0      138        97
  MAGISTER SCIENTIARUM      1      561       424
  PHILOSOPHUS SCIENTIARUM   0       70        73
```

3. Conclusiones

Este documento muestra un ejemplo muy somero de los análisis posibles con los datos disponibles. La comparación de estas tendencias con las de otros postgrados nacionales o internacionales puede ser de gran utilidad para analizar el impacto y la efectividad de las políticas nacionales en materia de ciencia y tecnología (Medina y Lindorf, 2011).

Referencias

Lemon, J (2006). "Plotrix: a package in the red light district of R". En: *R-News* 6.4, págs. 8-12.

Medina, Ernesto y Helga Lindorf (2011). "Desarrollo De La Ecología En Venezuela: Una Perspectiva Desde El Inicio Hasta La Consolidación De Los Estudios De Postgrado". En: *Ecotrópicos* 24.2, págs. 123-144. URL: http://www.saber.ula.ve/bitstream/123456789/36398/1/Articulo1.pdf.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Wickham, Hadley (2015a). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.1. URL: https://CRAN.R-project.org/package=rvest.

— (2015b). *xml2: Parse XML*. R package version 0.1.2. URL: https://CRAN.R-project.org/package=xml2.