Investigación social con R

Daniela de los Santos

Investigadora | Área de Desarrollo y Género | CIEDUR

Metodologías de investigación en ciencias sociales

(Mis) Primeros pasos en R

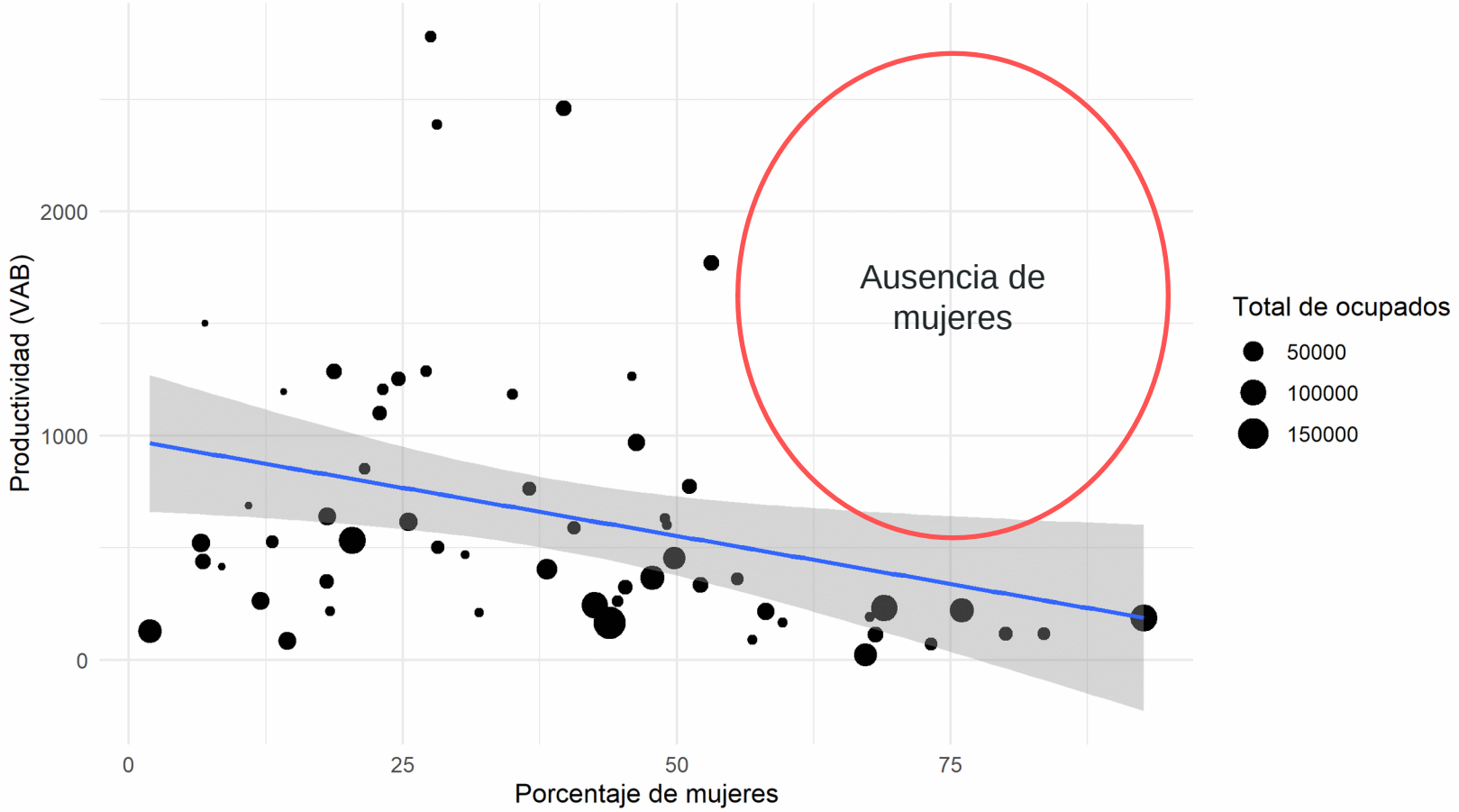
- Formación en procesamiento de datos, pero con herramientas específicas de análisis estadístico.

 En la facultac  En el trabajo

- Empecé a explorar las posibilidades de R cuando necesité generar visualizaciones para relaciones complejas entre algunas variables. Puntualmente, estudiando la relación entre segregación ocupacional y productividad.

Nivel de productividad de sectores económicos (rama de actividad + tamaño de empresa)

Según % de mujeres ocupadas



Fuente: elaboración propia en base a datos de ECH 2014 y BCU

R y el análisis **cuantitativo**

Fuentes de información:

Bases de datos estructuradas (encuestas, censos, GIS)
Datos no estructurados (webscrapping – Big Data)

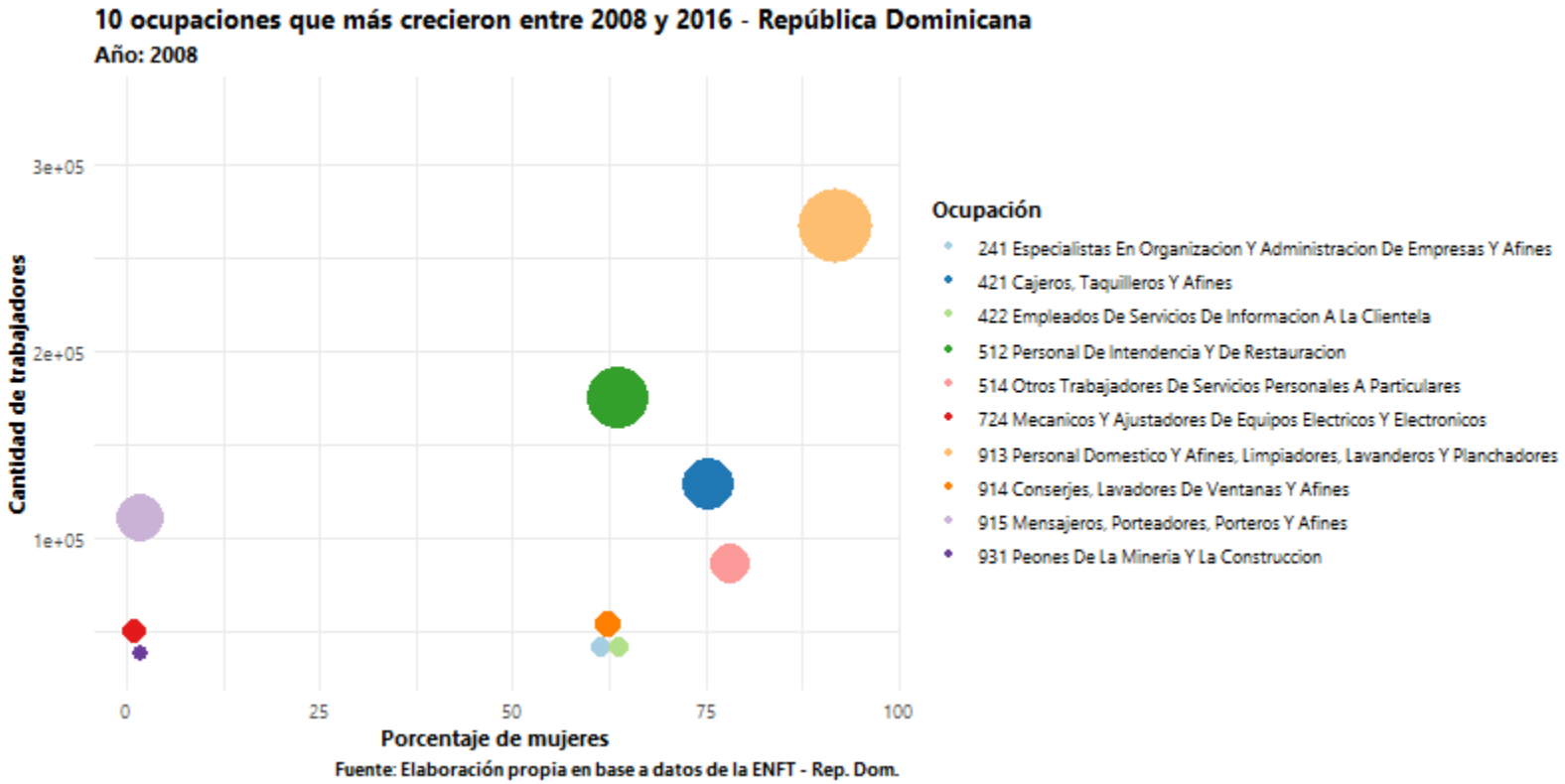
¿Para qué usar R:

- **Importación de datos** (foreign, XML, rjson, readxl)
- **Exploración y procesamiento** (tidyverse, Base R)
- **Modelización** (tradicional y *machine learning*)
- **Visualización** (ggplot2, r2d3, plotly, gganimate, leaflet, tmap y muchas más)
- **Presentación y publicación de informes reproducibles** (RMarkdown), **aplicaciones**

Un ejemplo a partir de datos estructurados:

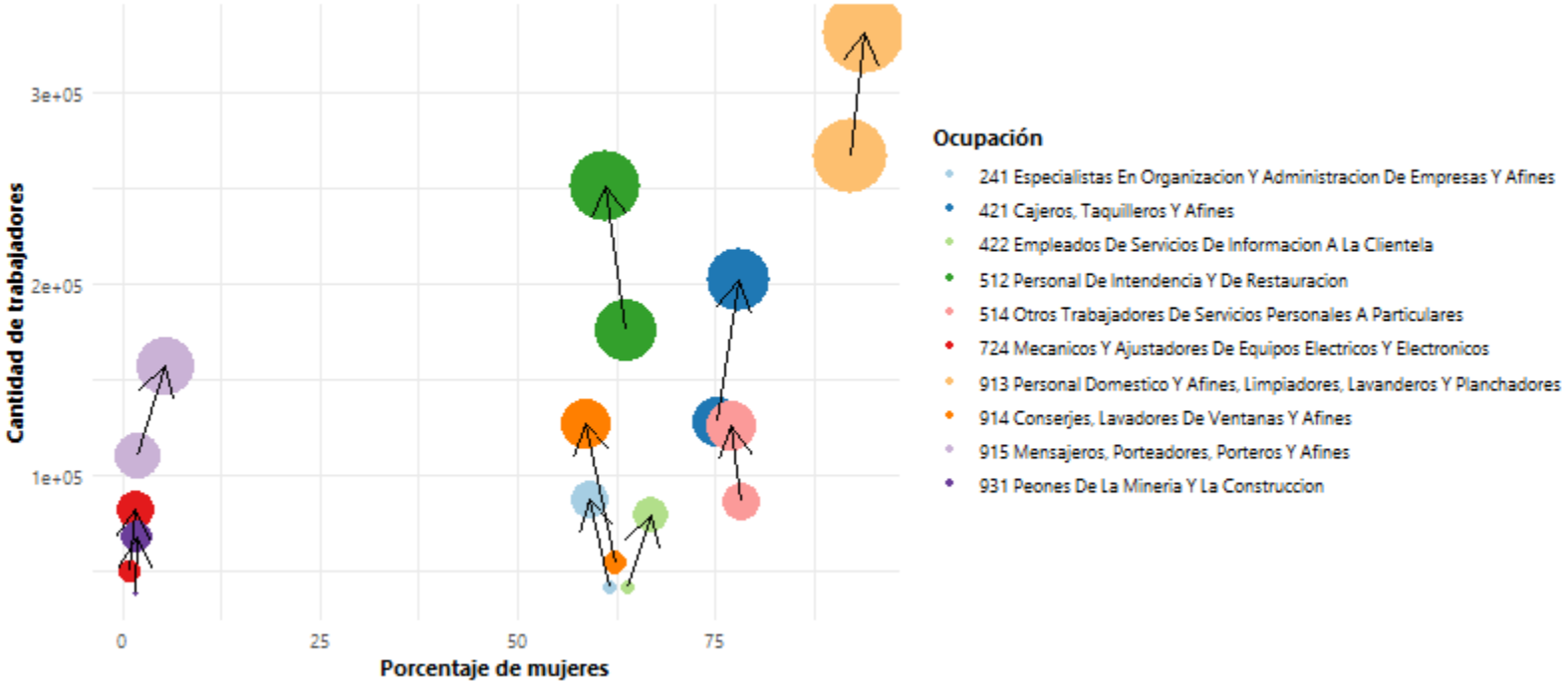
- Estudio sobre segregación ocupacional en República Dominicana: cambios en el mercado laboral entre 2000 y 2016, utilizando la Encuesta Nacional de Fuerza de Trabajo del Banco Central de RD.

- - ¿Qué ocupaciones crecieron? ¿Por qué?
  - ¿Cómo se distribuyen hombres y mujeres entre las ocupaciones en crecimiento?
  - ¿Hay mejoras en la segregación ocupacional?
  - ¿Las mujeres han empezado a insertarse en forma importante en otros campos no tradicionales?


10 ocupaciones que más crecieron entre 2008 y 2016 - República Dominicana
Año: 2008
Cantidad de trabajadores
3e+05
2e+05
1e+05
0
25
50
75
100
Porcentaje de mujeres
Fuente: Elaboración propia en base a datos de la ENFT - Rep. Dom.
Ocupación
241 Especialistas En Organizacion Y Administracion De Empresas Y Afines
421 Cajeros, Taquilleros Y Afines
422 Empleados De Servicios De Informacion A La Clientela
512 Personal De Intendencia Y De Restauracion
514 Otros Trabajadores De Servicios Personales A Particulares
724 Mecanicos Y Ajustadores De Equipos Electricos Y Electronicos
913 Personal Domestico Y Afines, Limpiadores, Lavanderos Y Planchadores
914 Conserjes, Lavadores De Ventanas Y Afines
915 Mensajeros, Porteadores, Porteros Y Afines
931 Peones De La Mineria Y La Construccion

10 ocupaciones que más crecieron entre 2008 y 2016 - República Dominicana

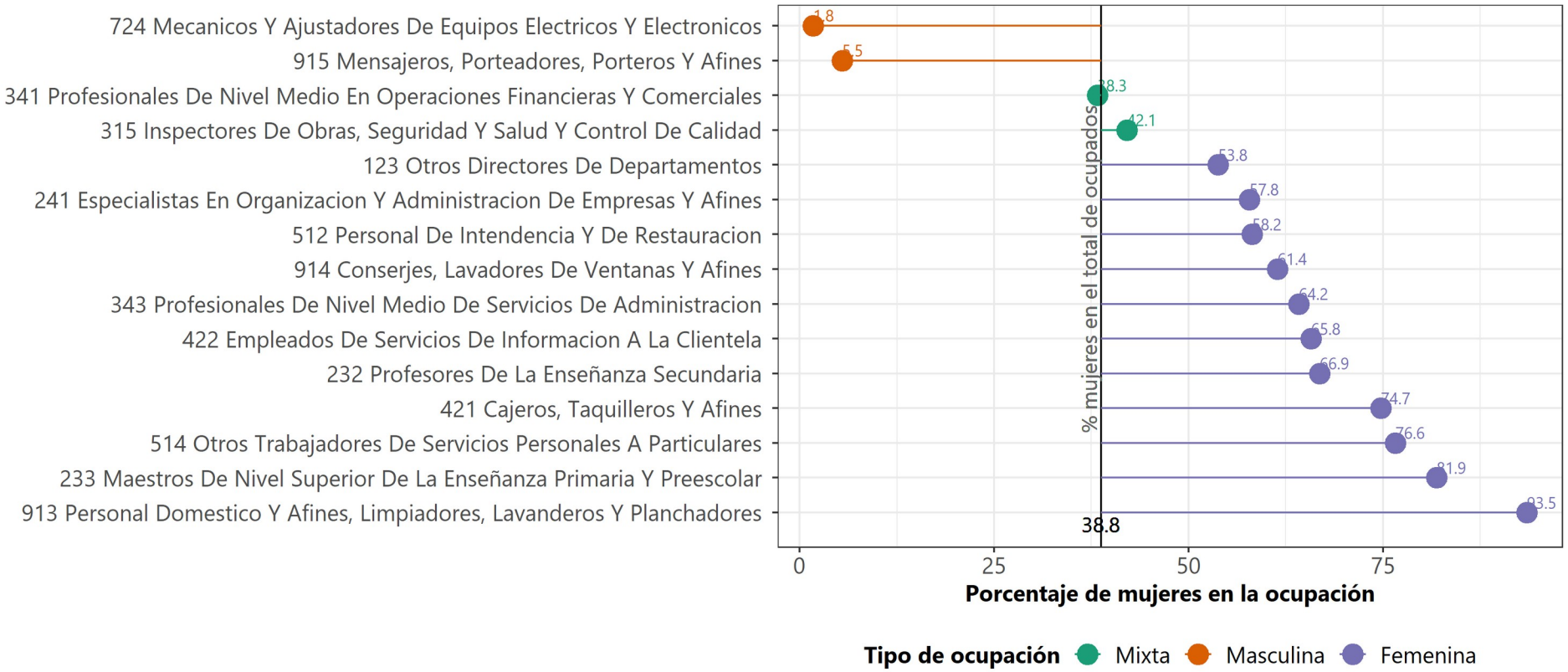
Años 2008 y 2016



Fuente: Elaboración propia en base a datos de la ENFT - Rep. Dom.

Ocupaciones que más crecieron en Rep. Dom.

Entre 2000 y 2016



Elaboración propia en base a datos de las ENFT 2000 y 2016 - Rep. Dom.

R y el análisis **cualitativo**

Fuentes de información:

- Entrevistas
- Grupos focales
- Revisión documental
- Revisión de prensa
- Etnografía (observación participante)

¿Para qué usar R?

- Recolección y pre-procesamiento de información (webscrapping/ APIs)
- Procesamiento y análisis (stringr, stringi, tm)
- Presentación de información (ej.: wordcloud2, ggwordcloud)

Procesamiento y análisis de datos cualitativos

Análisis de texto y contenido: es una aplicación cuantitativa a las herramientas cualitativas clásicas.

“Un análisis sistemático, objetivo y cuantitativo de las características de los mensajes” (Neuendorf 2002)

Este análisis no reemplaza el análisis tradicional cualitativo e interpretativo, pero puede sumar, servir para un análisis global y primario, y facilitar el trabajo del analista para encontrar los puntos clave.

Herramientas que van un poco más allá (IA): sentiment analysis,

Un ejemplo:

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<rss xmlns:atom="http://www.w3.org/2005/Atom" xmlns:slash="http://purl.org/rss/1.0/modules/slash/" version="2.0">
  ▼<channel>
    <title>Lo último</title>
    <description>Descripción</description>
    <pubDate>Tue, 20 Aug 2019 19:07:35 +0000</pubDate>
    <generator>El País Feed Generator (http://www.elpais.com.uy/)</generator>
    <link>http://www.elpais.com.uy/</link>
   ▼<item>
     ▼<title>
        Nahitan Nández y Boca en desacuerdo por un porcentaje del pase al Cagliari
      </title>
     ▼<description>
       ▼<![CDATA[
          El presidente del club argentino, Daniel Angelici, dijo: "Nahitan se habrá olvidado, Boca le había hecho un préstamo de 600 mil dólares q
        ]]>
      </description>
      <pubDate>Tue, 20 Aug 2019 18:50:00 +0000</pubDate>
     ▼<link>
        http://www.elpais.com.uy/ovacion/futbol/nahitan-nandez-boca-desacuerdo-porcentaje-pase-cagliari.html
      </link>
     ▼<guid>
        http://www.elpais.com.uy/ovacion/futbol/nahitan-nandez-boca-desacuerdo-porcentaje-pase-cagliari.html
      </guid>
      <enclosure type="image/jpeg" length="315566" url="http://www.elpais.com.uy/files/rss_thumbnail/uploads/2019/08/12/5d51afca79e18.jpeg"/>
      <subtitle>ARGENTINA</subtitle>
      <category>Fútbol</category>
      <categorySlug>ovacion/futbol</categorySlug>
```

Un ejemplo:

En una revisión de prensa, el trabajo con webscrapping + RSS resulta práctico para no perder tiempo buscando todos los días en todos los portales que nos interesan las palabras claves que nos interesan.

RSS - futuro del trabajo

```
library(rvest)

library(stringr)
library(XML)

    url <- "https://www.elpais.com.uy/rss/"
    download.file(url, "RSSElpais.html", quiet = FALSE, mode = "w", cache
OK = TRUE, extra = getOption("download.file.extra"))
    page <- read_html("RSSElpais.html", encoding = "UTF-8")

 extract<-html_nodes(page,"item")
 html_children(extract[3])

html_text(html_nodes(extract[3],"entity"))

    #FUNCION PARA CARGAR TITULOS Y FUENTES
    my.fun<- function(page=page){
      extract<-html_nodes(page,"item")
      N<-length(extract)
      my.title<- vector("list",N)
      my.source<- vector("list",N)
      my.entity<- vector("list",N)

      for(i in 1:length(extract)){
        my.source[[i]]<-html_text(html_nodes(extract[i],"source"))
        my.title[[i]]<-html_text(html_nodes(extract[i],"title"))
        my.entity[[i]]<-html_nodes(extract[[i]],"entity")
      }
      myObjects <- NULL
      myObjects[[1]] <- my.source
      myObjects[[2]] <- my.title
      myObjects[[3]] <- my.entity
myObjects
     }

    #EXTRAER DATA
    results <- my.fun(page)

grep(pattern="automatización", results)
```

```
## integer(0)
```

Revisión de prensa – Futuro del trabajo

- Actores sindicales

- Actores empresariales

Fuente: elaboración propia.

Conclusiones

- R tiene muchas aplicaciones en la investigación social, y no solo en el campo “cuantitativo”.
- Los mecanismos de *machine learning,* poco explorados desde las ciencias sociales tradicionales (en parte porque “predicen”, pero no necesariamente “explican”), abren un campo nuevo de posibilidades que vale la pena explorar!

Gracias!

Contacto: dlsantos.daniela@gmail.com


danidlsa