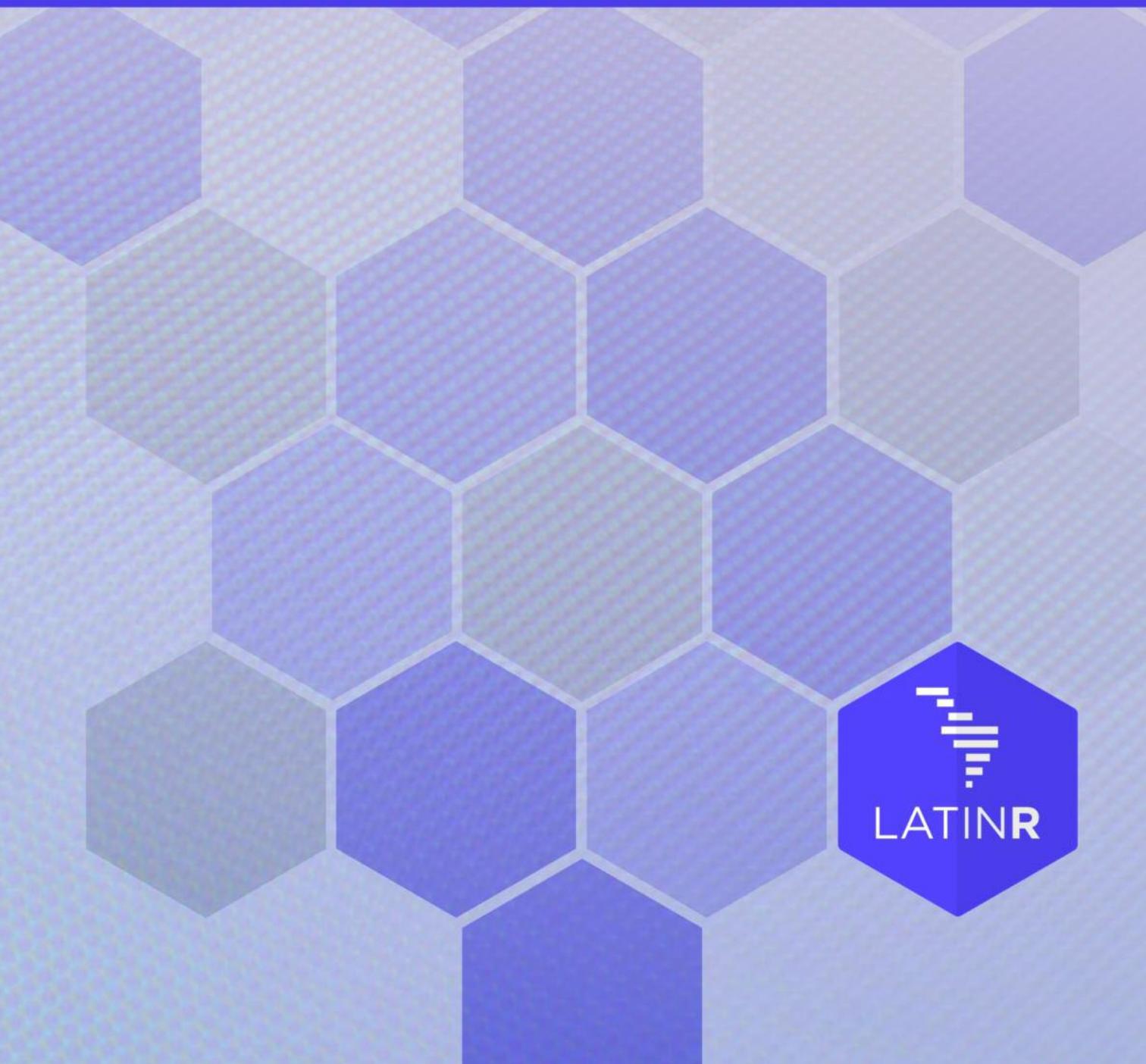


Actas de LATINR 2021

Tercera Conferencia Latinoamericana
sobre el uso de R en Investigación + Desarrollo

Editoras: Yanina Bellini Saibene, Riva Quiroga,
Natalia da Silva



Diseño Gráfico
Dis. Gráf. Francisco Etchart

Febrero de 2022

Equipo

Chairs

Yanina Bellini Saibene

- Instituto Nacional de Tecnología Agropecuaria
- R-Ladies Santa Rosa, Argentina + Global Team
- MetaDocencia

Natalia da Silva

- Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República
- R-Ladies Montevideo, Uruguay

Riva Quiroga

- Facultad de Letras, Universidad Católica de Chile
- R-Ladies Santiago / Valparaíso, Chile

Comité Científico

Laura Ación

- LatinR Inaugural Local Chair / UBA-CONICET, Argentina

Marcela Alfaro

- Department of Statistics, University of California, Santa Cruz

Virginia Alonso

- CONICET

Ignacio Alvarez-Castro

- IESTA-Universidad de la República

Zulema Bazurto Blacio

- Universidad de Guayaquil

Mathias Bourel

- Facultad de Ingeniería, UDELAR

Naim Bro

- Instituto Milenio Fundamento de los Datos (IMFD)

Marina Cock

- Universidad Nacional de La Pampa

Juan Cruz Rodriguez

- FAMAF - Universidad Nacional de Córdoba

Jairo Cugliari

- ERIC-Université Lumière Lyon 2

María José García-Zattera

- Departamento de Estadística, Pontificia Universidad Católica de Chile

Adriana Gili

- Universidad Nacional de La Pampa

Juan José Goyeneche

- IESTA-Universidad de la República

Carmen Le Foulon

- Centro de Estudios Públicos (CEP)

Vianey Leos

- North Carolina State University

Natalia Moradeira

- UNSAM - CONICET

Leonardo Moreno

- IESTA-Universidad de la República

Victoria O'Donell

- Universidad de Buenos Aires

Ricardo Olea

- Pontificia Universidad Católica de Chile

Pamela Pairo

- Comisión Nacional de Actividades Espaciales (CONAE)

Lucia Rodriguez Planes

- Universidad Nacional de Tierra del Fuego - CONICET

Germán Rosati

- IDAES-UNSAM / CONICET, Argentina

Andrea Sánchez-Tapia

- Laboratorio Independiente de Informática de la Biodiversidad

Alejandra Tapia Silva

- Universidad Católica del Maule

Francisco Urdinez

- Pontificia Universidad Católica de Chile

Inés Varas

- Pontificia Universidad Católica de Chile

Luis D. Verde Arregoitia

- Instituto de Ecología A.C.

Comité Organizador**Monica Alonso**

- RLadies Buenos Aires
- Banco Ciudad

Linda Cabrera Orellana

- R-Ladies Guayaquil
- Universidad Técnica de Machala

Samuel Carleial

- University of Konstanz (GER)

Joselyn Chavez

- R-Ladies Cuernavaca
- Bioconductor
- CDSB México

Paola Corrales

- Universidad de Buenos Aires, Argentina
- R-Ladies BsAs, Argentina

Andrea Gómez Vargas

- Universidad de Buenos Aires
- R Ladies Buenos Aires

Beatriz Milz

- Universidad de São Paulo
- R-Ladies São Paulo

Macarena Quiroga

- CIIPME-CONICET

Pamela Pairo

- Comisión Nacional de Actividades Espaciales (CONAE)

Luis D. Verde Arregoitia

- Instituto de Ecología A.C.
- UNaHur

Prólogo

La cuarta edición de LatinR fue planificada desde un inicio para ser realizada a de forma virtual entre el 10 y 12 de noviembre de 2021, atendiendo a que la situación sanitaria hacía difícil la realización de eventos presenciales.

En esta ocasión hubo dos charlas invitadas:

- Beautiful Tables in R by Tom Mock
- R Development Guide: Motivation, First Draft, and the Way Forward by Saranjeet Kaur

Al igual que en ediciones anteriores, en las semanas previas a la conferencia se organizaron tutoriales en línea, los que fueron hospedados por capítulos de RLadies y grupos de usuarios de R de la región. Los cuatro tutoriales ofrecidos este año atendieron a los resultados de la encuesta realizada a la comunidad sobre sistemas de interés. Tres tutoriales se impartieron en español y uno en portugués:

- Desarrollo de Aplicaciones de Shiny a nivel comercial, a cargo de Oriol Senan, Agustín Pérez Santangelo y Federico Rivadeneira
- "Introducción al análisis de redes con R", a cargo de María Ramos
- "gg+: paquetes para extender las capacidades de ggplot2", a cargo de Luis Verde.
- "Modelos de equações estruturais (Structural Equation Models; SEM) in R", a cargo de Samuel Carleial

En esta cuarta edición se alcanzó la cifra récord de 1000 personas inscritas. Se recibieron un total de 64 propuestas, de las cuales 48 se incluyeron en el programa final, en las modalidades charla regular y charla relámpago. Al igual que el año anterior, las charlas fueron grabadas y publicadas en los días previos a la conferencia en el canal de YouTube de LatinR, ordenadas en listas de reproducción según las sesiones temáticas del evento. Estas sesiones, que se realizaron de forma sincrónica, tenían por objetivo generar una instancia de diálogo en torno a los temas de las presentaciones. Al inicio de cada sesión, los autores y las autoras de los distintos trabajos, hacían un breve resumen de su trabajo, y luego se daba paso a las preguntas de la audiencia y del resto de los participantes.

Agradecemos a nuestros sponsors AppSilopn, RStudio, RConsortium y a la Universidad Nacional de la República y al INTA por su apoyo.

Índice

Desarrollo de paquetes	9	Análise exploratória de despesas com educação: avaliando impactos em um indicador de rendimento educacional Fernando Barbalho, Tiago Pereira, Lucas Leite, Jordão Gonçalves
Divide y vencerás: de polAr al 'polArverse' Juan Pablo Ruiz Nicolini		
Paquete Calidad, para la evaluación de la precisión de estimaciones provenientes de encuestas de hogares Ricardo Pizarro and Klaus Lehmann		Desarrollo de una herramienta predictiva de calidad de agua para la gestión: modelos de machine learning + shiny Andrea Cardoso, Matías Muñoz Wolf, Juan José Lagomarsino, Lucía González-Madina, Mathias Bourel, Juan Pablo Pacheco, Rafael Terra, Gustavo Mendez, Néstor Mazzeo, Gonzalo Perera and Carolina Crisci
The {botmaker}: automatically build R-based bots, the result of creating the @RStatsJobsBot Twitter bot Juan Cruz Rodriguez		
speech: extracción, disponibilización y análisis de discursos parlamentarios en Uruguay Elina Gómez , Nicolás Schmidt		
Paquetes, modelos y aplicaciones en ciencias	17	Aplicaciones en salud pública 36
AMALIA: R, shiny y minería de texto para el análisis masivo de archivos de la dictadura uruguaya Elina Gómez		Uso de R en la validación de un modelo predictivo con aplicación a la enfermedad del asma Alejandra Tapia Silva
ANÁLISIS DE LA RED DE INVESTIGADORES DEL IESTA Pablo Mones and Ramón Álvarez-Vaz		Ciencia de datos con R con impacto en salud pública: una experiencia de uso de tidyverse para la detección de embarazos Sabrina Laura López, Carolina Mengoni Goñalons, María Cristina Nanton and Manuel Rodríguez Tablado
Políticas económicas frente al COVID-19: índice de Shannon para la diversidad de agendas de gobiernos locales Enrique García-Tejeda		¿Por qué los funcionarios públicos de salud deberían saber R para analizar sus datos? Paulo Villarroel Tapia
Elaboración de redes analíticas interactivas con el paquete netCoin de R Modesto Escobar and Cristina Calvo López		Estimating remaining life expectancy free of anxiety/depression in Argentina: trends and application of an algorithmic stepwise decomposition for demographic change, 2005-18 Octavio Nicolas Bramajo
Retos y R en conteos rápidos electorales Maria Teresa Ortiz Mancera, Michelle Anzarut and Luis Felipe González Pérez		Enseñanza de R 46
ANÁLISIS DEL TIPO DE CONSUMO CULTURAL SEGÚN NIVEL SOCIOECONÓMICO DE LOS BENEFICIARIOS DEL PROGRAMA PASE CULTURAL Jonathan A. Modernel, Magdalena Cornejo		Scaling feedback using learnr and gradethis in a introductory R course Beatriz Milz y Fernando Correa
		Aplicaciones shiny para modelos de crecimiento de ecología de poblaciones: una propuesta

simple y no simplista para animar al uso de R en cursos introductorios	Un cuento digital desde R: cómo crear un relato situado con Leaflet	70
Lucía Rodríguez Planes	Natalia Morandeira	
Karel la robot enseña R: un paquete para la enseñanza de programación	Periodismo de datos, datos abiertos y visualización	71
Marcos Prunello		
Un conjunto de paquetes para generar tutoriales interactivos para enseñar R	JFemDados: democratizando informação	
Yanina Bellini Saibene	Matheus Valentim e Marcello Filgueiras	
Comunidad de R	A Base dos Dados+: acesso fácil a dados públicos de qualidade	
56	Rodrigo Dornelles , Matheus Valentim , Fernanda Scovino , Pedro Cavalcante	
Conociendo el camino para aprender a usar R en Latinoamérica: desafíos para promover la inclusión y diversidad	Alavancando o poder do RMarkdown com as linguagens da Web e D3.js para produzir histórias de dados envolventes sobre finanças públicas	
Claudia Alejandra Huaylla, Paola Corrales, Andrea Gómez Vargas, Joselyn Chávez, Denisse Fierro Arcos, Virginia García Alonso	Tiago Pereira, Fernando Barbalho, Jordao Goncalves and Lucas Leite	
Juntas podemos más, corta historia de cómo la pandemia nos incentivó a colaborar	Collect and use open access World Bank data to know your country	
Denisse Fierro Arcos, Danisse María Carrascal Polo, Linda Cabrera Orellana, Mary Jane Rivero Morales	Bruno Thiago Tomio	
Investigación y comunicación de resultados	R en producción, computación y flujos de trabajo	80
61		
Tablas reproducibles, presentables y con formato numérico local con gtsummary	R en producción: aprendizajes, retos y mejores prácticas	
Eva Retamal Riquelme	Ángel Escalante , Nancy Morales	
Uso de R y Youtube para reporte de protocolos: experiencia en laboratorio de física de suelos	Analogsea: using R for big data analytics	
Cristina Contreras, Sara Acevedo, María Jesús Melej, Edouard Acuña, Carolina Giraldo, Carlos Ávila and Carlos Bonilla	Mauricio Vargas, Scott Chamberlain, Hadley Wickham and Bob Rudis	
Un viaje a la ecología del movimiento a través de la minería de texto	IBM Cloud Functions com R	
Rocío Joo, Simona Picardi, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Vilma S. Romero-Romero, Mathieu Basille	Thiago Pires	
Datos espaciales	Más velocidad y menos colapsos: preprocesamiento de archivos con utilidades del sistema operativo	
66	Luis Verde	
Soy naturalista y quiero pasear en mi país, ¿dónde hay más oportunidades de llenar vacíos de información?	Uso de R como front-end en un datawarehouse de gestión académica universitaria	
Florencia Grattarola, Juan Manuel Barreneche	Daniel Alessandrini, Pablo Martínez, Óscar Montañés, Juan Manuel Serralta	

Desarrollo de paquetes y modelos

missMSPC: un paquete de herramientas gráficas para aplicar MSPC con datos faltantes

Julia Inés Fernández , Diego Marfetán Molina , José Alberto Pagura , Marta Beatriz Quagliino

QR: un paquete para la factorización QR sin rotación

Juan Claramunt González

Estimación de un modelo computacional mediante computación Bayesiana Aproximada

Juan Ignacio Baccino Costa , Mauro Loprete , Alvaro Valiño , Daniel Ciganda

Feature and variable selection in complex data classification

Manuel Oviedo de la Fuente, Manuel Febrero Bande

Optimizando @RStatsJobsBot: un modelo de aprendizaje automático para clasificar tweets de ofertas de empleo

Martin Rodriguez Nuñez, Juan Cruz Rodriguez

Cómo implementar algunos modelos de imputación múltiple para datos de panel

Ramón Álvarez-Vaz, Diana Del-Callejo-Canal, Margarita Edith Canal-Martínez, Elena Vernazza and Alar Urruticoechea

Años potenciales de vida perdidos por siniestros de tránsito en Uruguay

Gonzalo De armas , Mauro Loprete , Ramón Álvarez-Vaz

Desarrollo de paquetes y aplicaciones en ciencia

agromet: un paquete para el análisis de datos meteorológicos

Natalia Gattinoni, Paola Corrales, Elio Campitelli, Yanina Bellini y Gabriel Rodriguez

Rocc: gestão e análise de dados de ocorrências de espécies

Sara Mortara and Andrea Sánchez-Tapia

94

Interoperabilidad y grandes volúmenes de datos: como potenciar el diseño de políticas públicas basada en evidencia

Juan Pablo Zumárraga, Fernando Ashbey, Julieta Coll and Adrian Ibarra

Aplicación de R para analizar la perspectiva del consumidor sobre la consistencia de alimentos

Franco Della Fontana, Margarita Armada and María Cristina Goldner

Ensembles conformacionales de Proteínas intrínsecamente Desordenadas moldean las velocidades de evolución dando origen a patrones conformacionales

Julia Marchetti, Nicolas Palopoli, Alexander Miguel Monzon, Diego Javier Zea, Maria Silvina Fornasari, Silvio C.E. Tosatto and Gustavo Parisi

Sesión Desarrollo de paquetes

Divide y vencerás: de {polAr} al 'polArverse'

Anónimo

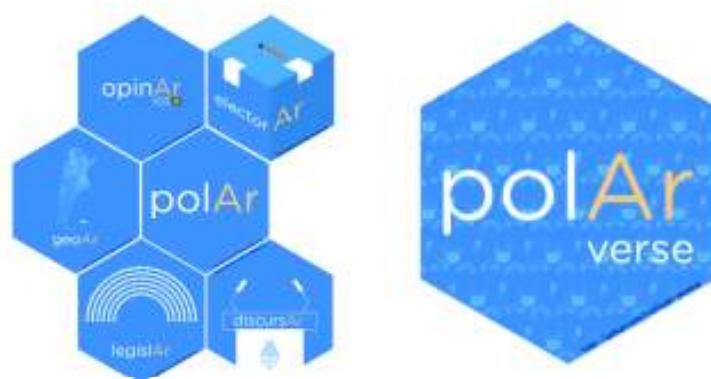
Palabras clave: política - argentina - elecciones - discursos - votaciones - opinión - geo

Abstract

{polAr} nació como un paquete que buscaba facilitar el acceso y herramientas para el análisis de datos electorales en Argentina (Ruiz Nicolini 2020). El 16 de mayo de 2020 el proyecto en desarrollo viajaba a los servidores del *Comprehensive R Archive Network (CRAN)*. Con el tiempo se fueron integrando otros flujos de trabajo y fuentes de datos (geográficos, discursos presidenciales o votaciones legislativas), con un correlativo incremento de dependencias y dificultad de administrar cada vez más funciones.



En octubre de 2020, durante el panel *Desarrollo de Paquetes* del III LatinR¹, se nos consultó sobre estrategias para la administración, desarrollo y mantenimiento de paquetes; y, particularmente, sobre qué hacer cuando un paquete crece mucho. Sí esa discusión sembró la semilla del '*polArverse*', la decisión de CRAN de archivar {polAr} abrió una ventana de oportunidad de avanzar en un proyecto ambicioso y extenso, pero compuesto por pequeños módulos. Un universo de paquetes para la política Argentina: divide y reinarás.



¹LatinR 2020 - Viernes 9 de Octubre - Desarrollo de paquetes: <https://www.youtube.com/watch?v=UYvSv8StDa8&t=10872s>

*polArverse: un universo de paquetes en desarrollo*²

El objetivo de la partición de {polAr} en un conjunto de paquetes más pequeños es brindar una multiplicidad de herramientas para el análisis político de Argentina, al tiempo de facilitar la administración, mantenimiento y crecimiento de los posibles flujos de trabajo que lo integran.

Hasta el momento *polArverse* esta compuesto por:

- **{geoAr}** - que facilita el acceso a geometrías de Argentina a distintos niveles (provincias, departamentos, radios censales), el diseño de grillas como si fueran mapas (para usar con geofacet) y otras herramientas para el trabajo geo.
- **{electorAr}** - que facilita el acceso a datos electorales de Argentina y **{legislAr}** - para datos de votaciones en las cámaras legislativas de Argentina-, ambos basados en el trabajo previo de Andy Tow³.
- **{opinAr}** - que facilita el acceso a datos de opinión pública del *Indice de Confianza en el Gobierno*⁴ y herramientas para trabajar con ellos.
- **{discursAr}** es un paquete aún en desarrollo que procura facilitar el acceso a discursos presidenciales. Además de los discursos de inauguración de sesiones que integraban la vieja versión disponible en {polAr}, en este paquete se está trabajando sobre discursos de gestión de presidentes contemporáneos.

Por último **{polArverse}**, basado en la idea y código de **tidyverse**, es un meta paquete auxiliar que permite el acceso a todo el universo **polAr** al mismo tiempo.

```
library(polArverse)
```

Referencias

- 10 Ruiz Nicolini, Juan Pablo. 2020. "polAr: Argentina Political Analysis." <https://github.com/electorArg/polAr>.
- . 2021a. "discursAr: Argentina's Presidential Speeches Toolbox." <https://github.com/PoliticaArgentina/discursAr>.
- . 2021b. "electorAr: Toolbox for Argentina's Electoral Data." <https://politicaargentina.github.io/electorAr/>.
- . 2021c. "geoAr: Argentina's Spatial Data Toolbox." <https://politicaargentina.github.io/geoAr/>.
- . 2021d. "legislAr: Argentina's Legislative Data and Tools." <https://politicaargentina.github.io/legislAr/>.
- . 2021e. "opinAr: Argentina's Public Opinion Toolbox." <https://politicaargentina.github.io/opinAr/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.

²Se está avanzando además en el desarrollo de un segundo paquete auxiliar, {polArViz}, que incluye todas funciones de visualización de datos del resto de los paquetes.

³Para {electorAr} utilizamos bases de datos y tablas disponibles en el *Atlas Electoral*. Para {legislAr} utilizamos los datos compartidos en el proyecto *Década Votada*.

⁴Escuela de Gobierno. Universidad Torcuato Di Tella www.utdt.edu/icg

Paquete Calidad, para la evaluación de la precisión de estimaciones provenientes de encuestas de hogares.

Anónimo

Abstract En base al estandar de calidad publicado por el Instituto Nacional de Estadísticas de Chile (INE), se ha desarrollado el paquete “Calidad”, para la implementación de dicho estandar para la evaluación de la precisión de las estimaciones provenientes de encuestas de hogares.

Palabras clave: encuestas - hogares - calidad estadística

Modalidad presentación: Comunicación Oral

Origen

El desarrollo del paquete “Calidad”, se da en el marco del estándar de calidad estadística definido por el Instituto Nacional de Estadísticas de Chile (INE), para determinar qué tan precisa y confiable es la información de encuestas de hogares que se publica, respecto a los atributos de la población que se pretende caracterizar.¹

La relevancia de este estándar y su implementación permitirá a la oficina de estadísticas de Chile estandarizar sus rutinas de control de calidad de los datos publicados y ofrecer un modelo de control de calidad de las estimaciones, que además, puede ser utilizado por otras instituciones tanto a nivel nacional como internacional.

El objetivo inicial del paquete es entregar a los usuarios de las estadísticas oficiales, tanto de las instituciones gubernamentales como investigadores y estudiantes, una herramienta de código abierto y sencilla utilización, que permita implementar de manera práctica los criterios de calidad estadística definidos.

Para lograr este objetivo y alineado a la política institucional del INE Chile, de impulsar el trabajo con R como principal herramienta para el procesamiento y análisis de datos, se han construido en entorno de R los criterios de calidad propuestos por el estándar.

Este paquete permite el trabajo con diversas encuestas de relevancia nacional como:

- La Encuesta de Caracterización Socioeconómica Nacional (CASEN)
- La Encuesta Nacional de Empleo (ENE)
- La Encuesta Suplementaria de Ingresos (ESI)
- La Encuesta de Presupuestos Familiares (EPF)
- La Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC)

Entre otras encuestas con diseño estadístico multietápico, no solo de Chile.

Es relevante mencionar que el paquete “Calidad” usa en su base el paquete “Survey” de Thomas Lumley, utilizado para el trabajo con encuestas con diseño complejo.

Principales características

“Calidad” posee dos familias de funciones:

- `create_`: permiten crear los insumos para la evaluación del estándar
- `evaluate_`: permiten ejecutar la evaluación del estándar

Con las familia “create” Podemos hacer los siguientes cálculos:

1. `calidad::create_mean`: calcular la media (ingreso)
2. `calidad::create_prop`: proporción o razón (ocupación)
3. `calidad::create_tot`: conteo de unidades (ocupación)
4. `calidad::create_tot_con`: suma de variables continuas (ingreso)
5. `calidad::create_median`: mediana (ingreso)

¹<https://www.ine.cl/docs/default-source/institucionalidad/buenas-pr%C3%A1cticas/clasificaciones-y-estándares/est%C3%A1ndar-evaluaci%C3%B3n-de-calidad-de-estimaciones-publicaci%C3%B3n-27022020.pdf>

A continuación un ejemplo, calculamos la media de la edad en la Encuesta Nacional Urbana de Seguridad Ciudadana 2019 (ENUSC):

```
library(survey); library(calidad)

dc <- svydesign(ids = ~Conglomerado, strata = ~VarStrat, weights = ~Fact_Pers, data = enusc)
options(survey.lonely.psu = "certainty")

calidad::create_mean(rph_edad, disenio = dc)

##           mean      se   gl     n   coef_var
## rph_edad 44.27603 0.2804819 486 24465 0.006334848
```

Como se puede observar, adicionalmente a la estimación se nos entrega el Error Estandar (SE), los Grados de Libertad (GL), la cantidad de casos muestrales (n) y el coeficiente de variación (coef_var). Estos son los insumos que serán evaluados con las funciones de la siguiente familia

Los argumentos recibidos por estas funciones son:

- var** = Variable de interés, ej: desocupación.
- denominador**² = denominador de la división en estimaciones de razón, ej: mujeres ocupadas / hombres ocupados.
- dominios** = variables de desagregación, ej: sexo+region.
- subpop** = Posibles variables de sub poblaciones, fuerza de trabajo.
- disenio** = diseño complejo generado con el paquete survey.
- ci** = para obtener intervalos de confianza.

Como ya se mencionó con la segunda familia de funciones se pueden evaluar los insumos creados. Existe una función de evaluación para cada tipo de estimación:

- calidad::evaluate_mean**
- calidad::evaluate_prop**
- calidad::evaluate_tot**
- calidad::evaluate_tot_con**
- calidad::evaluate_median**

Estas funciones se aplican sobre los insumos generados por las funciones *create*, obtenemos 3 columnas nuevas de evaluación y una cuarta columna que nos advierte si la estimación es *fiable*, *poco fiable* o *no fiable*.

A continuacion un ejemplo de evaluación, evaluando un estimador de media con “evaluate_mean”:

```
calidad::evaluate_mean(calidad::create_mean(rph_edad, disenio = dc))
```

```
##           mean      se   gl     n   coef_var      eval_n      eval_gl
## rph_edad 44.27603 0.2804819 486 24465 0.006334848 n suficiente gl suficiente
##           eval_cv calidad
## rph_edad cv <= 15  fiable
```

Obtenemos 3 columnas nuevas de evaluación y una cuarta columna que nos advierte si la estimación es fiable.

Próximos pasos

Actualmente el paquete se encuentra en el desarrollo de nuevas funcionalidades, orientadas a flexibilizar los análisis según las preferencias de los investigadores, así como para lograr ser útil a nivel Latinoamericano, para esto se está colaborando con la Comisión Económica Para Latinoamérica (CEPAL) en los ajustes y funcionalidades necesarias. Esperamos que “calidad” sea útil para diferentes oficinas nacionales de estadísticas y usuarios independientes.

²disponible solo para estimadores de razón

The `{botmaker}`: Automatically build R-based bots, the result of creating the @RStatsJobsBot Twitter bot.

Juan Cruz Rodriguez^a

^a*FAMAF, Universidad Nacional de Córdoba, Argentina*

Abstract

Key words: Automation, Github Actions, Free, Server

Twitter is one of the social networks most used by the R users community. And possibly it is the social network that offers the greatest flexibility for its programmatic access (`{rtweet}`) (Kearney 2019)). In this regard, Twitter bots result as an excellent tool to promote our product or tool. However, compared to other topics or programming languages, it is not usual to find a wide variety of bots related to R. This is not due to an increased difficulty itself, but rather due to unfamiliarity with automated bots in R.

In this flash talk, I will show the learning path I took to bring an idea to @RStatsJobsBot (<https://twitter.com/RStatsJobsBot>), a Twitter bot that currently has over 900 followers. The @RStatsJobsBot runs entirely on R and is deployed and continuously running on Github Actions at no cost.

In this learning path, I noticed a need for a tool that eases the deployment of an R script as a bot. As a result, I built the `{botmaker}` (<https://github.com/jcrodriguez1989/botmaker>), an R package that allows deploying an R script as a Github Actions scheduled job. During this talk, I will also present a complete usage example of the `{botmaker}`, which will teach how an R script can be turned into an automatically running R bot with a few lines of codes.

References

- Kearney, Michael W. 2019. “Rtweet: Collecting and Analyzing Twitter Data.” *Journal of Open Source Software* 4 (42): 1829. <https://doi.org/10.21105/joss.01829>.

Email address: jcrodriguez@unc.edu.ar (Juan Cruz Rodriguez)

speech: extracción, disponibilización y análisis de discursos parlamentarios en Uruguay

Elina Gómez , Nicolás Schmidt

Palabras clave: discursos parlamentarios - ocr - minería de texto

Introducción

speech es un paquete que permite convertir diarios de sesiones del parlamento uruguayo en formato PDF (URL o archivo local) a bases de datos ordenadas en la que cada fila es la intervención de cada uno/a de los/las legisladores/as que interviene en esa sesión. Se presentarán tres ejes de trabajo entorno al paquete y potencialidades.

Ejes de trabajo:

- Extracción:

El paquete desarrollado incluye un conjunto de funciones que permiten extraer el texto de las menciones parlamentarias incluidas en los diarios de sesión a partir de técnicas de OCR (reconocimiento óptico de caracteres), de forma ordenada (desagregada o agrupada) en formato data.frame y recuperando información anexa como el nombre de el/la legislador/a, sexo, número de legislatura, fecha, cámara a la que pertenece el documento. Asimismo, dado que los diarios de sesión muchas veces son imágenes escaneadas y que en el proceso de OCR se puede perder o dañar la información recuperada, el paquete provee otro conjunto de funciones que ayudan a mejorar estos problemas (`speech_check()`, `speech_legis_replace()`), así como indicadores que dan cuenta la calidad de la recuperación. Por su parte, a partir de la integración con otro paquete (`puy`) que contiene información sobre políticos/as uruguayos/as, es posible anexar integrar fácilmente información sobre el partido político al que pertenece.

- Disponibilización:

El paquete se encuentra en CRAN para su instalación y uso por parte de usuarios/as de R, sin embargo para ampliar la disponibilidad se ha desarrollado una aplicación Shiny que permite la descarga de bases de datos de menciones a partir de introducir URL y/o archivos locales que contengan diarios de sesiones en formato PDF. La misma permite descargar las menciones de forma agrupada o desagregada, y contiene todas las variables anexas para favorecer el análisis de los datos textuales.

- Análisis:

Por último, se encuentran en desarrollo varias líneas de análisis a partir de la información que se obtiene con el paquete `speech`. Por un lado, la clasificación automática de temas a partir de la aplicación de modelos y algoritmos de aprendizaje automático, que permiten un análisis agregado a partir de las variables que recupera el paquete. Por otro lado, se proyecta desarrollar, a partir de una base de datos histórica y acumulada de intervenciones parlamentarias construida a partir del paquete, una aplicación que permita el procesamiento y visualización interactivas mediante técnicas de minería de texto, e incluyendo filtros dinámicos según variables de agregación recuperadas.

Elina Gómez
UMAD (FCS) - Udelar
elina.gomez@cienciassociales.edu.uy

Nicolás Schmidt

UMAD (FCS) - UdelarR

nicolas.schmidt@cienciassociales.edu.uy

Sesión Paquetes , modelos y aplicaciones en ciencias

AMALIA: R, Shiny y minería de texto para el análisis masivo de archivos de la dictadura uruguaya

Elina Gómez

Palabras clave: minería de texto - shiny - memoria histórica

Introducción

AMALIA es una aplicación que surge como una iniciativa en el marco del proyecto CRUZAR (*Sistema de Información de Archivos del Pasado Reciente*), y que a través de técnicas de minería de texto, busca aportar en el análisis masivo e interactivo de documentos de la dictadura uruguaya (1973-1985) que han sido digitalizados y convertidos a texto mediante técnicas de OCR. La misma ha sido desarrollada utilizando el lenguaje R y se encuentra en su versión de prueba. Se nutre de más de 100000 imágenes que conforman el denominado *Archivo Berruti*, y permite realizar búsqueda de términos y palabras, analizar las inter-conexiones entre los mismos, así como el contexto en que son mencionadas en los archivos.

Estructura general:

- Buscador:

El buscador parte de un listado de palabras validadas previamente a partir de su inclusión en los diccionarios pre-definidos y permite evaluar tanto la frecuencia de aparición de un término como la co-ocurrencia entre diferentes palabras en las unidades de agregación. Así también plantea la posibilidad de analizar su contexto de mención en el texto bruto.

- Explorador:

El explorador permite realizar un análisis partiendo de lo general a lo particular ya que es posible seleccionar un sub-conjunto de documentos y explorar las temáticas que incluye a partir de las frecuencias de términos, nubes de palabras, co-ocurrencia, redes de palabras y asociaciones. También es posible dirigir el análisis seleccionando los diccionarios de interés.

- Analizador:

El analizador plantea un análisis centrado en el contexto en que se menciona una determinada palabra o conjunto de palabras previamente validadas por los diccionarios, a partir de la frecuencia, redes de términos y asociación entre las palabras que forman parte de dicho contexto.

Paquetes utilizados:

- *shiny* , *shinythemes*, *shinyWidgets*, *shinycssloaders* , *wordclouds2* , *DT* para diseñar la estructura de la aplicación, formato, visualizaciones.
- *dbplyr* para hacer conexión con la base en Postgres y optimizar las búsquedas SQL.
- *quanteda* (*quanteda.textmodels*, *quanteda.textplots*, *quanteda.textstats*) para visualización y cálculos de co-ocurrencias y distancias entre términos.

- Otros: *dplyr* , *ggplot2*, *seededlda*

Elina Gómez
UMAD (FCS) - Udelar
elina.gomez@cienciassociales.edu.uy

Análisis de la Red de Investigadores del IESTA

Anónimo

Abstract El presente trabajo constituye un primer avance en el trabajo elaborado para obtener una primera visualización del cuerpo de investigadores del Instituto de Estadística (IESTA) de la Facultad de Ciencias Económicas y Administración (FCEA) de la Universidad de la República mediante el Análisis de Redes Sociales (ARS). El objetivo de este trabajo es generar un diagnóstico parcial del modo de trabajo de los investigadores del IESTA, generando un insumo muy importante para el actual Plan Estratégico del IESTA. Se buscarán construir distintas redes de acuerdo a diferentes características de los investigadores en cuanto a la visibilidad de sus diferentes perfiles académicos (Cvuy, ORCID, ResearchGate y Publons) y parte de su producción bibliográfica bajos los formatos de artículos, libros, documentos de trabajo, proceedings, exposición en jornadas académicas de FCEA, exposición en jornadas académicas externas a FCEA, y tutorías de trabajos finales de carrera o tesis de posgrado. Finalmente, se deja a disposición de los lectores el código y los datos utilizados para llevar a cabo el análisis en un repositorio de dominio público con el fin de que el mismo sea reproducible (ver <https://gitlab.com/iestafcea.udelar/red-de-investigacion-del-iesta>).

Palabras clave: Análisis de Redes - Detección de comunidades - Perfiles de investigación

Referencias

- Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2015. "Systemic Risk and Stability in Financial Networks." *American Economic Review* 105 (2): 564–608.
- Álvarez-Vaz, Ramón, and Silvia Altmark. 2019. "ESTUDIO Del Gasto En Turistas de Cruceros En Uruguay Para La Temporada 2010-2011 Mediante El anÁLISIS de Redes." *Cuadernos Del CIBAGE* 1 (21): 27–64.
- Barrat, Alain, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2004. "The Architecture of Complex Weighted Networks." *Proceedings of the National Academy of Sciences* 101 (11). National Acad Sciences: 3747–52.
- Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10). IOP Publishing: P10008.
- Cárdenas, Julián. 2016. "El análisis de Redes: Qué Es, or'genes, Crecimiento Y Futuro." *Pensando Psicolog'a* 12 (19): 5–10.
- Clauset, Aaron, Mark EJ Newman, and Christopher Moore. 2004. "Finding Community Structure in Very Large Networks." *Physical Review E* 70 (6). APS: 066111.
- Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal Complex Systems*: 1695. <https://igraph.org>.
- Elsner, James B, and Thomas H Jagger. 2013. *Hurricane Climatology: A Modern Statistical Guide Using R*. Oxford University Press.
- Erdős, Paul, and Alfréd Rényi. 1959. "On Random Graphs Publ." *Math. Debrecen* 6: 290–97.
- Jackson, Matthew O. 2010. *Social and Economic Networks*. Princeton university press.
- Krackhardt, David, and Jeffrey R Hanson. 1993. "Informal Networks." *Harvard Business Review* 71 (4): 104–11.
- Newman, M. 2010. *Networks: An Introduction*. Oxford University Press.
- Pons, Pascal, and Matthieu Latapy. 2005. "Computing Communities in Large Networks Using Random Walks." In *International Symposium on Computer and Information Sciences*, 284–93. Springer.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Team, R Core, and others. 2013. "R: A Language and Environment for Statistical Computing." Vienna, Austria.
- Tindall, David B, and Barry Wellman. 2001. "Canada as Social Structure: Social Network Analysis and Canadian Sociology." *Canadian Journal of Sociology/Cahiers Canadiens de Sociologie*. JSTOR, 265–308.

Tumminello, Michele, Fabrizio Lillo, and Rosario N Mantegna. 2010. "Correlation, Hierarchies, and Networks in Financial Markets." *Journal of Economic Behavior & Organization* 75 (1). Elsevier: 40–58.

La diversidad de las agendas locales frente al Covid-19

Enrique García-Tejeda

Doctor en Políticas Públicas

Las políticas económicas frente a la pandemia

En el inicio de la pandemia de Covid-19, los [pronósticos internacionales](#) auguraban la contracción de la Economía Mundial. Los gobiernos nacionales y locales implementaron herramientas regularmente utilizadas frente a las crisis económicas anteriores:

- Política fiscal
- Política monetaria
- Política cambiaria
- Política financiera

En México, los 32 gobiernos locales que conforman la federación utilizaron políticas económicas orientadas a sobrellevar el confinamiento y las medidas de distanciamiento social. La condonación del pago de impuestos, los préstamos a negocios y las transferencias directas fueron algunas de las herramientas más utilizadas para tratar de minimizar el impacto de la crisis económica.

La diversidad de las políticas locales

En la literatura de políticas públicas, Boydston et al. (2014, 182) proponen un índice para estimar la diversidad de las agendas públicas. El índice de Shannon propuesto (*Shannon's H*) está dado por el negativo de la sumatoria de la i -esima(x) proporción(p) multiplicado por su logaritmo natural (\log):

Índice de Entropía de Shannon

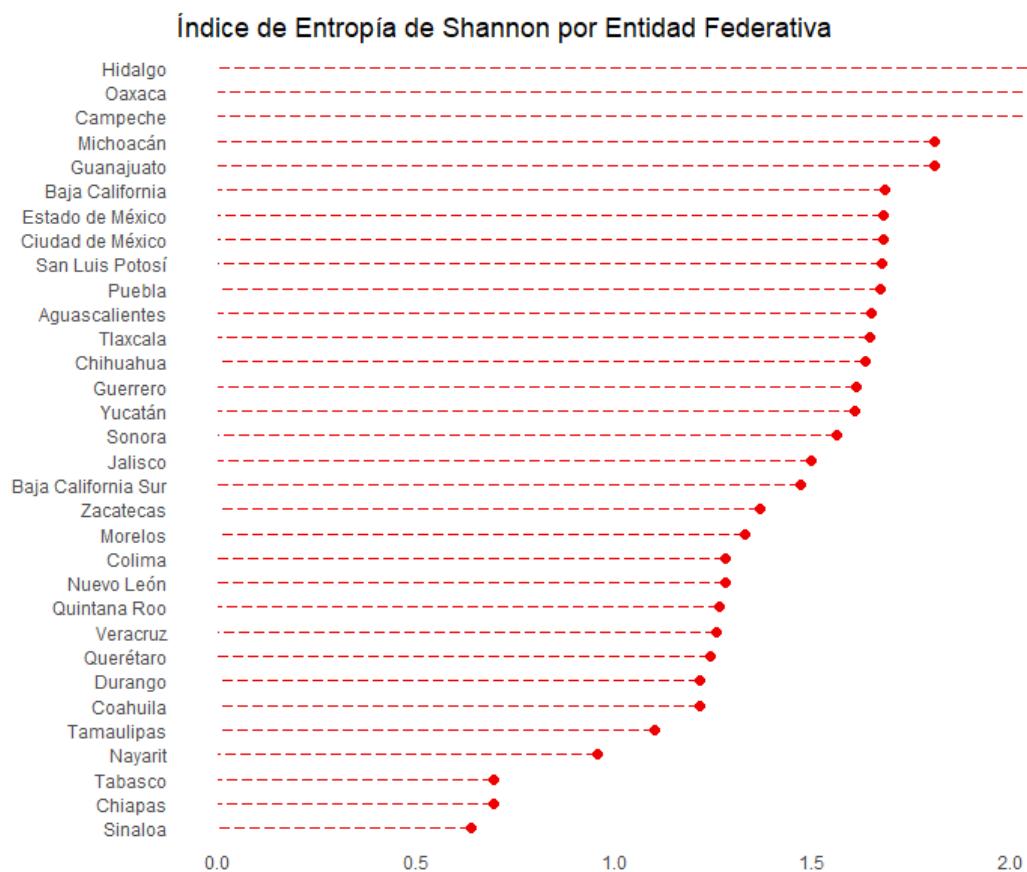
$$Shannon = - \sum_{i=1}^n ((p(x_i) * \log(p(x_i)))$$

En el caso de las políticas económicas de los gobiernos locales, la proporción es el número de medidas en uno de los n rubros de clasificación. Esta clasificación proviene de las categorías teóricas y prácticas en las cuales pueda categorizarse una política económica. Por ejemplo, los préstamos a las pequeñas empresas para el pago de renta de locales y para cubrir los salarios de los empleados pueden cuantificarse como 2 políticas económicas de la categoría de créditos empresariales frente al Covid-19. De esta manera, el número de intervenciones y las categorías de cualquier gobierno local o nacional pueden utilizarse como data para el cálculo del índice de Shannon.

La implementación de Shannon's H en R

En esta investigación se utilizó la base de datos del Laboratorio Nacional de Políticas Públicas [LNPP](#) sobre las políticas de los 32 gobiernos locales en México para estimar su diversidad con

R. La implementación del índice no utilizó librerías adicionales de R-base y se basó en 10 categorías de análisis para estimar la diversidad de las agendas locales en México. En el gráfico siguiente se presentan los resultados del índice de Shannon's que clasifica las agendas más diversas a menos diversas en forma descendente de los gobiernos locales en México.



El índice de Shannon's H para LatinR 2021

La participación consiste en presentar el desarrollo de una función de Shannon's H generalizada para cualquier conjunto de datos que posea una estructura de *data frame* y que contenga la información sobre agendas de gobiernos, partidos, organizaciones de la sociedad civil en categorías de análisis de interés ($n > 0$). El cálculo del estadístico se realiza sobre la columna que contenga la información de la frecuencia absoluta de las políticas analizadas. Adicionalmente es posible implementar una prueba no paramétrica para probar la [diferencia significativa](#) de las agendas locales, como complemento de la función diseñada.

El trabajo utiliza como ejemplo la base de datos de los gobiernos locales en México (marzo-mayo 2020), pero puede usarse cualquier data generado por sujetos de interés para el análisis de diversidad de Shannon. Puede parecer difícil de replicar, sin embargo, el código de la función permite estimar el índice con un mínimo de información en la base de datos (organizaciones, categorías y frecuencias). Los datos utilizados de México para mostrar la función son de libre acceso, por lo que pueden

utilizarse y replicarse fácilmente. Un problema a considerar es que no todos los países latinoamericanos dan seguimiento a las políticas y agendas públicas de sus gobiernos, por lo que puede ser difícil obtener datos para otro país.

Respecto a la interpretación del estadístico de Shannon en el ejemplo en concreto utilizado, la mayor diversidad de una agenda aumenta la posibilidad de que la intervención gubernamental cubra los aspectos más importantes de las problemáticas públicas que ha provocado la pandemia y reactive su economía.

Referencias bibliográficas

Las referencias se encuentran dentro de los hipervínculos del texto.

Elaboración de redes analíticas interactivas con el paquete netCoin de R

Anónimo

Abstract En esta presentación se expone un paquete para producir visualizaciones dinámicas e interactivas mediante el análisis de estadístico de datos. Se cubrirán los fundamentos del análisis de coincidencias y el uso de netCoin, un paquete para crear gráficos de análisis de redes. Este tipo de gráficos se ha empleado no sólo para resolver problemas topográficos y representar estructuras de red, como las de las redes sociales, sino también para mostrar la correlación entre variables según modelos casuales. El análisis de trayectorias y los modelos de ecuaciones estructurales son, en efecto, bien conocidos por los científicos sociales, pero ambos estaban restringidos a las variables cuantitativas en sus primeras etapas. En esta presentación, propondremos una nueva forma de mostrar las redes sociales y las conexiones entre variables cualitativas de forma similar al análisis de correspondencias, pero utilizando otro conjunto de técnicas multivariantes, como la regresión lineal y logística, mezcladas con el análisis de redes. Por ejemplo, uno de los usos específicos de esta técnica de análisis consiste en la caracterización de diferentes perfiles de respuesta por diversas variables sociodemográficas. El ARC (análisis reticular de coincidencias) explora los vínculos o co-ocurrencias de personas, características o eventos en determinadas circunstancias. El paquete R netCoin implementa este análisis de coincidencias y genera atractivas visualizaciones interactivas de los datos. Este paquete permite analizar las relaciones en los datos de las encuestas, las conexiones a través de los datos de las redes sociales o cualquier otro tipo de datos de redes dibujando gráficos interactivos. Incluso, netCoin también puede utilizarse para visualizar modelos estadísticos como las regresiones lineales o los modelos de ecuaciones estructurales. En el breve tiempo disponible se incluirán ejemplos con datos de redes sociales y de encuestas.

Palabras clave: análisis reticular de coincidencias - paquete netCoin - análisis de redes - gráficos dinámicos e interactivos

Introducción

A la presente propuesta se la denominará Análisis Reticular de Coincidencias (ARC), puesto que su principal objetivo es descubrir una serie de fenómenos, opiniones o características que en un determinado campo suelen aparecer conjuntamente. Muchos estadísticos persiguen la ilusión de dar con las causas de los fenómenos a partir de la información. Sin embargo, hay que ser cautos pues, salvo que se aplique con rigor el método experimental, las herramientas estadísticas son muy limitadas en el estudio de causas y efectos. Por ello, se propone una serie de análisis que no tienen como meta el descubrimiento de las "verdaderas" causas de los fenómenos en estudio, sino sus pautas de concurrencia con el fin de proporcionar al investigador posibles sugerencias de cómo está estructurada la realidad. El primer uso que se le puede dar a este análisis proviene de la dificultad de trabajar en cuestionarios con preguntas multirespuesta. Un siguiente empleo del análisis reticular de coincidencias es el análisis de contenido. También este análisis ha sido empleado para analizar documentos fotográficos y para ver la evolución de los retuits de un determinado tema en la red de Twitter. El análisis reticular de coincidencias tiene como finalidad descubrir las pautas de concurrencia de una serie de sucesos en un conjunto de escenarios. Su objetivo es descubrir cómo se distribuyen conjuntamente una serie de características dispuestas en distintas unidades en las que pueden o no estar presentes. Pueden distinguirse diversos grados de coincidencias: la nula, la simple, la probable, la condicional, la estadísticamente condicional, la subtotal y la total. Además, para estimar qué grado de coincidencias presentan dos sucesos puede emplearse un amplio rango de estadísticos: desde la mera frecuencia, hasta medidas de distancias más complejas, como la de Haberman o la de Rusell. Ahora bien, no solo es importante representar las coincidencias numéricamente con estadísticos. Una buena representación gráfica puede ayudar sobremanera a entender mejor la distribución de las coincidencias de un conjunto múltiple de sucesos.

El paquete netCoin para representar redes de coincidencias

netCoin es un paquete escrito en R que permite al usuario generar matrices de coincidencias con sus correspondientes grafos, así como crear una página web interactiva. En esta página interactiva pueden cambiarse un gran conjunto de elementos de los grafos, así como generar tablas y gráficos descargables.

Entre los elementos modificables se citan los siguientes:

- a. La etiqueta, el tamaño, el color y la forma de los sucesos o nodos, en función de sus propiedades. De igual modo se pueden representar áreas de nodos con las mismas características (conglomerados) e incluso reemplazar las formas geométricas de los nodos con imágenes.

- b. La etiqueta, el grosor y el color de las aristas que representan las coincidencias entre los sucesos, en función de las propiedades de los vínculos (frecuencias, grado de coincidencia, significación, ...)
- c. Se puede hacer una selección de nodos manual o en función de sus atributos.
- d. Se permite realizar una selección de las aristas en función de sus propiedades.
- e. De modo similar, puede cambiarse la disposición de los nodos del grafo según los diversos algoritmos que calculan otros paquetes de redes existentes en R.

Además de la representación gráfica, se obtienen dos tipos de tablas de atributos: la de los nodos y la de enlaces y en cada una de estas tablas aparecen las correspondientes propiedades de unos y otros. El paquete supone tiene dos características que lo distinguen de otras representaciones de redes entre los paquetes de redes. La primera es la elaboración de redes interactivas que pueden ubicarse en un servidor web a fin de que estén accesibles a cualquier persona con solo conocer su URL y la segunda es que permite una representación dinámica de las redes y el usuario puede ver cómo evolucionan los nodos y los enlaces a lo largo del paso del tiempo. En la presentación se empezará con una demostración con un par de ejemplos de resultados obtenidos con el paquete y se terminará con varios ejemplos de código para representar redes sencillas a través de las funciones del paquete.

Retos y R en conteos rápidos electorales

Los conteos rápidos son métodos poderosos para monitorear elecciones, su importancia radica en que proveen de información oportuna a la sociedad y fomentan la transparencia en el proceso electoral. En México, el objetivo del conteo rápido es publicar predicciones del porcentaje de votos a favor de cada candidato la misma noche de la elección. Las estimaciones utilizan como insumo los conteos de votos finales de un conjunto de casillas seleccionadas de acuerdo a un diseño muestral probabilístico.

El conteo rápido conlleva varios retos, entre ellos sobresalen los siguientes:

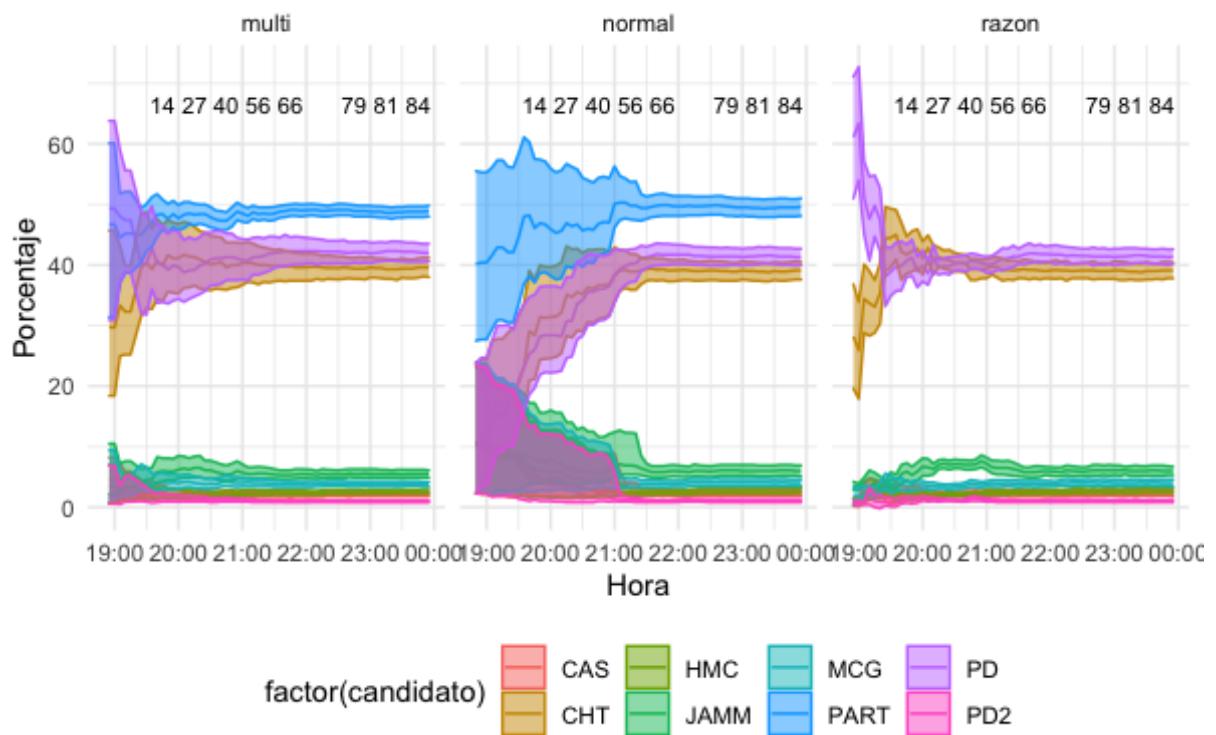
1. Por cuestiones de seguridad las estimaciones se hacen desde un cuarto aislado sin acceso a internet o teléfono. Esto limita las herramientas computacionales, ya que no se puede utilizar cómputo en la nube, y exige una preparación exhaustiva antes de la jornada electoral, pues no podremos descargar herramientas o buscar ayuda en la red.
2. A pesar de que se publica una única estimación al fin de la jornada electoral, el comité del conteo rápido acuerda actualizar las predicciones conforme se recibe nueva información, esto es cada 5 minutos, por lo que los métodos deben ser rápidos y robustos.
3. Las muestras diseñadas nunca se reciben en su totalidad, y muchas veces es necesario estimar con un alto porcentaje de casillas faltantes. Estas muestras tienen sesgos asociados al patrón de llegada de la información, sesgos que se deben considerar en el desarrollo de la metodología de estimación.

Los primeros dos retos hablan tanto de la elección de un modelo factible como de las herramientas computacionales óptimas para hacerlo. Platicaremos de por qué R es una herramienta ideal para afrontar estos desafíos, promoviendo transparencia y reproducibilidad. Hablaremos también de la importancia de tener un flujo de trabajo robusto, siguiendo buenas prácticas de desarrollo de software (control de versiones, revisión de código, uso de contenedores, etc.), y el porqué en nuestro equipo decidimos desarrollar un paquete de R para organizar el código final.

El tercer reto se tradujo en cómo producir estimaciones confiables, con buenas propiedades estadísticas, utilizando muestras parciales y cómo desarrollar un marco para validar nuestra metodología en escenarios realistas. Hablaremos de los aspectos que consideramos en este problema, y de los aspectos más generales del proceso de construcción de modelos en aplicaciones. En esta etapa de desarrollo, R también fue fundamental, pues las paqueteterías estadísticas nos permitieron construir modelos en R y hacer interfaz a Stan, donde escribimos el modelo bayesiano con el que estimamos la noche de la elección.

Como conclusión compararemos métodos a través de ejercicios de simulación y aplicaremos nuestra propuesta a los datos de las elecciones de junio del 2021.

Votacion para gobernador de Michoacan



Impacto de gasto em educação no ensino fundamental municipal: uma aplicação shiny

Fernando Barbalho *

Tiago Pereira *

Lucas Leite *

Jordão Gonçalves *

* Secretaria do Tesouro Nacional

O painel indicadores IDEB é uma aplicação shiny que permite a execução de análises exploratórias a partir do cruzamento de dados de despesa de educação dos municípios brasileiros com o desempenho das escolas públicas municipais brasileiras e também com os microdados do Censo Escolar Brasileiro. O desempenho é medido através da nota média das escolas no exame nacional do ensino básico, conhecido como IDEB.

O conjunto de dados gerado pelos cruzamentos de bases foi submetido a algoritmo de classificação para identificação de elementos importantes na determinação de notas do IDEB. A figura 1 mostra a tela inicial do aplicativo.

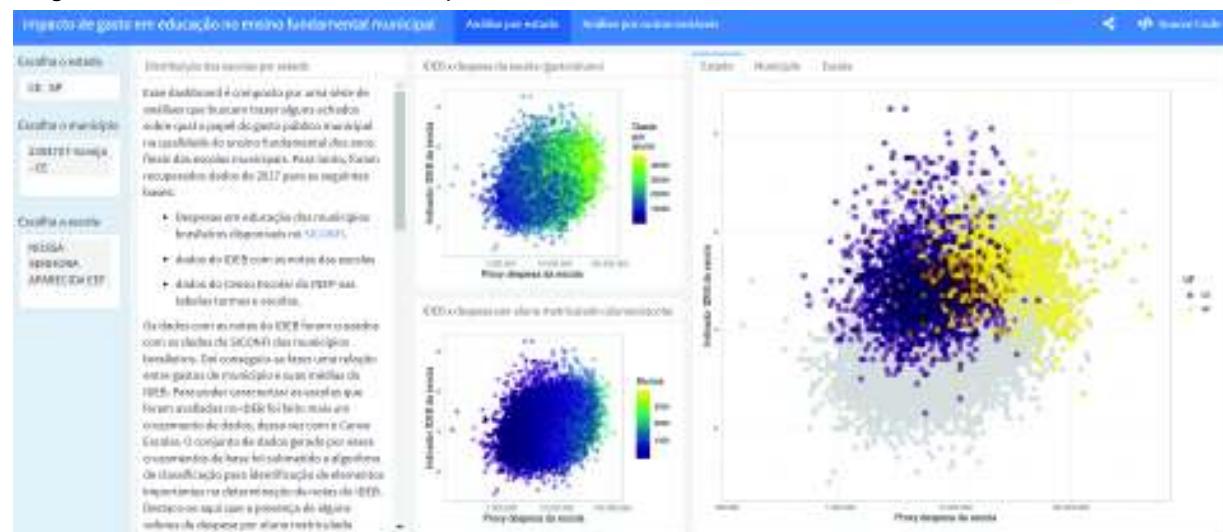


Figura 1: Tela inicial do aplicativo

Na última coluna da Figura 1 há três abas relacionadas a gráficos dinâmicos. A aba Estado começa mostrando a distribuição de notas e gastos por escola para Ceará e São Paulo. A condição mais importante para a determinação da média do IDEB das escola municipais foi o fato da escola estar localizada no estado do Ceará, enquanto que a segunda condição é a escola estar localizada no estado de São Paulo. Para o caso do Ceará percebe-se a predominância de escolas num quadrante em que estão as que desempenham melhor e gastam menos. Já São Paulo se destaca no quadrante de gastos mais altos e desempenho também mais alto. Como o gráfico é dinâmico, o usuário pode selecionar outros estados para fazer outras comparações usando o menu.

A segunda aba reflete o filtro sobre municípios e tem como valor default a cidade com a escola de maior pontuação. A terceira aba permite acompanhar a posição de escolas que o usuário selecionar. A escola selecionada por default é a de maior nota de IDEB.

Outras análise são feitas na segunda página do aplicativo, mostrada na Figura 2.



Figura 2: Segunda página do aplicativo

Além da variável Estado, o uso do algoritmo de classificação permitiu identificar outras variáveis importantes. Os quatro gráficos ao lado mostram por default essas variáveis.

A variável mais importante é a presença de quadra de esportes nas escolas. Observa-se que as quadras de esportes estão relacionadas principalmente às escolas com maior gasto e além disso, parece haver um relacionamento forte entre a presença de quadra de esportes e um maior indicador de IDEB para as séries finais do ensino fundamental municipal.

A segunda variável de destaque é a que trata da presença de internet. Percebe-se que esse recurso está presente em um número grande de escolas, com exceção das que têm baixa despesa. Nota-se ainda que as escolas que não têm internet estão bastante associadas a baixas notas do IDEB.

A terceira variável trata da existência de biblioteca. Nesse caso percebe-se que há uma predominância um pouco maior de escolas sem esse equipamento. As escolas que possuem biblioteca estão associados a uma maior despesa e também a médias mais elevadas de IDEB, considerando aqui que há um ponto de corte entre notas altas e notas baixas em torno da média 4.

A última variável relevante é a presença de água de rede pública. Percebe-se que a quantidade de escolas que é servida desse recurso é maior do que as que possuem quadra de esportes e biblioteca e menor do que as escolas com internet. Para essa variável, nota-se uma predominância nas escolas com maior despesa e o efeito parece ser favorável a maior média de IDEB.

As quatro variáveis citadas nos parágrafos anteriores podem ser substituídas por outras que compõem a tabela de Escolas do Censo Escolar. As outras variáveis não são tão importantes para análise do gasto das escolas nos equipamentos escolares e também para a determinação

de nota do IDEB, porém são úteis para ter uma ideia da universalização e/ou ausência de elementos que ajudam a caracterizar as escolas brasileiras.

O produto está disponível no seguinte link:

https://fabdev.shinyapps.io/dashboard_IDEB

Análisis del tipo de consumo cultural según nivel socioeconómico de los beneficiarios del programa Pase Cultural

*Jonathan A. Modernel, Magdalena Cornejo

Introducción

El acceso y la participación en la vida cultural están consagrados en la Declaración Universal de los Derechos Humanos de 1948. Aún con definiciones dinámicas acerca de lo que se entiende como cultura e incluso participación cultural, los Estados han tomado un rol central en la promoción y sostenimiento de la vida cultural así como también de la inclusión de distintos grupos en actividades culturales. En este contexto, el Ministerio de Cultura de la Ciudad Autónoma de Buenos Aires implementó en el año 2018 una política de fomento al consumo cultural entre adolescentes de 16 a 19 años de escuelas secundarias públicas. El Programa Pase Cultural consiste en una transferencia monetaria que puede utilizarse exclusivamente para la compra de bienes y servicios culturales, en una cartera de comercios y espacios habilitados. Si bien existen otras experiencias similares en el mundo, como los programas “18app” en Italia y “Pass Culture” en Francia, no se conocen publicaciones sobre los alcances y usos de estos programas por parte de los jóvenes.

El principal objetivo de este trabajo es realizar un análisis empírico de la dinámica temporal en el consumo de los beneficiarios del Programa Pase Cultural con especial énfasis en la determinación que el nivel socioeconómico pueda tener sobre los consumos culturales. Este análisis será abordado desde dos perspectivas: un análisis descriptivo de la población de beneficiarios segmentando por nivel socioeconómico y otros factores sociodemográficos; y, un análisis multivariado, en el cual se focalizará en los tipos de consumos realizados en función del rubro (en particular en cines y librerías) así como el nivel de uso del programa a lo largo del tiempo, nuevamente evaluando la significancia estadística que el nivel socioeconómico pueda tener en la determinación del consumo cultural. Conocer la composición de la población de beneficiarios así como también sus principales comportamientos permitirá dotar de información a los responsables del Programa Pase Cultural para la toma de decisiones futuras.

El análisis de datos se realizó utilizando el lenguaje R, siendo la primera experiencia personal y un punto de inicio en la formación para la evaluación de políticas públicas a futuro. La comunidad de R así como la enorme cantidad de información y ayuda disponible para el uso del lenguaje, hicieron de esta una experiencia de gran aprendizaje.

Metodología

Los datos utilizados para la descripción del perfil de los beneficiarios corresponden al período de octubre 2018 a

septiembre 2019. Como fuentes de datos se encuentran: la base administrativa de beneficiarios, el registro de transacciones para cada beneficiario; la base de establecimientos educativos; y, la base de comercios adheridos.

La mayoría de las bases utilizadas fueron generadas a partir de carga manual de datos (inscripción online, inscripción presencial) con distintos criterios a lo largo del tiempo. Por esta razón fue necesario unificar distintas variables categóricas (por ejemplo, en tipo de inscripción; inicialmente figuraban los distintos programas o lugares de inscripción y se unificó en inscripción online, en escuelas o eventos). Para la limpieza y manipulación de las bases de datos fue fundamental el uso de las librerías dplyr y tidyr, así como también las librerías lubridate y stringr para manipular las variables temporales y extraer información a partir de los campos de texto. Para armar el panel desbalanceado (cada fila correspondía a un consumo con los datos del beneficiario) también se utilizaron los paquetes de tidyverse. Para el análisis econométrico, se estimaron modelos lineales utilizando los paquetes AER, plm, stargazer y lmtest. Para la realización de gráficos se utilizó ggplot2 y viridis, para la elección de paletas de colores.

A partir de la consolidación de las bases mencionadas anteriormente, para la estimación de los distintos modelos se construyó un panel desbalanceado con la información de cada uno de los 7541 beneficiarios del programa para los que se conoce el nivel socioeconómico (NSE) y las transacciones que hicieron con fecha de consumo, importe y rubro y otras variables de control tales como género, edad, tipo de inscripción a lo largo de los 12 meses comprendidos entre octubre de 2018 y septiembre de 2019. De esta manera, los modelos estimados contemplan tanto la dimensión transversal (intentando explotar la heterogeneidad entre los beneficiarios) como temporal (intentando incorporar la dinámica temporal en el consumo). El panel es desbalanceado ya que no todos los beneficiarios tienen la misma cantidad de tiempo en el programa ni todos registraron consumos a lo largo de los meses analizados.

Análisis

La población de beneficiarios en el período comprendido entre octubre de 2018 y septiembre de 2019 en la base final es de 7.541 estudiantes. La distribución de éstos de acuerdo al NSE se muestra en la figura 1, donde puede observarse la cantidad de beneficiarios según el decil asignado a cada uno. El decil 1 de NSE se corresponde con el nivel socioeconómico más bajo; el decil 10 se corresponde con el nivel socioeconómico más alto. Se indica también si los beneficiarios contaban ya con su tarjeta Pase Cultural o no, debido a que una parte de los mismos aún no la había retirado por la sede comunal seleccionada o no se le había hecho entrega de la misma en la escuela.

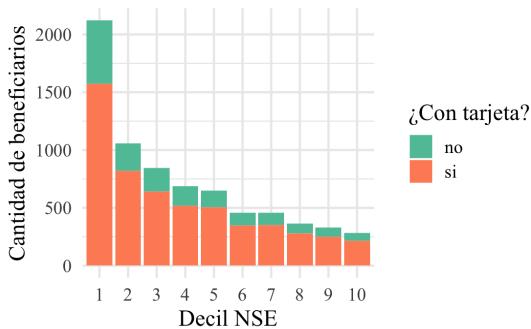


Figura 1. Distribución de beneficiarios según decil NSE según tengan o no la tarjeta.

Al analizar en la figura 2 la proporción de transacciones de acuerdo al rubro del consumo por decil de NSE, puede observarse que en todos los deciles el cine es el consumo mayoritario seguido por librerías. Además, en los deciles más bajos de NSE esta proporción es mayor. Al contrario, los deciles más altos muestran un mayor consumo de librerías en comparación con los de deciles más bajos.

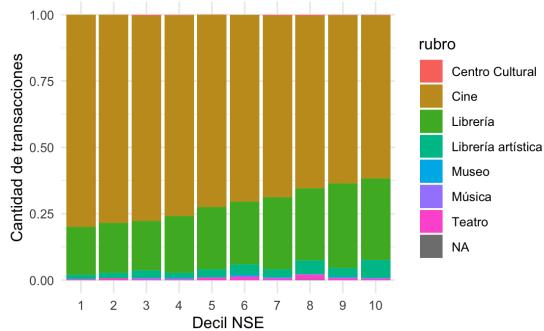


Figura 2. Proporción de transacciones según rubro de consumo para cada decil de NSE.

Si se analiza el consumo por el promedio del importe gastado, como se ve en la figura 3, no se encuentran diferencias significativas en el promedio del importe total entre los deciles de NSE. Sin embargo, si se analiza el importe promedio gastado según el rubro, en este caso en cines y librerías, se observa nuevamente que los deciles más bajos presentan un importe promedio gastado menor para librerías y considerablemente mayor en cines. Este importe promedio se iguala para los deciles más altos de NSE.

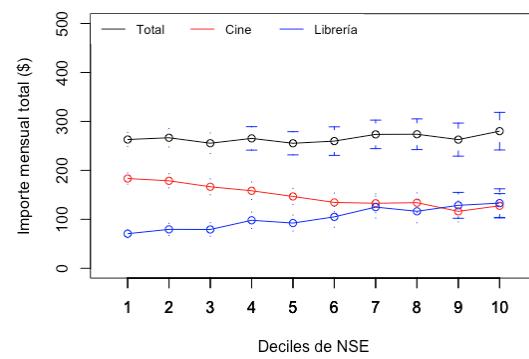


Figura 3. Importe mensual gastado en total, cine y librerías, de acuerdo al decil de NSE.

Conclusiones

El nivel socioeconómico permitiría explicar con distintos niveles de significancia el nivel de consumo de los beneficiarios. Si bien no se observaron diferencias significativas en el nivel de consumo total entre los beneficiarios activos del programa, sí las hay al considerar a la población completa de beneficiarios (aquellos que hicieron al menos una transacción y los que no hicieron ninguna). Esto muestra que los beneficiarios de mayor nivel socioeconómico tienen en promedio un mayor consumo en relación a los beneficiarios de menor nivel socioeconómico. Estas diferencias también se observan, y en algunos casos se acrecentan, cuando se compara según el tipo de consumo (cine o librería).

Con visiones divergentes sobre el rol de las políticas culturales, las políticas que facilitan y mejoran el acceso a la cultura de grupos que ven restringidas sus actividades culturales por recursos económicos son una tendencia en distintos países del mundo, como en Italia, Francia y Uruguay. Aún así, no se encuentran estudios sistemáticos y rigurosos sobre el impacto de estas políticas en los niveles de acceso y frecuencia, y menos aún sobre otras externalidades positivas de la participación cultural de los ciudadanos.

Desarrollo de una herramienta predictiva de calidad de agua para la gestión: modelos de Machine Learning + Shiny

El deterioro de la calidad del agua en Uruguay constituye un problema a nivel nacional. La Laguna del Sauce (Maldonado, Uruguay), segunda fuente de agua potable del país, no escapa a esta problemática. En este sentido, los gestores requieren de herramientas que les permitan anticiparse a condiciones no deseadas de calidad de agua (e.g. floraciones de cianobacterias, presencia de especies potencialmente tóxicas y de toxinas, la presencia de mal olor y sabor) para optimizar el proceso de potabilización y garantizar un adecuado abastecimiento a la población. La utilización de herramientas predictivas utilizando modelos de Machine Learning (ML) para la predicción de calidad del agua en el ámbito de la gestión, es hoy día muy incipiente a nivel nacional, e inclusive, a nivel internacional. La incorporación de estas herramientas acopladas a una interface interactiva, en el proceso de toma de decisión, presenta gran potencial dada su fácil utilización y costos prácticamente nulos. Es claro que la incorporación de estas herramientas en este ámbito, dependerá del buen desempeño de los modelos utilizados y de la disponibilidad de recursos humanos para su ejecución y mantenimiento.

En este contexto, el presente trabajo muestra los resultados de la generación de una aplicación en Shiny recientemente desarrollada para la predicción de atributos de calidad del agua en Laguna del Sauce usando modelos de ML. Dicha aplicación se encuentra instalada en el servidor de OSE y disponible para los gestores de la planta de OSE-UGD ubicada en Laguna del Sauce.

Dentro de los modelos de ML, se consideraron a los Random Forests (RF) dado que fueron en todos los casos los modelos que presentaron el mejor desempeño entre una amplia variedad de modelos de ML. Las variables de respuesta consideradas fueron: niveles de clorofila-a (proxy de biomasa del fitoplancton), niveles de biovolumen de cianobacterias, presencia de Grupos Funcionales Basados en Morfología (Kruk *et al.*, 2010) (en concreto se consideraron los grupos III y VII que son los que comprometen más severamente a la calidad del agua) y presencia/ausencia de la especie de cianobacteria potencialmente tóxica *Microcystis aeruginosa*. Todas las variables son categóricas (2 a 3 clases) y en muchos casos desbalanceadas. Se aplicó la técnica SMOTE para lidiar con el problema del desbalance de datos en la variable de respuesta. Las variables predictoras fueron variables de calidad de agua, meteorológicas e hidrológicas. Para entrenar los modelos, se contó con una base de datos a paso diario en el período 2002-2019 (17 años de datos). La evaluación de los modelos se hizo calculando el error de clasificación sobre una muestra test independiente, no utilizada para nada en el proceso de entrenamiento. Los modelos se evaluaron no solo mediante su error global, sino que también por su capacidad de predecir las diferentes clases por separado. Cabe aclarar que los modelos predicen las variables de respuesta mencionadas del día siguiente a partir de información de las variables predictoras del día de hoy. Las principales librerías utilizadas en este trabajo fueron: DMwR para lidiar con el desbalance de datos, randomForest para ajustar los modelos predictivos y ggplot2 para los gráficos.

Todos los modelos presentaron un desempeño global (porcentaje de casos correctamente clasificados) sobre la muestra de testeo independiente de entre 78 y 94%. El desempeño por clase también fue muy bueno, presentando en general valores

por encima del 80%, salvo algunas excepciones. En concreto, los niveles superiores de las variables (niveles que comprometen en mayor medida la calidad de agua) fueron muy bien clasificados.

Todo el desarrollo descrito, se incorporó en una aplicación Shiny donde los gestores pueden implementar los modelos en tiempo real obteniendo predicciones para el día siguiente. Además de las predicciones se brinda información gráfica para visualizar la evolución temporal de variables predictoras relevantes, así como la distribución y relación entre variables. También se encuentra accesible en la página de la aplicación una breve descripción de todo lo que la misma contiene, los resultados más relevantes de los modelos y los códigos para generarlos. Vale mencionar que un resultado que deriva del uso de la aplicación es la actualización permanente de una base de datos con todas las variables predictoras y de respuesta. Eso hace que en un futuro se puedan actualizar los modelos sin mayores esfuerzos además de tener una base actualizada disponible. Aún es muy reciente el uso de la herramienta en la planta de OSE-UGD, la evaluación de su incorporación en el proceso de toma de decisión, será un insumo muy importante para valorar la importancia de este tipo de desarrollos en el ámbito de la gestión.

La aplicación está disponible en la siguiente url:

https://matias-mw.shinyapps.io/Laguna_del_Sauce_2/S

Sesión **Aplicaciones** **en salud** **pública**

Uso se R en la validación de un modelo predictivo con aplicación a la enfermedad del asma

Alejandra Tapia Silva

Palabras clave: datos del asma; modelo logístico de efectos mixtos; desempeño predictivo; diagnósticos de influencia global y local; Método de Metropolis-Hastings; R

Resumen

En este trabajo se propone el uso de R para la validación de un modelo predictivo ampliamente usado en las áreas de las ciencias médicas. Esta validación se realiza a través del desarrollo de una metodología simultánea de la técnica de influencia local y global para el modelo logístico de efectos mixtos, con el objetivo de detectar observaciones influyentes que puedan afectar sus inferencias y desempeño predictivo. Para la implementación de la metodología se creó un código en R, utilizando funciones de paquetes tanto para el ajuste del modelo como para la obtención de las medidas del desempeño predictivo. Un estudio con datos reales de asma en niños y adolescentes, recolectados en un hospital público de São Paulo, Brasil, fue realizado para la aplicación. Los resultados muestran que esta metodología es útil para obtener un modelo predictivo preciso que proporcione evidencia científica cuando se toman decisiones médicas basadas en datos.

Introducción

El asma es una de las enfermedades crónicas más importante que afecta a millones de personas en todo el mundo. El asma se describe como una enfermedad heterogénea por la Iniciativa Global para el Asma (GINA: <https://ginasthma.org>) y se caracteriza como una inflamación crónica de las vías respiratorias. En las últimas décadas, la prevalencia del asma está aumentando en muchos países, especialmente entre los niños y adolescentes (GINA). Por lo tanto, estrategias basadas en la evidencia científica son cruciales para generar mejores medidas preventivas, así como un mayor acceso y adherencia a tratamientos que reduzcan la carga económica. La evidencia científica sobre el asma está fuertemente relacionada con el uso de modelos predictivos, que proporcionan información valiosa en las diferentes áreas de las ciencias médicas para la toma de decisiones basadas en datos. En esta dirección, uno de los modelos predictivos ampliamente utilizado para ajustar la presencia o ausencia de una enfermedad con datos agrupados corresponde al modelo logístico de efectos mixtos. Este modelo presenta desafíos estadísticos que tienen una fuerte implicación en los resultados y pueden comprometer las inferencias y predicciones y, consecuentemente, las conclusiones para la toma de decisiones basadas en datos. Así, una vez que el modelo se ha ajustado, es fundamental evaluar la calidad de éste, para comprobar la validez de su ajuste.

Metodología

En este trabajo se propone una metodología para la validación del ajuste de un modelo logístico con efectos mixtos, la cual consiste en el desarrollo simultáneo de las técnicas de influencia local y global que a menudo se utilizan por separado (ver Tapia et al. 2020 y referencias en el mismo), con el objetivo de detectar observaciones influyentes en las inferencias y desempeño predictivo. Además, de identificar observaciones que tengan un comportamiento demasiado diferente en relación a las demás; ver más en Tapia et al. (2020). Esta metodología es implementada con un código R, cuyo algoritmo es descrito a continuación.

- Paso 1: Ajustar el modelo logístico de efectos mixtos utilizando la función `glmer` del paquete `{lme4}`. Calcular las medidas del desempeño predictivo: sensibilidad (Sen), especificidad (Esp) y porcentaje de clasificación correcta (PCC), utilizando las funciones `sensitivity`, `specificity` y `pcc` del paquete `{PresenceAbsence}`.
- Paso 2: Fase I: Basados en la estimación de parámetros obtenida en Paso 1, implementar un código R para muestrear observaciones aleatorias desde la función de densidad condicional dada en ec. (6) de Tapia et al. (2020), a través del método de Metropolis-Hastings. Luego, basados en el método de Monte Carlo, con estas observaciones aproximar las esperanzas condicionales de las matrices dadas en ec. (7) y (8) de esta misma referencia.
- Paso 3: Fase II: Con las matrices obtenidas en Paso 2, implementar un código R para calcular la medida de influencia global, junto con el punto de corte asociado, y detectar grupos y observaciones influyentes. Realizar un análisis post-eliminación, es decir, eliminar estos grupos y observaciones, y recalcular las estimaciones de los parámetros y medidas de Sen, Esp y PCC.
- Paso 4: Fase III: Con las matrices obtenidas en Paso 2, implementar un código R para calcular la medida de influencia local, junto con el punto de corte asociado, y detectar observaciones influyentes. Realizar un análisis post-eliminación, es decir, eliminar estas observaciones, y recalcular las estimaciones de los parámetros y medidas de Sen, Esp y PCC. Tanto en el Paso 3 como 4 para calcular las medidas de influencia global y local se utilizaron algunas funciones para cálculo matricial del paquete `{matrixcalc}`.
- Paso 5: Fase IV: Basados en los resultados de las estimaciones de los parámetros y medidas de Sen, Esp y PCC de la Fase II y Fase III, determinar los grupos y/o observaciones para un nuevo análisis post-eliminación.
- Paso 6: Fase V: Basados en los resultados de la Fase IV realizar un nuevo y final análisis post-eliminación.

Aplicación

Un estudio con datos reales de asma en 362 niños y adolescentes, recolectados en un hospital público de São Paulo, Brasil, fue realizado para la aplicación; ver más información en Tapia et al. (2020). El objetivo de este estudio consistió en construir un modelo predictivo preciso para la probabilidad de que un paciente presente obstrucción fija de la vía aérea dado el grupo de gravedad del asma en el que fue clasificado y covariables de interés. Específicamente, la variable respuesta Y_{ij} corresponde al registro de que si los pacientes tenían ($Y_{ij} = 1$) o no ($Y_{ij} = 0$) una obstrucción fija de la vía aérea (FAO). Además, estos pacientes se clasificaron en cuatro grupos según su Gravedad del asma (Grupo 1 - Asma intermitente, Grupo 2 - Asma persistente, Grupo 3 - Asma persistente moderada, Grupo 4 - Asma grave persistente), incorporando esta variable como intercepto aleatorio, u_i . De esta forma, se asume un modelo logístico de efectos mixtos dado por: $Y_{ij}|u_i \sim Bernoulli(\pi_{ij})$, con $u_i \sim N(0, \sigma^2)$ y

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \beta_1 Trat_{ij} + \beta_2 Eosi_{ij} + \beta_3 Alergia_{ij} + u_i, \quad j = 1, \dots, n_i, i = 1, \dots, 4, \quad (1)$$

donde $Trat$ corresponde a la duración del tratamiento en años, $Eosi$ a la presencia o ausencia de eosinofilia en los análisis de sangre y $Alergia$ a la suma de todos los niveles de todos los factores que producen alergia siguiendo la prueba radioalergosorbente (RAST). El Cuadro 1 muestra las estimaciones, errores estándar (ES), valores-p, cambios relativos (CR) y medidas de Sen, Esp y PCC para el modelo inicial (1) ajustado. Este modelo es denominado inicial debido a que es ajustado con todas las observaciones del conjunto de datos.

Para validar el ajuste de este modelo, se aplica la metodología propuesta, donde los resultados obtenidos en la Fase V son mostrados en Modelo Fase V del Cuadro 1. Se puede observar que al eliminar los casos del Grupo 4 para $y_{ij} = 0$ y Grupo 3 para $y_{ij} = 1$, las covariables $Eosi$ y $Alergia$ dejan de ser significativas al 5% y 10%, respectivamente, y las medidas de Sen, Esp y PCC aumentan sustancialmente. Sin embargo, si se considera el Modelo Fase V reducido, es decir, sin esas covariables, la estimación de la varianza asociada a la Gravedad del asma aumenta y las medidas Sen, Esp y PCC disminuyen. Por lo tanto, estas covariables deben permanecer en Modelo Fase V, siendo éste el modelo predictivo final. En conclusión, con el uso de R se implementa una metodología que nos permite obtener un modelo preciso con mayor capacidad predictiva y algunos cambios inferenciales, proporcionando evidencia científica mejorada para los datos de la enfermedad del asma en la toma de decisiones médicas. Cabe destacar que esta metodología permite identificar situaciones que no podrían detectarse si utilizáramos estas técnicas por separado. Además, de identificar observaciones que tengan un comportamiento demasiado diferente en relación a las demás con las que se pueda realizar un escrutinio adicional.

Cuadro 1: Resultados de las estimaciones, errores estándar (ES), valores-p, cambios relativos (CR) y medidas de Sen, Esp y PCC para los modelos ajustados.

Casos eliminados	Efecto	Estimación (ES)	valor-p	CR	Sen	Esp	PCC
Modelo inicial							
Ninguno	Intercepto	-4.0430 (0.7343)	<0.0001	-	0.6969	0.7598	0.7541
	Tratamiento	0.1871 (0.0579)	0.0012	-			
	Eosinofilia	0.1101 (0.0481)	0.0220	-			
	Alergia	-0.7006 (0.4026)	0.0817	-			
	Gravedad del asma	0.5417	-	-			
Modelo Fase V							
Group 4 - $y_{ij} = 0$	Intercept	-4.7920 (3.2765)	0.1435	18.5261	0.8750	0.8860	0.8851
Group 3 - $y_{ij} = 1$	Tratamiento	0.2987 (0.1181)	0.0114	59.6048			
	Eosinofilia	0.0916 (0.0863)	0.2883	16.7512			
	Alergia	-0.4026 (0.7687)	0.6004	42.5413			
	Gravedad del asma	5.5658	-	927.5726			
Modelo Fase V reducido							
Group 4 - $y_{ij} = 0$	Intercepto	-4.2963 (3.2379)	0.1845	6.2670	0.8333	0.8382	0.8378
Group 3 - $y_{ij} = 1$	Tratamiento	0.2869 (0.1159)	0.0133	53.3056			
	Gravedad del asma	5.7493	-	5121.0510			

Tapia, A., Giampaoli, V., Leiva, V., and Lio, Y. (2020). "Data-Influence Analytics in Predictive Models Applied to Asthma Disease." *Mathematics*, 8, 1587. <https://doi.org/10.3390/math8091587>.

GINA. The Global Strategy for Asthma Management and Prevention; GINA Report: Fontana, WI, USA, (2020). Available online: <https://ginasthma.org/ginareports> (accessed on 11 September 2020).

Alejandra Tapia Silva
Universidad Católica del Maule, Chile
alejandraandreatapiasilva@gmail.com

Ciencia de datos con R con impacto en salud pública. Una experiencia de uso de tidyverse para la detección de embarazos.

Abstract

La implementación de una historia clínica electrónica en el sistema de salud pública de la Ciudad de Buenos Aires derivó en una base de un gran volumen de datos. No obstante, al ser una base transaccional cuyo objetivo es mejorar la atención de pacientes, el uso secundario de análisis de datos presenta un desafío por el tipo de datos registrados (semi estructurados, incompletos, subjetivos). Al día de la fecha la historia clínica no cuenta con un módulo destinado a identificar y caracterizar embarazos. La detección de embarazos es de gran relevancia ya que facilita el acceso a derechos de la salud sexual y reproductiva, y permite destinar fondos para ello. Durante el 2021, integrantes del equipo de ciencia de datos de la Agencia Operativa de Gestión de Información y Estadísticas en Salud del Ministerio de Salud de la Ciudad Autónoma de Buenos Aires ejecutaron un proyecto de mejora del proceso existente de detección de embarazos y de consolidación de una base específica con actualización periódica. Este proyecto fue llevado a cabo por un equipo interdisciplinario y desarrollado en el lenguaje R utilizando principalmente la librería *tidyverse* para la exploración, manipulación y visualización de datos masivos. El primer paso del proceso resultó novedoso ya que consistió en utilizar una nueva fuente de información, el sistema de admisión, pases y egresos sobre internaciones hospitalarias. Así, se integró esta información sobre eventos obstétricos a los registros obtenidos a partir de la historia clínica mediante el uso de expresiones regulares de los campos de texto estructurado y texto libre. Los siguientes pasos consistieron en la extracción y el cálculo de variables, la clasificación de cada registro, la discriminación entre embarazos de una misma persona y la caracterización de cada uno.

El uso de *tidyverse* fue central para el éxito del proyecto, que se encuentra en una etapa de cierre y evaluación.

Keywords: Expresiones regulares, Salud Pública, Historia clínica electrónica, *A*, *tidyverse*, Software R.

Este trabajo repone una experiencia de uso de R¹ en producción con impacto en políticas públicas de acceso a la salud por parte de un equipo de ciencia de datos del Gobierno de la Ciudad de Buenos Aires (CABA).

A partir del año 2011 el CABA comenzó a implementar una historia clínica electrónica (Historia Integral de Salud, HIS) en los centros de salud del primer nivel de atención. La HIS consiste en una base transaccional de las atenciones que se dan de manera ambulatoria y cuyo principal objetivo es mejorar la atención de los pacientes de manera de constituir una verdadera red de atención de salud. A pesar de la ventaja de contar con un gran volumen de información para su análisis, el uso secundario de los datos presenta ciertas dificultades ya que el registro es incompleto, subjetivo, poco sistemático y semi estructurado².

Dentro de este gran volumen de información los embarazos tienen gran relevancia, ya que su detección temprana permite facilitar la accesibilidad a derechos de la salud sexual y reproductiva, como por ejemplo controles de embarazo oportunos. La detección de este tipo de eventos también reviste de importancia económica en tanto son objeto de políticas públicas de financiamiento.

Si bien está previsto su desarrollo a mediano plazo, al día de la fecha la HIS no cuenta con un módulo vinculado a los embarazos que permita facilitar la identificación y extracción masiva de información vinculada a estos eventos. Es por ello que durante el 2021, dentro de la Agencia Operativa de Gestión de Información y Estadísticas en Salud (AGES) del Ministerio de Salud de la Ciudad Autónoma de Buenos Aires se destinaron recursos humanos a la mejora

del proceso existente de detección de embarazos y a la consolidación de una base específica con actualización periódica.

Este proyecto fue llevado a cabo por un equipo de ciencia de datos interdisciplinario y desarrollado en el lenguaje R mediante RStudio Server³, que se volvió clave en el contexto de teletrabajo impuesto durante la pandemia.

Para este desarrollo se utilizaron múltiples paquetes R. Parte de la colección de paquetes tidyverse⁴ (*lubridate*, *dplyr*, *tidyr*, *stringr* y *ggplot2*) para el proceso de *data wrangling* y visualización de datos *properties*, *dbplyr* y *odbc* para gestionar las conexiones con bases de datos S L y *agiseR*, librería desarrollada internamente por el equipo de Ciencia de datos de ES, para extraer información del *Data Warehouse* de la gerencia.

La decisión de dar inicio al proyecto respondió a diversos factores técnicos y sanitarios.

Respecto de los primeros, se destacan la búsqueda de reproducibilidad y de optimización de recursos dado que esta temática es transversal a varios proyectos, la reciente disponibilidad de una nueva fuente de información (admisión, pases y egresos, ADT), y la necesidad de adecuación del circuito de procesamiento de datos debido a la implementación de una nueva versión de la HS (tabla 1).

En cuanto a los factores sanitarios, se destaca la relevancia de estrategias de acompañamiento y búsqueda activa debido al impacto que tuvieron las medidas de aislamiento social en particular durante el comienzo del 2020 sobre el acceso de las personas a los centros de salud, acciones que dependen fuertemente de la identificación de los grupos de interés.

	HIS	ADT
Acto que registra	Atención ambulatoria longitudinal (consultas)	Internación: ingreso, pase y egreso
Ámbito	Atención primaria de la salud (APS) y consultorios externos y guardias de algunos hospitales.	Internación Hospitalaria
Tipo de información	Clínica y psicosocial	Administrativa
Tipo de datos recolectados	Semiestructurado (motivos de consulta) y campo de texto libre (evolución)	Estructurado (por ejemplo, tipo de evento y fecha)

tabla 1. Fuentes de información. Comparación entre las dos fuentes de información utilizadas para el proyecto: módulo de Historia Integral de Salud (HS) y módulo de admisión, pases y egresos (ADT).

El primer paso del proceso consistió en recopilar todos los registros por embarazo a partir de las dos fuentes de información (figura 1). Mediante el uso de expresiones regulares⁵ se detectaron consultas en la HS relativas a embarazos que incluyeran información sobre edad gestacional. A partir de este dato se calcularon la fecha de Última Menstruación (M) y la fecha Probable de Parto (PP) para todos los embarazos que se encontraran dentro del período de búsqueda. En el sistema ADT se registra la fecha de parto y la edad gestacional a ese momento: a partir de estos datos, calculamos la M.

Los siguientes pasos consistieron en clasificar cada registro (primera consulta o control) para poder discriminar cada embarazo (considerando la posibilidad de más de un embarazo por persona dentro del periodo de interés) y finalmente caracterizarlo. Las variables de interés refieren a la delimitación temporal del embarazo, a su finalización efectiva, la edad gestacional al primer registro y al momento de la finalización. Por último, conservamos los embarazos en curso durante el periodo de interés definido al inicio del proceso.

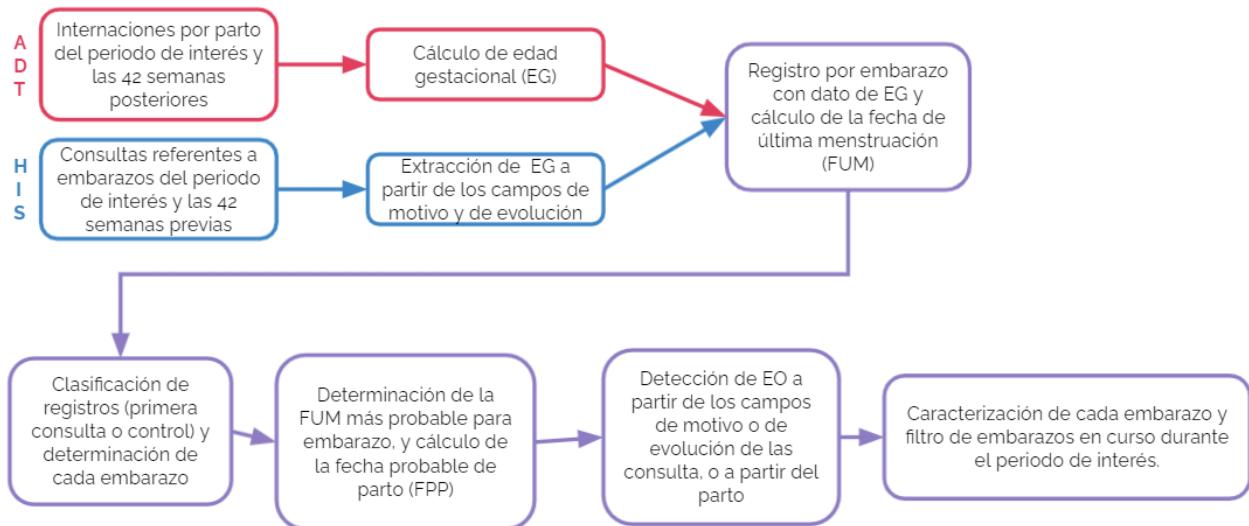


Figura 1. Esquema del proceso de detección de embarazos.

Como resultado se obtuvo un algoritmo que genera una tabla con los embarazos activos durante el período de interés y las variables comúnmente utilizadas en los pedidos de información relativos a la temática. Esto marca un avance importante dado que no es posible contar con esta información de manera directa a partir de los datos transaccionales.

Referencias

- (1) R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org>.
- (2) Otsis, L., Hartvigsen, J., Chen, Y., Meng, C. (2010). Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010, 1.
- (3) RStudio Team (2020). RStudio: Integrated development for R. RStudio, P. C., Boston, MA www.rstudio.com.
- (4) Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, (43), 1.
- (5) Shivade, C., Raghavan, P., Osler Lussier, E., Embi, P., Elhadad, N., Johnson, S., Lai, A. M. (2019). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2), 221-230.

¿Por qué los funcionarios públicos de salud deberían saber R para analizar sus datos?

Palabras clave: educación, salud pública, comunidad, ciencia de datos, colaboración

Autor: Paulo Villarroel Tapia (Enfermero y fundador OpenSalud LAB)

Resumen

Durante el último año hemos visto cómo el análisis de datos ha sido clave para combatir la pandemia COVID-19. Si bien la pandemia hizo más evidente la importancia del uso de datos, esta necesidad ha estado presente desde hace décadas. En particular, las instituciones de salud son entidades que generan una gran cantidad de datos a alta velocidad y de gran complejidad. Estos datos son claves para mejorar los procesos internos de las instituciones, entregar mejores tratamientos a los pacientes y tomar decisiones en salud pública basados en la evidencia. Pero para lograr aquello, es necesario contar con funcionarios con las competencias necesarias para explotar y extraer valor desde sus datos.

Por lo anterior, diseñamos un programa de formación (de acceso abierto, gratuito y colectivo) para el sector público de salud, buscando suplir la brecha que existe a la hora de pensar en qué problemas de diseño y gestión de políticas públicas son adecuados de resolver con datos.

Sobre el programa de formación

En OpenSalud LAB diseñamos un programa de formación que busca formar a distintos profesionales en variados aspectos de la ciencia de datos aplicado a salud. Entendemos que estas temáticas de análisis de datos tienen una tremenda aplicación y un gran potencial, pero siguen siendo poco utilizadas de forma regular dentro de las instituciones públicas en la toma de decisiones. Por otro lado, hemos visto que las personas que tienen conocimientos y competencias suficientes para desarrollar este tipo de proyectos son pocas o no están participando en las instituciones públicas. Por esos motivos, queremos acercar este conocimiento a los funcionarios públicos, sin barreras geográficas ni económicas.

El curso tiene un importante enfoque práctico y participativo, además, de ser realizado gracias a la colaboración de muchas personas y comunidades. En particular, contamos con la ayuda de la comunidad de R-Ladies Concepción, de docentes de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile y de otras personas que ayudaron como expositores.

El programa de formación fue realizado de forma online, con más de 100 inscritos en su inicio y con casi 6 meses de duración, mezclando sesiones en vivo, actividades asincrónicas, foros, documentación y de conversaciones offline. Las personas inscritas eran de diversas regiones de Chile, incluso de otros países latinoamericanos como Argentina, Colombia y Perú, en su mayoría del sector público de salud, pero también de instituciones privadas. Un dato interesante, es que casi la mitad de los inscritos eran mujeres y cerca del 15% tenía conocimientos previos en programación.

Programáticamente, el curso abordó 2 grandes áreas temáticas. El primero, llamado "Gestión y estrategia", dado que comprendemos que este tipo de temáticas están poco desarrolladas

dentro de las organizaciones y, por tanto, el grado de conocimientos específicos puede no ser muy elevado. En la segunda área, llamada "Análisis de datos y programación", se abordaron en detalle cómo diseñar un proyecto de datos y llevarlo a cabo. Se revisaron distintos aspectos metodológicos y técnicos para el desarrollo de este tipo de iniciativas, con un enfoque lo más práctico posible. Acá aplicamos el aprendizaje de R y sus distintas aplicaciones en salud.

Por su diseño, el curso es más parecido a una comunidad de práctica. Una comunidad de práctica se puede definir como un “grupo de personas que comparten una preocupación o una pasión por algo que hacen y aprenden a hacerlo mejor, interactuando con regularidad.” (Wenger, 2014).

El curso planteaba el desarrollo de un proyecto, por parte de los alumnos, que se iba trabajando a lo largo del tiempo. Entre ellos, se desarrolló un *dashboard* para visualizar las tablas quirúrgicas de un hospital público pediátrico, un gráfico interactivo para visualizar las iniciativas públicas y sus presupuestos de Chile durante de las últimas décadas. Además, a otras personas le ayudó a procesar la información de personas vacunadas por COVID-19 en un Servicio de Salud y desarrollar un script similar a un ETL en su departamento para consolidar información desde distintas fuentes.

El curso fue muy bien evaluado por los alumnos, con una alta valoración y nivel de recomendación.

Fue una experiencia grandiosa y de alto valor, logramos generar el interés de muchas personas sobre la importancia del análisis avanzado de datos y que el lenguaje R provee una gran herramienta para esos fines, constituyéndose en un aliado a la hora de resolver distintos desafíos en salud.

Actualmente se está desarrollando la segunda versión del curso, esta vez con varias novedades, más temáticas y tratando de incorporar las mejoras y aprendizajes de la versión anterior. Entre las más importantes están las de realizar más cantidad de talleres prácticos, aumentar las sesiones asincrónicas para estudio e incluir sesiones de tutorías 1-1 y grupales. Además, de generar un mayor énfasis en las primeras etapas de la Ciencia de Datos (importar, ordenar, transformar y visualizar) que notamos son más relevantes de desarrollar dado el perfil de los alumnos al cual está dirigido el curso y el nivel de madurez en temas de análisis de datos de las instituciones en las cuales habitualmente trabajan.

Referencias

1. OpenSalud LAB <https://www.opensaludlab.org>
2. Programa y clases del Bootcamp <http://datascience.opensaludlab.org/>
3. Listado de colaboradores https://github.com/opensaludlab/ciencia_datos#hugs-agradecimientos

Título: Aplicación del algoritmo de descomposición *stepwise replacement* para la estimación de la esperanza de vida libre de ansiedad/depresión en Argentina

Autor: Octavio Nicolas Bramajo (Centre d'Estudis Demogràfics/Universitat Autònoma de Barcelona, Bellaterra, España) obramajo@ced.uab.es

Resumen:

En ocasiones, dada la manera en que se construyen ciertos indicadores sociales, sanitarios y económicos, es necesario tener en cuenta que el valor de estos indicadores puede estar afectado por ciertos elementos que hacen a la composición de la población en cuestión. Un ejemplo de esto es el indicador conocido como esperanza de vida saludable/libre de enfermedades: se trata de un indicador de la morbilidad/salud de una población (que requiere tasas de prevalencia), pero que se deriva de la esperanza de vida (un indicador de mortalidad), para el cual son necesarias tasas de mortalidad). Por lo tanto, cambios en la mortalidad pueden tener cambios en la esperanza de vida saludable, pero sin afectar realmente la carga de la enfermedad en la población (Andreev et al., 2002). Es por esto que se realizan técnicas de descomposición demográfica, algunas de ellas incorporadas en el software libre R en el paquete *DemoDecomp* (Riffe, 2018)

Con el comando de descomposición *stepwise replacement* se realiza una descomposición de reemplazo de algoritmos, habitual cuando se trata de intervalos de tiempo discretos (el paquete también incluye la descomposición linear integral de Horiuchi, Wilmoth y Pletcher, habitual cuando se asume que los intervalos de tiempo entre los períodos son continuos en lugar de discretos). A modo de ejemplo, se han estimado las esperanzas de vida libre de ansiedad/depresión a los 20 años (EVLTAD20) para Argentina a partir de la Encuesta Nacional de Factores de Riesgo, calculadas con el método Sullivan (Jagger et al., 2006), y se han calculado las diferencias de EVLTAD20 para los períodos 2005-2009, 2009-2013 y 2013-2018, para cada componente (morbilidad y mortalidad) y por grupos quinquenales de edad. La función de Sullivan devuelve un output que es la EVLTAD20 (<https://github.com/onbramajo> presenta una forma completa de la descomposición, como ex.health).

Para utilizar la descomposición es necesario presentar los datos de prevalencia y de mortalidad en un mismo vector para cada momento, para luego separar cada componente en una matriz:

Paso 1: la descomposición

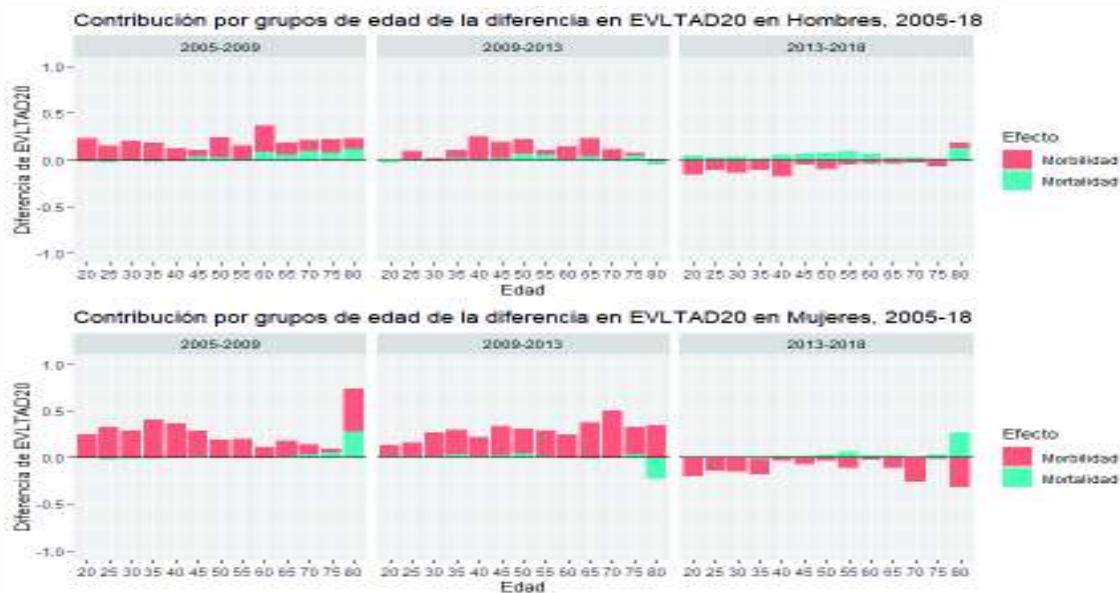
- Descomposición<- stepwise_replacement (func=Sullivan.fun,
pars1 =Vector1 c(Mortalidad2005,PrevalenciaDepresión2005),
pars2 =Vector2 c(Mortalidad2005,PrevalenciaDepresión2009)

Paso 2: separar los componentes en columnas

- Componentes
matrix(Descomposición,nrow=(length(Descomposición)/2),ncol=2,byrow=F)
colnames(Componentes)<- c("Mortalidad","Morbilidad")
• ### Paso 3: Hacer el data.frame para plotear
Componentesdf<- mutate(as.data.frame(Componentes),Edad=c(seq,20,80,5))
Componentesdf <- Componentesdf %>%
as.tibble() %>%
pivot_longer(c(1,2),
names_to = "Efecto",
values_to = "Contribución")

}

A partir de las EV20 y EVLTAD20 es posible obtener el tiempo vivido con ansiedad/depresión tanto en términos absolutos (expresados en años, como la diferencia entre EV20 y EVLTAD20) y graficarlos para ver la evolución en el tiempo del indicador: es decir, cuanto de la EVLTAD20 se debe a cambios en la mortalidad y cuanto en la morbilidad de las personas



Fuente: Elaboración propia en base a ENFRs, CELADE y DEIS

A través de la aplicación de un procedimiento de descomposición matemática disponible en el software libre R, se observó que durante el período 2005-2013 el aumento de la EVLTAD20 no sólo se debió a una mejora en la mortalidad, sino principalmente debido a una mejora general en la salud de la población en cuanto a la presencia de trastornos de ansiedad/depresión. Esto significa, que no sólo las personas vivieron más tiempo, sino también de manera más saludable durante ese tiempo, particularmente las mujeres, en donde el componente de morbilidad explicó casi la totalidad del cambio en la EVLTAD20. Sin embargo, durante 2013-2018 hubo un aumento generalizado en la prevalencia, acompañado de una disminución en la EVLTAD, que sería aún mayor si no fuera por un aumento en la LE general durante ese período.

Referencias bibliográficas:

- Andreev, E. M., Shkolnikov, V. M., & Begun, A. Z. (2002). Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies, healthy life expectancies, parity-progression ratios and total fertility rates. *Demographic Research*, 7, 499–521. <https://doi.org/10.4054/demres.2002.7.14>
- Jagger, C., Cox, B., & Le Roy, S. (2006). Health expectancy calculation by the Sullivan method. *EHEMU Technical Report*, June.
- Riffe, T. (2018). *DemoDecomp: Decompose Demographic Functions version 1.0.1 from CRAN*. <https://rdrr.io/cran/DemoDecomp/>

Sesión Enseñanza de R

Scaling feedback using learnr and gradethis in an introductory R course

Beatriz Milz , Fernando Correa

Palabras clave: Online courses - Formative assessment - Feedback

Several learning activities that were performed in person had to be changed into an online approach due to the COVID-19 pandemic. Several people that otherwise could not attend an in-person activity were able to join in the online activities. The increasing number of participants presents challenges, and the one we approach in this article is checking and giving feedback to homework assignments in R courses.

Checking and giving feedback to homework assignments usually are time-consuming activities, and as such, they might make it harder to give timely feedback. Mine Cetinkaya-Rundel gave a talk at the RStudio Global Conference 2021, in which she presented an approach for giving feedback at scale that is both meaningful and timely (Çetinkaya-Rundel 2021). We tried this approach in an introduction to programming with R class with 50 students, and here we present the main strengths and challenges we have faced so far.

In this course, we adopted interactive online homework assignments to provide formative assessment after each class. The process of reviewing and evaluating whether the students could answer the exercises correctly enabled instructors and students to qualify the understanding of contents discussed in class. Feedbacks and answers were steadily available while students answered to the assignments.

We structured our homework assignments using a blend of R packages to enable interactivity, steadily available feedbacks, and scalable review of several assignments. For the interactivity part, we used the package `{learnr}` (Schloerke et al. 2021), a tool that enables the creation of interactive tutorials by using R Markdown documents (Allaire et al. 2021) and Shiny Apps (Chang et al. 2021). The package `{gradethis}` (Aden-Buie et al. 2021) was designed to be used in `{learnr}` tutorials and was used to incorporate steadily available exercise feedback. The package `{learnrhash}` (Rundel 2020) was used to generate a compressed text-based representation of the answers (called *hash*), that students could copy and paste to submit their answers to the exercises. At the end of each homework, each student could send information such as their names, emails, and the hash code created by `{learnrhash}` (Rundel 2020) through Google Forms.

By using the package `{googlesheets4}` (Bryan 2021), we were able to import the answers sent by the students through Google Forms. We developed a reproducible report to present the results of homework assignments using R Markdown (Allaire et al. 2021) and the package `{pagedown}` (Xie et al. 2021). Two examples of the information presented in the report were the percentage of students that submitted their answers through Google Forms and the percentage of students that answered to each exercise correctly. This report was designed to be used only by the instructors of the course.

We asked the students to answer the homework assignments before the following class, so the instructors could review the report and identify concepts that should be revised in class. At the beginning of each class, we reviewed the concepts that we identified when using the reproducible report, and we also solved the exercises in order to explain and answer any questions left.

Distributing the `learnr` tutorials at scale can be a challenge, so we provided three options for the students:

1. An R package to store the `learnr` tutorials, available on GitHub (<https://github.com/>);
2. A project in RStudio Cloud (<https://rstudio.cloud/>), with the package installed, so students could make a copy of the project and use it;
3. A deployed version in `Shinyapps.io` (<https://www.shinyapps.io/>) using a Starter Plan (which costs \$9.00 per month + taxes).

Some of the strengths that we found while using this approach are:

1. We have data that can help identifying the misunderstanding gaps that should be clarified or filled through revision of contents.
2. Although we spend some time before the course adapting the exercises from .R files to structured exercises in a learnr interactive tutorial, during the course we did not spend time correcting and giving feedback on the homework for each student.
3. We chose to create hints in the exercises, but did not show the solutions within the learnr tutorial. The students reacted positively to the possibility to see the hints, which were not possible to create when we used the .R files for the homework.

We also faced some challenges that we are still working on improving:

1. Some of the students faced encoding errors when they used the tutorials directly in their RStudio sessions, when installing the package from GitHub.
2. Considering that it is an introductory course, some of the students faced difficulties installing the package from GitHub. We had to help each one of them to install it by using Zoom share screen feature, and talking through the installation problems.
3. A {learnr} interactive document is not the same environment in which usual R programming happens, such as an IDE like RStudio or VSCode, and some learners might have a hard time understanding how they are related. Learners who can correctly answer homework assignments might not realize how to apply their knowledge on longer codes, such as R scripts.
4. The packages {learnr} (Schloerke et al. 2021), {gradethis} (Aden-Buie et al. 2021) and {learnrhash} (Rundel 2020) were designed in English; therefore, all the messages, buttons, and other words are written in English. The maintainers of the package {learnr} (Schloerke et al. 2021) are making great efforts to support additional languages in the tutorials, and part of the interface already has support for it. Considering that this course is taught in Brazilian Portuguese, and the package {gradethis} (Aden-Buie et al. 2021) is still only available in English, we made a fork of the package and translated the internal messages into Brazilian Portuguese, called {gradethisBR}.

We are aware of the remaining work to do in order to improve the support of additional languages in the learnr tutorials that use this approach. However, this is important to facilitate the use of learnr by more non-English speakers in forthcoming courses and classes. We understand that {gradethisBR} is a temporary package and will be better if this feature is native in {gradethis}. Thus, we want to help improving {gradethis} to support additional languages as well, such as Brazilian Portuguese, Spanish, German and others.

References

- 10 Aden-Buie, Garrick, Daniel Chen, Garrett Grolemund, and Barret Schloerke. 2021. “Gradethis: Automated Feedback for Student Exercises in Learnr Tutorials.” <https://pkgs.rstudio.com/gradethis/>.
- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. “Rmarkdown: Dynamic Documents for r.” <https://pkgs.rstudio.com/rmarkdown/>.
- Bryan, Jennifer. 2021. “Googlesheets4: Access Google Sheets Using the Sheets API V4.” <https://github.com/tidyverse/googlesheets4>.
- Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. “Shiny: Web Application Framework for r.” <https://shiny.rstudio.com/>.
- Çetinkaya-Rundel, Mine. 2021. “Feedback at Scale.” Online. <https://www.rstudio.com/resources/rstudionglobal-2021/feedback-at-scale/>.

Rundel, Colin. 2020. “Learnrhash.” <https://github.com/rundel/learnrhash>.

Schloerke, Barret, JJ Allaire, Barbara Borges, and Garrick Aden-Buie. 2021. “Learnr: Interactive Tutorials for r.” <https://rstudio.github.io/learnr/>.

Xie, Yihui, Romain Lesur, Brent Thorne, and Xianying Tan. 2021. “Pagedown: Paginate the HTML Output of r Markdown with CSS for Print.” <https://github.com/rstudio/pagedown>.

Aplicaciones Shiny para modelos de crecimiento de ecología de poblaciones: una propuesta simple y no simplista para animar al uso de R en cursos introductorios
Lucía Rodríguez Planes

La ecología de poblaciones se apoya en modelos numéricos para organizar relaciones entre variables y parámetros, jerarquizar relaciones, hacer proyecciones y pronósticos, y poner a prueba hipótesis. Trabajar los conceptos tras estos modelos requiere de cierta fluidez en el pensamiento formal, la aplicación de conceptos matemáticos, el conocimiento de algún lenguaje de programación y el conocimiento ecológico conceptual en sí, que es el objetivo de enseñanza. Estudiantes de grado neófitos en herramientas numéricas y de programación pueden encontrar una carga cognitiva demasiado alta al ser introducidos al contenido curricular al mismo tiempo que a un lenguaje de programación. Por otro lado, existen programas en computadora que buscan facilitar estos contenidos al ocultar la parte formal y permitir la interacción solo con opciones y resultados. Estos programas, como el conocido “Populus”, no serían un paso intermedio en un proceso de comprensión ya que no permiten analizar las funciones detrás de los resultados. Ambas situaciones pueden fallar en motivar a estudiantes, por exceso de complejidad o por falta de ella.

La emergencia sanitaria mundial ocurrida durante 2020 forzó la virtualidad de los procesos de enseñanza-aprendizaje y extremó las dificultades al impedir la ocurrencia en el espacio de estudiantes y docentes, un elemento que nos era indispensable en las clases para asistir a quienes se iniciaban en el uso de herramientas de programación. De la necesidad de separar las dificultades del proceso de enseñanza-aprendizaje surgió la idea de diseñar una herramienta **simple** para la introducción a estos temas, y al mismo tiempo **no simplista**, y que invitara a estudiantes a explorar en profundidad el contenido desde la programación en una instancia posterior al manejo correcto del contenido ecológico conceptual.

Presento dos aplicaciones web Shiny desarrolladas para cursos que abordan la ecología de poblaciones y los modelos de crecimiento determinísticos de una población sin estructura de edades ni sexo, en versiones densoindependiente y densodependiente. A partir de la selección de los valores de constantes y variables en los modelos de crecimiento se construyen series de valores de abundancia en el tiempo que se visualizan en diferentes tipos de gráficos. Los modelos posibles son los abordados en los textos clásicos de ecología general, con alguna adición menos habitual entre los modelos densodependientes que se inspira en el programa de Ecología de Poblaciones, un curso avanzado de la carrera de Ciencias Biológicas de la Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Cada Shiny app consiste en un panel izquierdo para la selección de las opciones, y una sección amplia donde se genera el gráfico. Los primeros valores a elegir son la abundancia inicial y el tiempo total, y para el modelo densodependiente la capacidad de carga, comunes a todos los modelos. La botonera a continuación activa paneles con opciones para la tasa de crecimiento y parámetros adicionales de acuerdo con el modelo de crecimiento elegido: a tiempo discreto (con ecuaciones planteadas en diferencias de abundancia entre tiempos) o a tiempo continuo (basadas en modelos diferenciales y con resultados numéricos que surgen de la integración para diferenciales de tiempo de décima del paso de tiempo discreto). La última botonera permite modificar el tipo de gráfico que se desea ver, entre las opciones que resultan más importantes desde la comprensión de los modelos: gráfico de trayectoria de la

abundancia en el tiempo, gráfico del logaritmo natural de la abundancia en el tiempo, y el reclutamiento neto per cápita (continuo) o razón de abundancias de tiempos sucesivos (discreto) en función de la abundancia. Se pueden extraer conclusiones generales a partir de los valores elegidos y el comportamiento que despliegan, y también pueden elegirse valores publicados para alguna población real para proyectar su tendencia poblacional y discutir las implicancias de los diferentes modelos.

La Shiny app se basa en un solo archivo donde se distingue la interfaz del usuario y el servidor con comentarios dentro de un mismo código. Los paquetes necesarios y las funciones de crecimiento fueron definidas explícitamente en la sección inicial del código para que se carguen junto con la aplicación. A partir de la selección de los valores en las botoneras (interfaz de usuario) el servidor utiliza esos valores para alimentar las funciones de crecimiento elegidas. Los resultados los guarda en una *tibble* con la primera columna correspondiente a los valores de abundancia, la segunda al tiempo, y elabora las demás variables a utilizar usando ciclos *for* simples con reglas de subsetting de R *base*. La selección del tipo de gráfico y la información que toma de la *tibble* de resultados para cada eje (con solo ciertas combinaciones posibles) se basa en condiciones *if - else* de tres opciones. La decisión de priorizar código con estrategias en R *base* por sobre *tidyverse* se debió a que, a pesar de necesitar un código más largo, en este caso permite comprender casi sin información previa las instrucciones del código.

Las aplicaciones están disponibles en <https://www.shinyapps.io>: https://luciarp.shinyapps.io/pop_growth_indep y https://luciarp.shinyapps.io/pop_growth_dep. El código estuvo disponible en un repositorio con control de versiones https://github.com/luciairp/teaching_popecology del que estudiantes entusiastas y animados por la clase pudieran descargar y ejecutar desde su propia interfaz de RStudio, accediendo a cada una de las partes del código. Los estudiantes se mostraron interesados en el lenguaje que permitía generar la aplicación y motivados a ver el código en entorno R, lo que genera una expectativa que podría traducirse fácilmente en una introducción a un lenguaje de programación.

La primera versión de prueba se promocionó a la comunidad a través de la red social Twitter en agosto de 2020. Todos los comentarios recibidos señalaron dificultades y aspectos mejorables que fueron utilizados para generar la versión 1.0.0 actual. En esta versión ambas aplicaciones fueron utilizadas en la materia de grado “Ecología general” de la Universidad de Buenos Aires, “Ecología y Ambiente” de la Universidad Favaloro y “Ecología” de la carrera Ingeniería Ambiental de la Universidad Nacional de San Martín, en el segundo cuatrimestre 2020 y primer cuatrimestre 2021, en clases sincrónicas virtuales y que contaron con otras herramientas digitales integradas como la pizarra compartida de la suite Google: Jamboard. Los comentarios recibidos de su uso reciente se están incorporando a una nueva versión que será probada y lanzada hacia el final de 2021 y con versión trilingüe (español, portugués e inglés).

Karel la robot enseña R: un paquete para la enseñanza de programación

Marcos Prunello

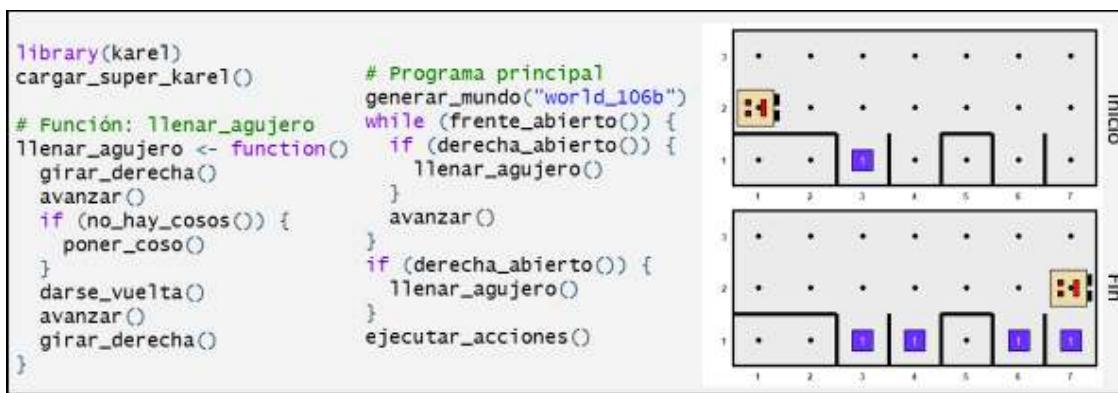
Palabras clave: educación - introducción a la programación - karel the robot - paquete bilingüe

En este trabajo se presenta `karel`, un nuevo paquete de R creado con el propósito de brindar un entorno que posibilite la enseñanza de principios de la programación en una forma dinámica e interactiva, para estudiantes sin experiencia previa que estén cursando el nivel secundario o los primeros años de su formación de grado. *Karel* es una robot que vive y camina por ciertos *mundos*, en los cuales puede realizar determinadas acciones si se lo pedimos, mientras aprendemos a programar en R.

La robot Karel acompaña en la enseñanza de conceptos fundamentales de la programación, por ejemplo, el de procesador (Karel), ambiente (su mundo), objetos (llamados *cosos*) y acciones (las actividades que puede realizar). Además, como se puede observar en las viñetas y en la guía de estudio que ejemplifica su utilización (disponible online en <https://mptru.github.io/introprog/>), junto con Karel es posible instruir sobre el concepto de la descomposición algorítmica: en numerosas oportunidades la robot debe cumplir objetivos cuya resolución requiere descomponer el problema en partes más pequeñas, para cada una de las cuales los estudiantes deben programar una función en R. También facilita la exemplificación del uso de estructuras de control de código secuenciales, condicionales (`if () {} else {}`) e iterativas (`for () {}, while () {}`).

El primer paso para programar con Karel es *generar un mundo* en el cual ella pueda andar, a través de la instrucción `generar_mundo()`, aclarando entre los paréntesis el nombre del mundo que queremos usar. El paquete trae incorporados unos cuantos pero los usuarios pueden optar por crear otros nuevos. Por ejemplo, con la instrucción `generar_mundo("mundo007")` se genera el mundo de la figura mostrada a continuación.

Todos los mundos de Karel son rectangulares, compuestos por calles que los recorren horizontalmente (filas) y avenidas verticales (columnas). Karel siempre se encuentra en la intersección entre una calle y una avenida (celda), mirando hacia una de las cuatro direcciones posibles: este, norte, oeste o sur. Los bordes negros representan paredes que Karel no puede atravesar, solo puede rodearlas. Además, en algunas celdas hay uno o varios *cosos*. Karel puede recorrer el mundo poniendo y juntando *cosos* por ahí; si los junta los guarda en su mochila y lleva un registro de cuántos tiene. Finalmente, Karel solo puede realizar estas actividades: `avanzar()`, `girar_izquierda()`, `juntar_coso()` y `poner_coso()`. Claro, agrupando ingeniosamente estas acciones básicas se pueden crear otras nuevas, por ejemplo, crear una función `girar_derecha()` que produzca tal efecto, haciendo girar a Karel tres veces a la izquierda. Por otro lado, Karel es capaz de evaluar ciertas características de su entorno a través de funciones que arrojan un valor lógico TRUE o FALSE, por ejemplo: `frente_abierto()`, `hay_cosos()` o `mira_al_sur()`, de manera que podemos condicionar las acciones que Karel realiza a la verificación de ciertos aspectos sobre su posición y su mundo. Una vez que se ejecuta el código con todas las acciones que Karel debe realizar, se debe correr la función `ejecutar_acciones()` y se puede ver el resultado en una animación creada con los paquetes `ggplot2` y `ganimate`. En el siguiente ejemplo se puede observar que, gracias a la creación de la función auxiliar `llenar_agujero()`, Karel puede reparar todos los baches que existen en la calle 1:



The screenshot shows the Karel the Robot software interface. On the left, there is a code editor with R-like pseudocode for a robot named 'karel'. The code includes functions like 'cargar_super_karel()', 'llenar_agujero()', and 'ejecutar_acciones()'. On the right, there are two world maps labeled 'Mundo' and 'Fín'. The 'Mundo' map shows a 7x7 grid with various obstacles (walls) and a starting point at (1,1). The 'Fín' map shows the same grid after the robot has completed its task.

```

library(karel)
cargar_super_karel()

# Función: llenar_agujero
llenar_agujero <- function()
  girar_derecha()
  avanzar()
  if (no_hay_cosos()) {
    poner_coso()
  }
  darse_vuelta()
  avanzar()
  girar_derecha()
}

# Programa principal
generar_mundo("world_106b")
while (frente_abierto()) {
  if (derecha_abierto()) {
    llenar_agujero()
  }
  avanzar()
  if (derecha_abierto()) {
    llenar_agujero()
  }
  ejecutar_acciones()
}

```

Con el objetivo de sortear algunas de las barreras idiomáticas que pueden hacer más desafiante poder dar los primeros pasos en el aprendizaje de programación, el paquete karel se diseñó de forma completamente bilingüe: todas las funciones tienen una versión en español y en inglés (por ejemplo, `girar_izquierda()` y `turn_left()`), así como todas las páginas del manual de ayuda y del sitio web están escritos en ambos idiomas.

Este paquete ha sido utilizado por primera vez en el año 2020 para la asignatura “Introducción a la Programación” de la Licenciatura en Estadística (Universidad Nacional de Rosario, Argentina). Se dicta en el primer cuatrimestre del primer año de la carrera, con una audiencia en la que predominan estudiantes recién graduados de su formación secundaria sin ningún tipo de experiencia en la programación. El uso de karel al inicio del cursado sirvió como herramienta para exemplificar y poner en práctica conceptos básicos de programación, ya que los estudiantes tuvieron la posibilidad de ver de manera interactiva el efecto del código escrito, al hacer que la robot, por ejemplo, repitiera ciertas estructuras o condicionara sus acciones a verificaciones lógicas. En el año 2021 se adelantó el empleo de karel al cursillo de nivelación para ingresantes, posibilitando la enseñanza de manera temprana y lúdica de nociones de control del flujo de código y de descomposición algorítmica.

La idea para la implementación de este paquete se basa en *Karel the Robot*, un lenguaje de programación creado con fines educativos por el Dr. R. E. Pattis de la Universidad de Stanford (California, EEUU), quien también escribió el libro *Karel the Robot: A Gentle Introduction to the Art of Programming*, en 1981. Su nombre es un homenaje a Karel Čapek, el escritor checo que inventó la palabra robot en su obra de ciencia ficción *R.U.R. (Rossum's Universal Robots)*. Su sintaxis se basaba en Pascal, pero a lo largo de las décadas esta estrategia de enseñanza fue implementada en distintos lenguajes como Java, C++, Ruby y Python. El paquete aquí presentado es la primera implementación de Karel para R. Algunos de los ejemplos incluidos en las viñetas son adaptaciones de aquellos publicados por Eric Roberts en su material *Karel the robot learns Java* (2005).

La página web del paquete karel es <https://mpru.github.io/karel> y se puede instalar desde CRAN mediante `install.packages("karel")`.

*Marcos Prunello
Escuela de Estadística
Facultad de Ciencias Económicas y Estadísticas
Universidad Nacional de Rosario
marcosprunello@gmail.com*

Un conjunto de paquetes para generar tutoriales interactivos para enseñar R

Yanina Bellini Saibene

Palabras clave: educación - enseñanza - learnr

Abstract

El paquete [learnr](#) nos permite desarrollar un nuevo nivel de programación literal al permitir generar lecciones interactivas utilizando como base documentos rmarkdown. Tanto sea como [material digital de un libro](#), como [ayuda de un paquete de R](#) o como material de ejercitación en cursos de grado y posgrado, los tutoriales interactivos se presentan como un recurso tecnológico interesante y poderoso para la enseñanza de R. Si además, se suman recursos pedagógicos para el diseño de lecciones y recursos abiertos para compartir y publicar las lecciones, el aporte al conjunto de recursos de la comunidad potencia la utilidad de estas herramientas. Este trabajo presenta un conjunto de paquetes de R para generar lecciones nuevas utilizando plantillas con aspectos pedagógicos, de diseño visual y en español y tutoriales listos sobre temas como visualización, iteración y trabajo con texto. Además se presenta un sitio web con recursos para aprender a generar estos materiales. Todos estos recursos son de acceso libre.

¿Por qué paquetes?

Empaquetar los tutoriales es una excelente alternativa para hacer llegar nuestras lecciones interactivas a un grupo grande de usuarios; una vez que se ha instalado el paquete, se puede ejecutar localmente. Esto evita los costos de publicar los tutoriales como aplicaciones Shiny y posibles problemas en la conexión a Internet. Como es un paquete, también puede contener datos propios para trabajar durante las lecciones.

El paquete [{learnres}](#)

Basado en learnr este paquete contiene dos plantillas para generar tutoriales: la misma plantilla que viene por defecto en learnr, pero traducida al español y una segunda plantilla con una sugerencia de estructura y elementos que contribuyen a que el material pueda ser compartido, reutilizado, citado y encontrado.

También tiene una hoja CSS con los estilos explicados para que se pueda editar y generar estilos propios y un JSON con todas las partes de la interfaz traducidas al español.

Paquetes de tutoriales

Por el momento se cuenta con tres paquetes:

- [**Tutorial interactivo sobre Minería de Texto**](#) que tiene como objetivo introducir los conceptos básicos de la minería de texto con un enfoque tidy basado en el libro Tidy Text de Silge y Robinson.
- [**Tutorial de introducción a la iteración con R:**](#) Este paquete contiene dos tutoriales para introducir el concepto de iteración en R. La iteración es la tarea de aplicar una función de forma iterativa a cada elemento de un vector. El primer tutorial explicará qué es un vector e introducirá dos formas de iterar en R: bucles 'for' y el paquete 'purrr'. El segundo tutorial introduce la familia de funciones map de purrr que hace que la iteración sea rápida y fácil. Aprenderás los secretos de `map()` y sus variantes.
- [**Tutorial Intro a ggplot2:**](#) incluye varios elementos pedagógicos como un mapa conceptual con los temas que cubre, dos personas tipo; elementos open access como la licencia de uso y luego una serie de explicaciones y ejercicios para hacer gráficos en R.

Todos estos tutoriales han sido utilizados en el dictado de clases de carreras de posgrado en dos universidades de Argentina. Las modalidades de uso de estos materiales fueron las siguientes:

- Durante la clase sincrónica por medio de trabajo en grupo los estudiantes realizan los tutoriales y luego se discute en el grupo completo, como les fue, qué conceptos fueron más complejos de aprender, sobre cuales hay dudas, etc.
- Como aula invertida, donde los alumnos realizan estos ejercicios por su cuenta y luego en clase discutimos los resultados, repasamos y despejamos dudas.
- Como material complementario o extra a la clase.

Las y los estudiantes destacan este recurso como uno de sus preferidos en la encuesta anónima de feedback (ticket de salida) posteriores a cada clase. El inconveniente más grande es la instalación de los paquetes cuando las dependencias son muchas. En el caso de Windows, si los usuario no cuentan con Rtools instalado, es necesario realizar estas tareas para que los paquetes funcionen correctamente,a compañando a las y los estudiantes en estas tareas.

En <https://learning-learnr.netlify.app/> se encuentra el acceso a todos estos recursos, con su código fuente, blog posts explicando cómo generar y utilizar cada herramienta, videos de cursos y charlas sobre el tema y proyectos donde se utilizan tutoriales interactivos.

Yanina Bellini Saibene
INTA-UNAB-MetaDocencia
yabellini@gmail.com

Sesión Comunidad de R

Autores: Claudia Alejandra Huaylla, Paola Corrales, Andrea Gómez Vargas, Joselyn Chávez, Denisse Fierro Arcos, Virginia García Alonso.

Título: Conociendo el camino para aprender a usar R en Latinoamérica: desafíos para promover la inclusión y diversidad

Palabras claves: Aprendizaje, Comunidades, Desigualdades, Relevamiento.

Resumen:

El entorno de R es una poderosa herramienta utilizada mundialmente para diversos fines. A pesar de la gran variedad de usos que se le da a esta herramienta, todas las personas que utilizan R tienen algo en común y es que en algún momento tuvieron que aprender dicho lenguaje. ¿Cuáles fueron los motivos por los cuales lo hicieron? ¿Dónde fue que aprendieron a usar R? ¿Qué desafíos enfrentan aquellas personas que utilizan R?

Estas son algunas de las preguntas incluidas en la primera encuesta Latinoamericana sobre el uso de R realizada en 2020. Creamos dicha encuesta para evaluar qué desafíos enfrentan los usuarios de R en una región donde el idioma, la infraestructura y las diferencias en acceso a recursos, información y capacitación pueden representar una barrera para el aprendizaje y uso de R. Solo conociendo cuáles son esas barreras y dificultades que enfrentaron (o no) al aprender y utilizar R es que se podrían implementar prácticas que mejoren la inclusión de les usuaries.

La encuesta fue compartida dentro de la comunidad de R desde Tijuana hasta Ushuaia y más de 900 personas nacidas y/o residentes de Latinoamérica contestaron las 31 preguntas planteadas. Aproximadamente el 19% de las personas que manifestaron pertenecer a una comunidad pertenecen a LatinR. Dentro de esta comunidad las redes más usadas son twitter y slack. Estos hallazgos pueden ser una herramienta útil para hacer difusión sobre algún evento y promover la inclusión en Latinoamérica.

Menos de un 25% de las personas encuestadas indicaron que les resultó difícil aprender R. La mayoría indicó que aprendieron a usar R, no por ser un requisito, sino por interés propio y de manera autodidacta utilizando material digital y/o cursos gratuitos online. Más aún, R fue el primer lenguaje de programación de casi la mitad de las personas encuestadas a pesar de que el 70% indicó conocer al menos otro idioma de programación.

Otro de los resultados más llamativos fue que el idioma inglés no fue identificado como una barrera para aprender R y resolver errores. De hecho, los artículos en inglés, la comunidad de Stackoverflow en dicho idioma y las documentaciones de los paquetes (que suelen estar en inglés) fueron identificados como los lugares más visitados para resolver problemas o aprender a usar un paquete en concordancia con lo mencionado anteriormente. Sin embargo, las personas encuestadas indicaron tener un alto grado de inglés y de educación en su mayoría, por lo que aún no podemos diferenciar si existe algún tipo de causalidad o es simplemente un reflejo de la situación de las personas que decidieron contestar la encuesta.

La baja representatividad de algunos grupos minoritarios históricamente excluidos (personas con algún tipo de discapacidad física o cognitiva, población LGBTIQ+, población

afrodescendiente y de descendencia indígena) también nos ofreció un primer acercamiento para dar cuenta de las desigualdades estructurales de la región.

A pesar de que aún falta información para responder algunas de las dudas iniciales, algo que sí queda claro de las respuestas es que las comunidades juegan un rol clave en el aprendizaje ya que la mitad de las personas indicaron que formar parte de alguna comunidad les ayudó a resolver problemas. A su vez, la mitad de los encuestados indicó que comparten su código, práctica que ayuda a mejorar la reproducibilidad y, por ende, la inclusión. De las personas que no lo comparten, la mitad indicó que es debido a que no saben cómo hacerlo, señalando una de las principales problemáticas que deberían ser abordadas para mejorar el aprendizaje y uso de R en Latinoamérica.

En esta charla presentaremos el análisis y los resultados de las preguntas antes mencionadas entre otras, e invitaremos a que más personas se sumen a esta iniciativa. La encuesta es abierta y reutilizable dado que buscamos inspirar a otras regiones desatendidas de todo el mundo a identificar sus fortalezas y desafíos y de dicha manera ayudar a aumentar la inclusión y la diversidad de la comunidad internacional de R.

Título: Juntas podemos más, corta historia de cómo la pandemia nos incentivó a colaborar

Autoras:

Denisse Fierro Arcos^{1,*}, Danisse María Carrascal Polo², Linda Jazmín Cabrera Orellana³, Mary Jane Rivero Morales²

¹R-Ladies Galápagos, Santa Cruz, Galápagos, Ecuador

²R-Ladies Barranquilla, Barranquilla, Atlántico, Colombia

³R-Ladies Guayaquil, Guayaquil, Guayas, Ecuador

* Autora de correspondencia: galapagos@rladies.org

Palabras/frases clave: Comunidad virtual, Aprendizaje virtual, Redes virtuales, Grupos minoritarios.

Resumen:

La pandemia del COVID-19 ha tenido un efecto negativo en la salud y economía de comunidades a nivel mundial. Las economías de América Latina y el Caribe se vieron especialmente afectadas, con la reducción más grande en el producto interno bruto a nivel mundial registrada en esta región en el 2020 (OCDE, 2020). Sin embargo, la pandemia de afectó de manera homogénea a la población, sino que mujeres y personas de pertenecientes a minorías de género (LGBTI), grupos que ya vulnerables debido a inequidades existentes previas a la pandemia (por ej., discriminación y marginalización que resulta en una mayor proporción de personas en estos grupos viviendo en la pobreza), se vieron desproporcionadamente afectadas y en múltiples frentes (Morgan et al, 2021; OCDE, 2020; UNCTAD, 2021).

Algunos de los impactos reportados entre estos grupos incluyen un incremento del 25% en la incidencia de violencia de género a nivel mundial desde el inicio de la pandemia (ONU Mujeres, 2020), y un incremento en el tiempo dedicado a trabajo no remunerado aumentó (por ej., cuidado de familiares, educación de hijos) y recayó principalmente en mujeres (OCDE, 2020; ONU Mujeres, 2020). Mientras que reportes sugieren que personas LGBTI tienen un riesgo más alto de mortalidad por COVID-19 debido al acceso desigual a salud y mayor probabilidad de tener condiciones de riesgo como asma, hipertensión y obesidad, así como resultados más deficientes en salud mental que personas de géneros binarios (Kuehn, 2021; Morgan et al., 2021). Todo esto resulta en que estos grupos tengan menor seguridad económica y peores resultados de salud (Kuehn, 2021; Madgavkar et al., 2020).

Estas estadísticas nos motivaron a crear oportunidades para que principalmente mujeres y personas pertenecientes a minorías de género puedan adquirir nuevas habilidades en programación de manera gratuita. Pero, preparar estas sesiones puede representar una cantidad importante de tiempo, lo cual limita la capacidad de grupos con pocos miembros organizadores de ofrecer eventos de manera regular. Inspiradas por la idea de sororidad en la que se basa R-Ladies, cuatro grupos (Barranquilla, Galápagos, Guayaquil y Milagro) en dos países diferentes (Colombia y Ecuador) decidimos trabajar juntas para crear un Club de Lectura basado en el libro *R para Ciencia de Datos*, el cual fue traducido al español por la comunidad latinoamericana de usuarios de R. Estas sesiones regulares atraen a un promedio de 20 participantes por sesión.

En esta charla compartiremos las lecciones que aprendimos durante los últimos meses, así como los errores cometidos, los obstáculos y soluciones que hemos encontrado durante la organización de estos eventos. Esperamos que esta información sirva de guía para grupos con poca experiencia organizando eventos en línea.

Referencias

- KUEHN, B.M. 2021. Sexual minorities gave greater COVID-19 risk factors [Online]. JAMA. Available: <https://jamanetwork.com/journals/jama/fullarticle/2777731> [Accessed October 20 2021].
- MADGAVKAR, A., WHITE, O., KRISHNAN, M., MAHAJAN, D. & AZCUE, X. 2020. COVID-19 and gender equality: Countering the regressive effects [Online]. McKinsey Global Institute. Available: <https://www.mckinsey.com/featured-insights/future-of-work/covid-19-and-gender-equality-countering-the-regressive-effects> [Accessed].
- MORGAN, R., BAKER, P., GRIFFITH, D.M., KLEIN, S.L., LOGIE, C.H., MWIINE, A.A., SCHEIM, A.I., SHAPIRO, J.R., SMITH, J., WENHAM, C., WHITE, A. 2021. Beyond a zero-sum game: How does the impact of COVID-19 vary by gender. *Frontiers in Sociology* 6, 126. doi: 10.3389/fsoc.2021.650729
- OCDE. 2020. COVID-19 en América Latina y el Caribe: Consecuencias socioeconómicas y prioridades de política [Online]. Organización de Cooperación para el Desarrollo Económico (OCDE). Available: https://read.oecd-ilibrary.org/view/?ref=134_134494-n1k7ww92ro&title=COVID-19-en-América-Latina-y-el-Caribe-Consecuencias-socioeconómicas-y-prioridades-de-política&ga=2.115982662.1985122934.1627028266-1676381724.1627028266 [Accessed July 23 2021].
- ONU MUJERES. 2020. Los efectos del COVID-19 sobre las mujeres y las niñas [Online]. ONU Mujeres. Available: <https://interactive.unwomen.org/multimedia/explainer/covid19/es/index.html> [Accessed July 23 2021].
- UNCTAD. 2021. COVID-19 threatens four ‘lost decades’ for gender equality [Online]. United Nations Conference on Trade and Development. Available: <https://unctad.org/news/covid-19-threatens-four-lost-decades-gender-equality> [Accessed October 17 2021].

Sesión Investigación y comunicación de resultados

Tablas reproducibles, presentables y con formato numérico local con {gtsummary}

Eva Retamal Riquelme.

Unidad de Neurología Adulto, Hospital Clínico La Florida. Santiago, Chile.

Abstract

El uso de tablas para resumir información es muy útil, especialmente para analizar valores uno a uno, y al comparar varias categorías. Con el paso del tiempo, frecuentemente es necesario generar nuevamente estas tablas. Hacer esto manualmente es un proceso tedioso y propenso a errores.

Por otra parte, la forma de escribir los números es distinta en diferentes idiomas. Si el software empleado utiliza un formato numérico diferente al idioma en que será presentado, es necesario cambiar este formato una vez concluido el análisis.

El paquete {gtsummary} permite elaborar tablas reproducibles que resumen conjuntos de datos. Estas tablas son altamente personalizables, permitiendo elegir el formato numérico del resultado. Esto disminuye la posibilidad de errores, y facilita la reproducibilidad de este tipo de análisis.

Usando el conjunto de datos *pinguinos*, del paquete {dados}, se ejemplifica cómo realizar tablas personalizadas, reproducibles, que pueden ser presentadas directamente e incorporan el formato numérico español.

Palabras clave: tablas - gtsummary - reproducibilidad - español

Demostración

Se muestran tablas realizadas con el paquete `gtsummary` (Sjoberg et al. 2021) a partir del conjunto de datos *pinguinos* (Quiroga et al. 2021), traducción al español del conjunto de datos *penguins* del paquete *palmerpenguins* (Horst, Hill, and Gorman 2020). Contiene datos de pingüinos de 3 islas del archipiélago de Palmer, en la Antártica. Para cada observación incluye: la especie del pingüino, la isla, sexo, año e información sobre su tamaño (*largo de aleta*, *masa corporal*, *largo y alto de pico*).

El análisis inicialmente será únicamente considerando la isla **Dream**. La función `tbl_summary` permite crear una tabla con medidas de resumen. Por defecto, aparece mediana y rango intercuartil para variables numéricas, y recuento (n) y porcentaje para variables categóricas (figura 1).

Figura 1

Characteristic	N = 124 ¹
largo_pico_mm	44.7 (39.2, 49.8)
alto_pico_mm	18.40 (17.50, 19.00)
largo_aleta_mm	193 (188, 198)
masa_corporal_g	3,688 (3,400, 3,956)
sexo	
hembra	61 (50%)
macho	62 (50%)
Unknown	1

¹ Median (IQR); n (%)

Personalizando las tablas

Para excluir variables que no sean de interés, basta con filtrar y seleccionar los datos de interés antes de ejecutar `tbl_summary`. El argumento `by` permite **separar** los datos en **categorías**.

Para agregar test de hipótesis se emplea la función `add_p` sobre resultado de `tbl_summary`. Por defecto, para variables numéricas entrega el valor p de test de hipótesis no paramétricos (figura 2).

Figura 2

Characteristic	hembra, N = 61 ¹	macho, N = 62 ¹	p-value ²
especie			>0.9
Adelia	27 (44%)	28 (45%)	
Barbijo	34 (56%)	34 (55%)	
Papúa	0 (0%)	0 (0%)	
largo_pico_mm	42.5 (37.0, 46.4)	49.1 (40.6, 51.2)	<0.001
alto_pico_mm	17.80 (17.00, 18.20)	19.00 (18.50, 19.70)	<0.001
largo_aleta_mm	190 (187, 195)	196 (191, 201)	<0.001
masa_corporal_g	3,450 (3,300, 3,650)	3,950 (3,756, 4,250)	<0.001

¹ n (%); Median (IQR)

² Fisher's exact test; Wilcoxon rank sum test

Formato numérico español

La función `theme_gtsummary_language` permite traducir el **formato numérico a español**, y permite elegir los separadores para miles y decimales.

Es posible modificar el texto que aparece en la tabla. En este caso se cambió la etiqueta de `especie`, aquellas relacionadas con el tamaño del pingüino, y `Characteristic` por `Variable`.

Se cambiaron las **medidas de resumen y test de hipótesis** para variables continuas a promedio, desviación estándar y prueba T de Student, respectivamente, y se agregó un **encabezado** para agrupar *macho* y *hembra* (figura 3).

Reproducibilidad

Si fuera necesario **replicar** esta tabla, incluyendo los datos de todas las islas, la tabla resultante sería la Tabla 4. En ella se mantienen las medidas resumen elegidas en el paso anterior.

Como el tamaño muestral es diferente, cambian los test estadísticos que aparecen por defecto. Para este caso, la prueba de hipótesis de la variable categórica `especie` cambia de prueba exacta de Fisher a prueba de Chi cuadrado.

Otras consideraciones

La **nota al pie** de la tabla que incluye por defecto las medidas resumen de estadística descriptiva y los test de hipótesis se puede eliminar o sobrescribir manualmente según sea necesario con la función `modify_footnote`.

En figura 5 se observa la parte inferior de una tabla similar a la figura 4, con una modificación de la nota al pie.

Referencias

- Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. “Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data.” Manual. <https://allisonhorst.github.io/palmerpenguins/>.
- Quiroga, Riva, Edgar Ruiz, Mauricio Vargas, and Mauro Lepore. 2021. “Datos: Traduce Al Español Varios Conjuntos de Datos de Práctica.” Manual. <https://github.com/cienciadedatos/datos>.
- Sjoberg, Daniel D., Michael Curry, Margie Hannum, Karissa Whiting, and Emily C. Zabor. 2021. “Gtsummary: Presentation-Ready Data Summary and Analytic Result Tables.” Manual. <https://CRAN.R-project.org/package=gtsummary>.

Figura 3

Variable	Sexo		
	hembra, N = 61 ¹	macho, N = 62 ¹	valor p ²
Especie			>0,9
Adelia	27 (44%)	28 (45%)	
Barbijo	34 (56%)	34 (55%)	
Papúa	0 (0%)	0 (0%)	
Largo del pico (mm)	42,3 (5,5)	46,1 (5,8)	<0,001
Alto del pico (mm)	18 (0,8)	19 (0,9)	<0,001
Largo de la aleta (mm)	190 (6)	196 (7)	<0,001
Masa corporal (g)	3.446 (270)	3.987 (350)	<0,001

¹n (%); Media (DE)
²test exacto de Fisher; t de Student

Figura 4

Variable	Sexo		
	hembra, N = 165 ¹	macho, N = 168 ¹	valor p ²
Especie			>0,9
Adelia	73 (44%)	73 (43%)	
Barbijo	34 (21%)	34 (20%)	
Papúa	58 (35%)	61 (36%)	
Isla			>0,9
Biscoe	80 (48%)	83 (49%)	
Dream	61 (37%)	62 (37%)	
Torgersen	24 (15%)	23 (14%)	
Largo del pico (mm)	42,1 (4,9)	45,9 (5,4)	<0,001
Alto del pico (mm)	16 (1,8)	18 (1,9)	<0,001
Largo de la aleta (mm)	197 (13)	205 (15)	<0,001
Masa corporal (g)	3.862 (666)	4.546 (788)	<0,001

¹n (%); Media (DE)
²prueba chi cuadrado de independencia; t de Student

Figura 5

Largo de la aleta (mm)	197 (13)	205 (15)	<0,001
Masa corporal (g)	3.862 (666)	4.546 (788)	<0,001

¹ Recuentos (%); media (desviación estándar)

² Prueba Chi cuadrado; prueba T para muestras independientes

Uso de R y Youtube para reporte de protocolos: experiencia en laboratorio de física de suelos

La pandemia del COVID-19 ha afectado de diversas formas a los investigadores que trabajan en laboratorios. Las medidas implementadas por cada laboratorio o institución en muchas oportunidades prohibía el acceso a los laboratorios, lo que impedía la realización de múltiples experimentos. En otros casos, los accesos se permitieron de manera parcial, lo que generó nuevas formas de trabajo para continuar con las investigaciones que se llevan a cabo en estos laboratorios. Esto además afectó el acceso a los datos por los investigadores, por lo que se priorizó la disponibilidad de estos vía digitalización y consolidación de bases de datos.

En particular, en física de suelos, las experiencias que se realizan en laboratorio toman semanas, y existen periodos críticos en donde se requiere de supervisión de parte del investigador experto para dar continuidad a las experimentaciones. Las limitaciones de acceso, y la necesidad de continuar con las investigaciones han motivado a que se generen diversas formas de aprendizaje, aprovechando las múltiples posibilidades que existen hoy en día en los medios digitales de comunicación. Ya que parte de los datos fueron disponibles en un repositorio online, se debió capacitar a los investigadores en temáticas de manejo de datos con R. A partir de lo anterior, es que en el Laboratorio de Biofísica de Suelos se generó un Canal en YouTube para explicar diversas metodologías utilizadas en laboratorio, y también el uso de diversos paquetes en R para trabajar con los datos que se generan en el Laboratorio. Hay dos tipos de visualizaciones de datos que son complejas para los alumnos nuevos del Laboratorio: gráficos ternarios y perfiles de suelo. Usando los paquetes `{flipbookr}`, `{ggtern}` y `{ggplot2}`, se realizaron talleres “paso a paso”, los cuales quedaron a libre disposición en Youtube. La respuesta ha sido mejor que lo que se esperaba, generando más de 50 suscriptores al Canal de YouTube y más de 100 visualizaciones a los videos en un mes, además de colaboración con investigadores extranjeros. Debido al buen recibimiento, se continuará con este esquema didáctico para visualización y análisis de datos.

Un viaje a la ecología del movimiento a través de la minería de texto

Rocío Joo* Simona Picardi† Matthew E. Boone‡ Thomas A. Clay§
Samantha C. Patrick¶ Vilma S. Romero-Romero|| Mathieu Basille**

El campo de la ecología del movimiento ha experimentado un crecimiento sin precedentes en la última década. El desarrollo masivo de dispositivos que permiten seguir animales y personas (e.g. GPS, TDRs, acelerómetros, cámaras de video), y de herramientas de programación y análisis, han originado como consecuencia un mayor acceso a grandes volúmenes de datos, generación de resultados y publicación de artículos científicos. Quisimos revisar de manera cuantitativa el estado de la ecología del movimiento como ciencia, analizando diferentes aspectos como los temas investigados, los softwares y los dispositivos de seguimiento utilizados.

En esta presentación se describe la metodología utilizada para lograr este objetivo. A través de criterios de búsqueda que explicaremos brevemente, identificamos más de 8000 publicaciones científicas en inglés en ecología del movimiento realizadas en la última década. Para el análisis utilizamos varias técnicas de minería de texto. El énfasis en esta presentación será principalmente en dos enfoques: “diccionario” y modelamiento de temas. Un enfoque tipo “diccionario” fue utilizado para revisar aspectos como los softwares y los dispositivos de seguimiento. Para cada aspecto se definió un diccionario como una lista de categorías (e.g. “R”, “Matlab”), en la que, a su vez, cada categoría estuvo definida por una lista de términos (e.g. “R Software”, “R Development Core”). Para cada documento, de encontrarse uno de los términos en el texto (e.g. en la sección Materiales y Métodos o en el Resumen), el documento fue asociado a la categoría correspondiente al término. Para estudiar los temas investigados en los artículos se utilizó un enfoque distinto, puesto que ellos no fueron definidos a priori. La identificación de temas latentes y desconocidos se realizó mediante el ajuste de modelos de Asignación Latente de Dirichlet (Latent Dirichlet Allocation) a los resúmenes de las publicaciones. Estos modelos son esencialmente modelos bayesianos jerárquicos de tres niveles (temas, palabras, resumen) para documentos. El modelo define cada tema como una mezcla de palabras, donde ciertas palabras tienen mayor probabilidad de ocurrencia en ciertos temas. Como cada resumen podría estar compuesto por uno o más temas, las palabras utilizadas en el resumen reflejan los temas subyacentes.

Los modelos de Asignación Latente de Dirichlet nos permitieron caracterizar 15 temas desarrollados en ecología de movimiento. Asimismo, encontramos una creciente hegemonía de R respecto a los otros softwares y de GPS respecto a otros dispositivos de seguimiento.

Palabras clave:

minería de texto - ecología del movimiento - modelamiento de temas

*University of Florida, USA, rocio.joo@globalfishingwatch.org

†University of Florida

‡University of Florida

§University of Liverpool

¶University of Liverpool

||Universidad de Lima

**University of Florida

Sesión Datos espaciales

Soy naturalista y quiero pasear en mi país, ¿dónde hay más oportunidades de llenar vacíos de información?

LatinR

2021

Florencia Grattarola^{1,2}, Juan Manuel Barreneche²

1 JULANA ONG, Alarcón 1392, 11300, Montevideo, Uruguay

2 Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha-Suchdol, 165 00, Czech Republic

Keywords: biodiversidad, ciencia ciudadana, iNaturalist, vacíos de información, informática de la biodiversidad, Uruguay.

La biodiversidad a nivel global está disminuyendo a un ritmo sin precedentes. En este contexto, uno de los principales desafíos que enfrentan los países alrededor del mundo es poder medir cuantitativamente la biodiversidad y monitorear sus cambios. Sin embargo, en diversas regiones de Latinoamérica y el Caribe los datos de biodiversidad disponibles de manera abierta son limitados. Para poder tomar mejores decisiones basadas en evidencia es sumamente crítico revertir la falta de datos primarios sobre la distribución geográfica de las especies. Para esto, la ciencia ciudadana se presenta como una herramienta comunitaria transformadora. Una de las plataformas globales más usadas en Latinoamérica y el mundo es iNaturalist (inaturalist.org). Esta plataforma web y aplicación para celulares tiene como principal funcionalidad el registro de organismos en el tiempo y el espacio y reúne, en torno a estos datos, a la comunidad de naturalistas más grande del mundo. Una característica de los datos que provienen de la ciencia ciudadana es que suelen centrarse en áreas de fácil acceso y próximas a centros poblados, carreteras y áreas de interés (como por ejemplo, áreas protegidas). Para poder generar datos más diversos y mejorar el conocimiento de la biodiversidad, quienes observan (y quienes usan los datos) se verían muy beneficiados de tener a disposición una herramienta que les permita decidir a dónde ir a registrar organismos y qué grupos observar en función de maximizar su aporte.

Usando como base los datos ingresados en la plataforma iNaturalist para Uruguay (descargados el 21 de octubre de 2021), nos propusimos generar un mapa interactivo que ordene las áreas con déficit de datos de biodiversidad y nos permita resaltar aquellas en las que registros adicionales de biodiversidad podrían ser particularmente valiosos para llenar los vacíos de conocimiento. Una vez descargados los datos, filtramos los registros de organismos cautivos o cultivados, los que no habían sido identificados con grado de investigación y aquellos con coordenadas obscurecidas, y nos quedamos con las observaciones identificadas a nivel taxonómico de especie, pasando de 28,215 registros a 17,398 (61.7% del total). Luego, dividimos el país en grillas hexagonales de 500km² y calculamos una serie de métricas por grilla (intensidad espacial y temporal), usando los paquetes `geouy`, `sf` y `tidyverse`, para luego integrarlas en una medida de prioridad. Intensidad espacial fue calculada como la cantidad de registros por unidad de área, e intensidad temporal como la cantidad de registros por cada mes/año de registro. Ambas métricas fueron re-escaladas entre 0 y 1, sumadas y vueltas a re-escalar para generar nuestro índice de prioridad. Finalmente, las grillas fueron categorizados como de prioridad ‘Muy Alta’, ‘Alta’, ‘Media’, ‘Baja’, ‘Muy Baja’, y ‘Sin registros’, en función de cada grupo taxonómico. El sistema de puntaje de prioridad construido, no es absoluto sino relativo al conjunto de celdas. De esta manera, siempre será posible encontrar celdas con mayor o menor oportunidad de hacer aportes valiosos. Los datos fueron luego usados como base para crear una app en Shiny, usando los paquetes `shiny` y `leaflet`. La app permite visualizar el mapa de Uruguay con las grillas hexagonales coloreadas según el orden de prioridad generado (Figura 1). Además, quienes consulten el mapa podrán seleccionar el grupo taxonómico de interés y elegir una grilla, para la cual se desplegará información sobre: el área de la grilla, los

valores de intensidad espacial y temporal, la cantidad de especies registradas para esa grilla, el número de especies nuevas registradas en el último año y la proporción de éstas sobre el total registrado para esa grilla.

Haciendo uso de esta herramienta, usuarios y usuarias de la plataforma iNaturalist podrán decidir el destino de sus paseos en función de dónde registrar observaciones de la biodiversidad es más urgente y así contribuir a mejorar el conocimiento sobre la distribución de especies en el país. Esta herramienta podría replicarse en otras partes del mundo, además, nuevas métricas podrían ser propuestas y formas de visualizar los datos acordes a las regiones (e.g. resaltando rutas, áreas administrativas o zonas de interés). Esta es la primera vez que hacemos pública la herramienta por lo que esperamos poder recibir devoluciones para seguirla mejorando a futuro.

El código (incluido el de la Shiny app) puede encontrarse aquí: <https://github.com/bienflorence/LatinR2021>

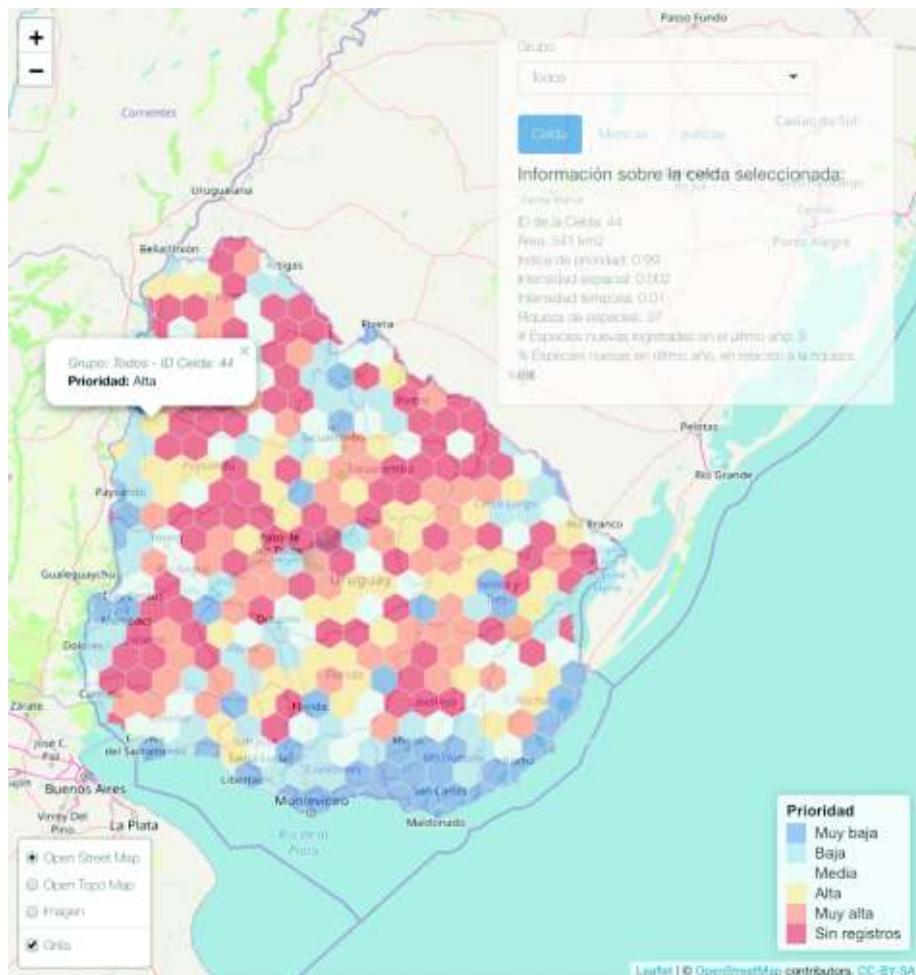


Figure 1: Visualización de la Shiny app

Un cuento digital desde R: cómo crear un relato situado con Leaflet

Natalia Morandeira

*Instituto de Investigación e Ingeniería Ambiental, Universidad Nacional de San Martín - CONICET.
General San Martín, Buenos Aires, Argentina. nmorandeira@unsam.edu.ar*

Escribo narrativa de ficción. Y además soy ecóloga, uso R y Sistemas de Información Geográfica. Hace poco se abrió la convocatoria de un concurso de cuento digital: se me ocurrió combinar estas artes y saberes para participar del certamen. Pensé en un relato situado. Cada escena sucede en un lugar de Argentina, cada párrafo está georeferenciado y las/los lectores pueden interactuar con el paisaje virtual. El cuento se llama "Detector de metales", está narrado en primera persona y es breve, de cinco párrafos. Si bien hay un orden ideal para leer el relato (que no es exactamente norte - sur), preferí no indicarlo: los párrafos pueden funcionar también como aguafuertes, retazos de la memoria de la narradora. En el proceso creativo, encontré que los lugares en los que elegía situar el relato eran disparadores, me permitían generar asociaciones libres con mi recuerdo de esos sitios o con el imaginario que tengo de ellos. Pienso que recorrer el mapa también puede ser un disparador para las/los lectores y que mezclar el orden de los párrafos puede generar nuevas historias. Desde el punto de vista de la escritura, otra ventaja es que el cuento puede crecer (sumar escenas, o incluso sumar puntos de vista si se agrega una capa adicional con la narración de otro personaje). Al trabajar con control de versiones en github, se podría armar también un proyecto de cuento colaborativo.

El cuento digital "Detector de metales" es un html originado a partir de una única diapositiva xaringan, con el título del cuento, el nombre de la autora (o seudónimo, mientras el cuento esté en evaluación en el concurso) y un mapa interactivo Leaflet. Seguramente se puede complejizar la visualización, pero preferí no explorar ese camino porque el lenguaje principal de mi contenido es la ficción narrada. Los datos necesarios son: un archivo vectorial de puntos con el ID de cada localidad y un archivo .txt con el relato, organizado en párrafos separados por algún delimitador (usé comillas y comas). En el .txt, se puede usar también lenguaje HTML para cuestiones de formato. El relato no puede incluirse desde el origen en el archivo vectorial, debido a que los campos de la tabla de atributos tienen una restricción de longitud. Además, tener el relato en un .txt separado permite corregir más fácilmente el relato.

El script usa RMarkdown y las librerías `{sf}`, `{dplyr}`, `{leaflet}` y `{htmltools}`. Los procesos realizados son: a) lectura del archivo vectorial y del relato; b) unión relacional de cada párrafo del relato a la tabla del objeto espacial; c) creación de un mapa base con uno o más proveedores de Leaflet; en mi caso usé Stamen.Watercolor como mapa base por defecto, por no tener texto, pero agregué otros mapas base como opciones seleccionables; d) despliegue de marcadores con las localizaciones de cada párrafo y etiquetas "pop-up" para cada párrafo del relato; e) algunas operaciones de archivos como renombrar el html producto del knitr y moverlo de directorio (docs/index.html; para publicar la página en githubpages).

Mientras creaba este cuento, encontré que hay varias posibilidades para contar historias con mapas, por ejemplo [leaflet-storymap](#) (ver la comparación con otros recursos en la sección [Compare with](#)). Mi elección priorizó un script con un flujo de trabajo sencillo y un producto en el que resalte la narrativa. Como ítems a mejorar, señalo la navegabilidad desde dispositivos móviles (el tamaño de tipografía adecuado para una computadora es chico en un teléfono, y resulta difícil agrandar la diapositiva) y la accesibilidad para personas ciegas, con visión reducida o con otras dificultades para navegar el mapa interactivo, que también parece ser un problema de los demás recursos. Aunque no soluciona totalmente el problema de accesibilidad, agregué en el pie de página un link con una versión alternativa accesible (lenguaje markdown), en la que comento la presentación original y presento el texto organizado por párrafos (se puede hacer click en el nombre de cada localidad).

Hoy el seudónimo "Silvia 79" es un homenaje a Silvia Matteucci, una de las primeras y más grandes ecólogas de paisajes de Argentina, más el número atómico del oro. Aquí les dejo el [repositorio](#) con el script que usé para generarlo y en el README.md encuentran un link al cuento en html y a la versión accesible. ¡Pasan y lean!

Palabras clave: Ficción; RSpatial; Leaflet; SIG; Cuento; Story Map; Visualización; Mapa

Tópico: Visualización de datos en R

Sesión Periodismo de datos, datos abiertos y visualización

JF EM DADOS: DEMOCRATIZANDO INFORMAÇÃO ATRAVÉS DA CIÊNCIA DE DADOS

Matheus Valentim e Marcello Filgueiras

A JFemDados

Com mais de 500 mil habitantes, o município de Juiz de Fora é um dos maiores do estado de Minas Gerais. Pólo educacional local, a cidade tem grande relevância na região, tendo vários municípios de menor porte dependendo diretamente de Juiz de Fora na saúde, na educação e no mercado de trabalho. Mesmo relevante, o município tem uma gestão por vezes omissa a questões relevantes e um debate público incipiente, que usa muito pouco de fontes oficiais e estatísticas.

Nosso objetivo é demonstrar, através de visualizações de dados públicos (mapas, gráficos e tabelas) todo tipo de temática que envolve o município. Objetivamos qualificar o debate das mais variadas maneiras, criando um jornalismo de dados local bem embasado e didático, que nutra o debate municipal com conteúdo quantitativo. Isso porque a Prefeitura, junto de outros órgãos municipais, apesar de disponibilizar os dados, o faz de maneira pouco organizada e de difícil consulta por parte da população. Nossa esforço é na direção contrária: extrair informações públicas e oficiais que ficam “arquivadas” no fundo de bases de dados ou de documentos oficiais e expô-las de uma maneira visual, acessível e didática. Tudo isso tendo como principal insumo a linguagem R, onde fazemos a maioria dos nossos trabalhos e também onde estamos desenvolvendo nosso blog via {blogdown} e um site interativo via {shiny}.

Nosso workflow e o uso do R

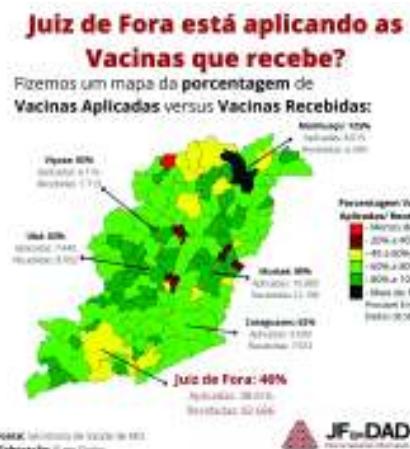
O fluxo de nosso trabalho envolve três grandes etapas: a extração/obtenção dos dados, o tratamento das bases e a disponibilização de visualizações e das bases tratadas para o público.

Os dados são extraídos de páginas oficiais de órgãos da cidade, como o site da Prefeitura e o site da Câmara Municipal, e também de páginas de órgãos públicos brasileiros, como IBGE e DataSUS. A extração envolve uso de técnicas de web scrapping através de pacotes como {httr}, {xml2}, {rvest} e {pdftools}, e é feita principalmente nos casos em que as fontes oficiais só disponibilizam pdfs ou tabelas virtuais dos dados, como foi o caso com óbitos e casos de COVID-19. Outra técnica de extração é a utilização da {basedosdados}, filtrando as bases nacionais para a cidade de Juiz de Fora. O tratamento dos dados consiste em transitar as bases para um formato tidy, sempre com o critério de padronizar as variáveis presentes em colunas. A visualização e a disponibilização das bases para o público é talvez a parte mais importante do nosso trabalho. Atualmente, as nossas visualizações são feitas através do {ggplot2} e do Flourish. Para atingir nosso público, temos publicações diretas em mídias sociais, como Twitter e Instagram (@jfemdados), além de uma parceria com a Tribuna de Minas, jornal estadual que publica os dados da cidade diretamente em cadernos impressos e virtuais. Todo nosso material usado, incluindo as bases que criamos no processo, as visualizações e os R scripts usados ficam no nosso [Github](#).

Temáticas abordadas e material produzido

Trazemos desde informações “quentes” e notícias que estão em pauta, como foram casos, ocupação de leitos e vacinação contra COVID-19, até dados ilustrativos “frios” como emissão de poluentes, focos de queimadas na cidade e censo escolar da cidade. Como exemplo desses temas quentes, onde os dados não estavam acessíveis para o público, temos o caso das mortes municipais por COVID-19. Caso emblemático, que envolveu

transformar um PDF 46 páginas de texto corrido, em um data frame de 1700 linhas, sendo necessário separar inicialmente cada caso por linhas, e cada variável em colunas específica, com uso denso de regex, {stringr} e {tidyverse}. Uma vez separadas, o processo de padronização das variáveis como a variável “comorbidades” foi necessário e feito também com uso de {stringr}. Para a padronização de datas, o {lubridate} foi usado. Tudo exigiu cerca de 400 linhas de código. Um [RMarkdown](#) desse processo está disponível, bem como o [primeiro post do Twitter](#). A [segunda postagem do Twitter](#) destacada reflete a análise da eficiência da ação da prefeitura em aplicar os estoques das vacinas disponibilizadas, comparando com outras cidades do Estado, usando mapas do pacote {geobr} e {sf}.



Por fim, apresentamos o nosso carro chefe: a análise da Câmara Municipal. Nesse projeto exploramos a atividade dos parlamentares e transparecemos isso à população. Buscamos demonstrar tanto o que está sendo votado, proposto e discutido por cada um dos vereadores, separando os por tema. Para obter os dados, usamos de web scraping com {httr} e {rvest} para raspar o site da câmara municipal para obter projetos de lei (PLs), moções e requerimentos. Para classificá-los, usamos regex via {stringr} para buscar por temáticas específicas, buscando por palavras ou expressões-chave recorrentes. Além disso, partindo da literatura da ciência política (CUNHA, Patrick Silva. “O Poder Legislativo Municipal: Estrutura, Composição e Produção”), analisamos qualitativamente os projetos que teriam caráter “simbólico”, de pouca relevância prática para a população, tais como criação de dias comemorativos e mudanças em nomes de ruas. O exemplo abaixo pode ser conferido [aqui](#).



A Base dos Dados+: acesso fácil a dados públicos de qualidade

Rodrigo Dornelles , Matheus Valentim , Fernanda Scovino , Pedro Cavalcante

Abstract Apresentamos o projeto “Base dos Dados Mais” da ONG Base dos Dados e seu novo pacote R, recentemente disponibilizado no CRAN. Esperamos poder apresentar para a comunidade R latino-americana novas possibilidades de análise de grandes bases de dados públicos, capazes de potencializar o uso de dados abertos. Nossa plataforma, que se utiliza do motor do Google BigQuery, reúne bases de interesse público já catalogadas, limpas, organizadas e compatíveis entre si. Pretendemos demonstrar aplicações do projeto com dados brasileiros e regionais e discutir formas de enriquecer e ampliar nosso impacto.

Palavras clave: dados públicos - dados abertos - gestão pública - políticas públicas baseadas em evidências - big data - Google BigQuery - R Stats

O que é a BD?

A [Base dos Dados](#) (BD) é uma organização sediada no Brasil que tem como missão universalizar o uso de dados, permitindo que a distância entre uma pessoa e uma análise seja apenas uma boa ideia. A BD [já mapeou mais de 950 conjuntos de dados de 504 organizações](#).

Para atingir esse objetivo, a BD **cataloga** e **trata** bases de dados de interesse público. A catalogação consiste em gerar metadados e informações sobre uma dada fonte dos dados. Nesse processo, indica-se aos usuários onde eles podem encontrar dados de uma dada temática e informa-se o que se espera encontrar nesses dados. Tudo isso através do mecanismo de busca do site.

O tratamento de dados é o carro chefe da organização. Para essa etapa, a Base dos Dados conta com uma equipe de assistentes e voluntários que cotidianamente limpam e organizam bases de dados complexas, transformando essas bases complicadas em estruturas simples e organizadas e disponibilizando-as em seu datalake público.

Essa equipe de colaboradoras(es) realiza a limpeza e adaptação dos dados utilizando linguagens como R, Python e Stata, de acordo com o conhecimento de cada pessoa: todos os scripts são disponibilizados, juntamente com os dados originais, no repositório do projeto no GitHub a fim de assegurar a reproduzibilidade e segurança no processo de limpeza.

Dessa forma, uma base que originalmente está segmentada em vários anos, tem nomes de variáveis não inteligíveis e que talvez seja muito grande para uso direto de pessoas comuns (o que é, em regra, muito comum em se tratando de dados públicos brasileiros) se torna, após o processo, uma base fácil e amigável.

Todos esses conjuntos de dados limpos são hospedados na nuvem do Google Cloud e podem ser facilmente acessados pelo nosso pacote do R [{basedosdados}](#). Essas bases já tratadas são coletivamente chamadas de **BD+**, e recebem essa tag no site, indicando que já estão limpas e prontas para uso.

Em resumo, a Base dos Dados disponibiliza gratuitamente um enorme datalake com bases já tratadas, normalizando variáveis chave entre tabelas, já em formato tidy e prontas para uso. É possível acessar dados de economia, saúde, segurança pública, orçamento, demografia, política, meio ambiente entre tantos outros, por meio do pacote [{basedosdados}](#) disponível no CRAN.

Mas por que isso é importante?

A importância disso é tornar trivial o manuseio de bases de dados complexas, porém muito úteis para a análise e elaboração de políticas públicas melhores, para a fiscalização da sociedade, jornalismo de dados e tantas outras aplicações. E, lembremos, de forma que seja possível cruzar entre si dados de eleições, economia, saúde, segurança etc já que as chaves e identificadores são padronizados em todas elas.

Embora muitas informações públicas tenham tido o acesso facilitado nos últimos anos, a leitura e manuseio delas ainda é complexo, instável e trabalhoso para a maior parte das pessoas interessadas.

Também, encontramos dificuldades em acessar e poder manusear bases de dados extremamente volumosas, com dezenas de GB de espaço em disco, como os dados do [Censo Demográfico](#), a [Pesquisa Nacional por Amostra de Domicílios \(PNAD\)](#), [Relação Anual de Informações Sociais \(RAIS\)](#), [micródados de vacinação contra COVID-19](#), [Censo Escolar](#) e outras. Esses dados, embora em tese acessíveis, são impraticáveis: além da dificuldade em interpretar e organizar os dados, há exigência de alta capacidade computacional, indisponível para a maior parte dos usuários.

Assim, através da BD+, é possível acessar facilmente esses dados, economizando muitas horas de trabalho que precisariam ser dedicadas a busca, abertura e limpeza desses dados.

O pacote {basedosdados}

Disponível no CRAN desde abril de 2021, o pacote concede acesso direto à API do BigQuery (mediada pelo {bigrquery}), permitindo sejam realizadas operações através R diretamente no serviço do Google, recebendo o resultado das operações localmente: Por exemplo:

```
# definir o billing_id correspondente ao projeto do BigQuery
basedosdados::set_billing_id("rfdornelles-bq")

# Download direto -----
basedosdados::download(query =
  SELECT * FROM `basedosdados.br_ana_atlas_esgotos.municipio`
  WHERE sigla_uf = 'AC'
  ,
  path = "esgotos_exemplo.csv")

# Executar query SQL -----
query <- "
  SELECT * FROM `basedosdados.br_ana_atlas_esgotos.municipio`
  WHERE sigla_uf = 'AC'
  "

esgotos_acre <- basedosdados::read_sql(query)
```

Conciliando todo o poder do R com o do BigQuery, o pacote a partir de sua versão 0.1.0 publicada em setembro/2021 faz uso do pacote {dbplyr} do {tidyverse}, de modo a permitir que as bases remotas sejam manuseadas mesmo sem qualquer conhecimento de SQL por parte dos usuários.

Basta criar uma variável usando uma das funções do pacote, que conectará com a base de dados remota, disponível no BiQuery, e que poderá ser manipulada usando as funções select, filter, mutate, summary entre outras e posteriormente baixadas para a memória ou para o disco.

```
# definir a tabela que vou operar:
nome_tabela <- "br_mc_auxilio_emergencial.microdados"

# fazer a conexão remota
base_remota <- basedosdados::bdplyr(nome_tabela)

# valor médio do benefício pago por estado
# as operações são executadas apenas com comandos do tidyverse
tabela_auxilio <- base_remota %>%
  select(mes, sigla_uf, valor_beneficio, enquadramento) %>%
  group_by(sigla_uf) %>%
  summarise(
    valor_total = sum(valor_beneficio, na.rm = TRUE),
    qnt_beneficiarios = n()
  )

# coletar os dados após concluída a análise
tabela_auxilio_coletada <- basedosdados::bd_collect(tabela_auxilio)

# ou salvar em disco
basedosdados::bd_write_rds(tabela_auxilio, path = "dados_auxilio.rds", overwrite = TRUE)
```

Neste exemplo, utilizando os verbos tradicionais do {tidyverse}, foi possível em pouco menos de 1 minuto realizar uma análise na base do Auxílio Emergencial pago no Brasil, contendo literalmente milhões de linhas.

Assim, mesmo iniciantes podem acessar, filtrar, cruzar, pivotar e organizar os dados utilizando a sintaxe comum do R e com as funções largamente conhecidas do {dplyr}. Ao final, os dados poderão ser coletados já no formato desejado para elaboração de gráficos, modelos, aplicativos shiny e o que a imaginação permitir.

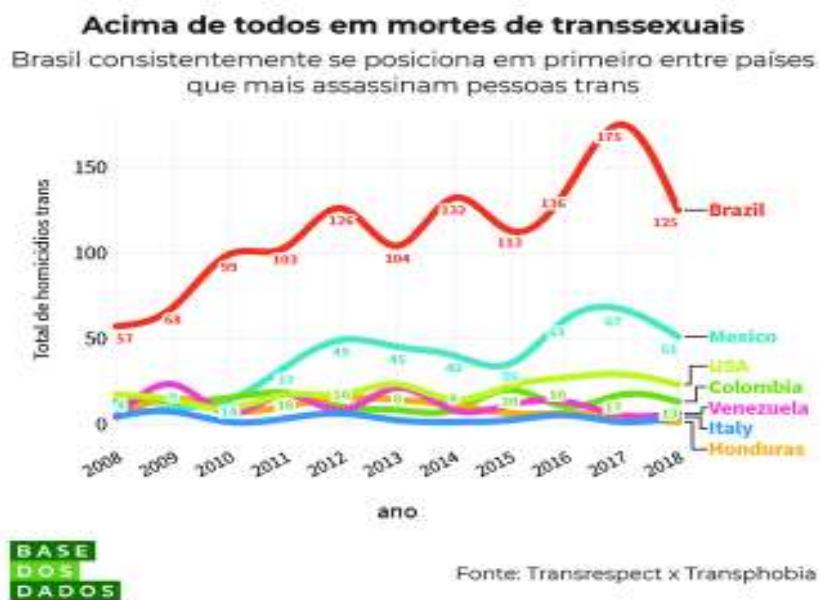
Conclusão: Algumas aplicações

Pessoas da academia, jornalistas, ativistas, pessoas da gestão pública etc podem se dedicar ao que mais importa: a análise, já que todo o trabalho de curadoria, atualização e limpeza já foi realizado por nossa comunidade.

Para que se tenha ideia do poder de tudo isso, agora é possível analisar uma [base de dados de 250GB como a da RAIS](#) (que contem dados trabalhistas e de atividade econômica de todo o país) em poucos segundos, utilizando os comandos tradicionais do {dplyr} e podendo, por exemplo, relacioná-la com dados eleitorais disponíveis na BD+ ou mesmo dados da pandemia de COVID [disponíveis em bases públicas do Google](#).

Em breve, será possível também buscar por tabelas e metadados por meio do próprio pacote, tudo isso sem necessário sair em nenhum momento do R.

Outros vários exemplos de análises que já foram realizados pela equipe com os dados da BD+ estão disponíveis no [repositório do GitHub](#).



Abaixo, algumas análises recentemente realizadas:

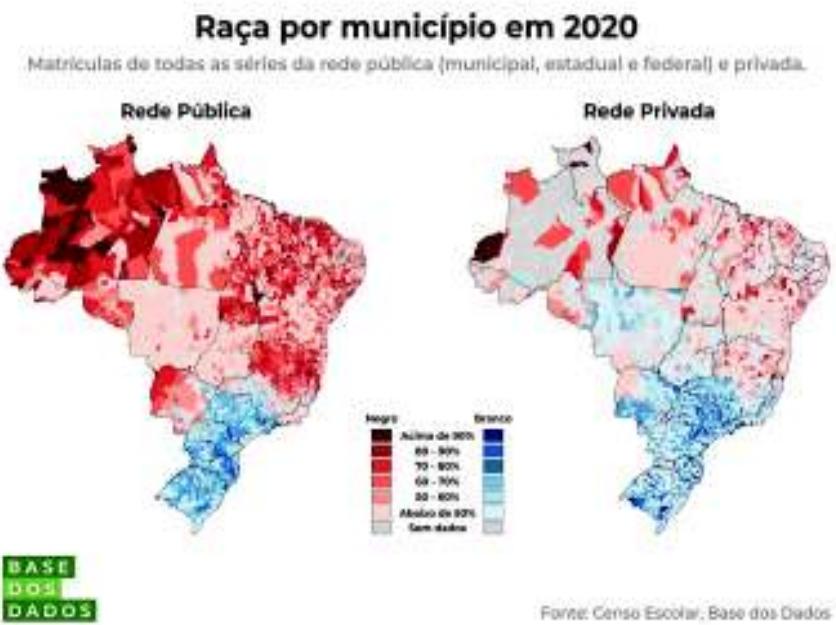


Figura 1: Gráfico comparando raça de estudantes das redes de educação pública e privada



Figura 2: Quantidade de funcionários(as) públicos(as) federais vinculados aos três poderes ao longo dos últimos anos

Alavancando o poder do RMarkdown com as linguagens da Web e D3.js para produzir histórias de dados envolventes sobre Finanças Públicas

Anônimo

Palabras clave: R Markdown - Data Visualization - Data Storytelling - Web Standards - r2d3 - Public Finance - Government

URL da Aplicação: <https://tchiluanda.github.io/DC/>

Introdução

O R Markdown permite integrar a análise e a comunicação de dados, costurando texto e códigos numa míriade de formatos de saída, desde artigos e slides até livros e aplicações web interativa.

A simplicidade da linguagem markdown, o poder do R e a versatilidade de formatos de saída fazem com que usuários de R possam produzir, num mesmo ambiente, um fluxo completo e reproduzível, desde a importação de dados brutos até a construção de um documento visualmente atraente (como ilustra Allison Horst em: https://github.com/allisonhorst/stats-illustrations/blob/master/rstats-artwork/rmarkdown_wizards.png).

Em particular, o formato de saída HTML permite que usuários sem conhecimentos de web design produzam páginas web completas, com design agradável e prontas para publicação e comunicação do resultado de suas análises.

Indo além

No entanto, é possível criar histórias de dados ainda mais envolventes e atraentes mergulhando um pouco mais fundo nas linguagens e tecnologias envolvidas na geração de documentos HTML para a Web a partir do R Markdown.

Uma das características do R Markdown que lhe conferem tamanha versatilidade é a possibilidade de inclusão de “chunks” de códigos escritos em outras linguagens além de R.

Por exemplo, para personalizar um pouco mais a aparência da página web, é fundamental usar CSS (“Cascading Style Sheets”), a linguagem usada para descrever a apresentação de páginas Web, incluindo cores, layouts e fontes. Códigos em CSS podem ser incluídos no próprio arquivo R Markdown. No documento HTML gerado, esses chunks serão incluídos como blocos `<style>`.

Para incluir novos elementos visuais que possam ser formatados pelo código em CSS (boxes explicativos em meio ao texto, como no caso do nosso projeto), é possível utilizar a própria sintaxe do Pandoc, em muitos casos, como no caso de “fenced divs” e “bracketed spans”. Para os demais casos, existe a possibilidade de se incluir o código HTML puro no corpo do texto.

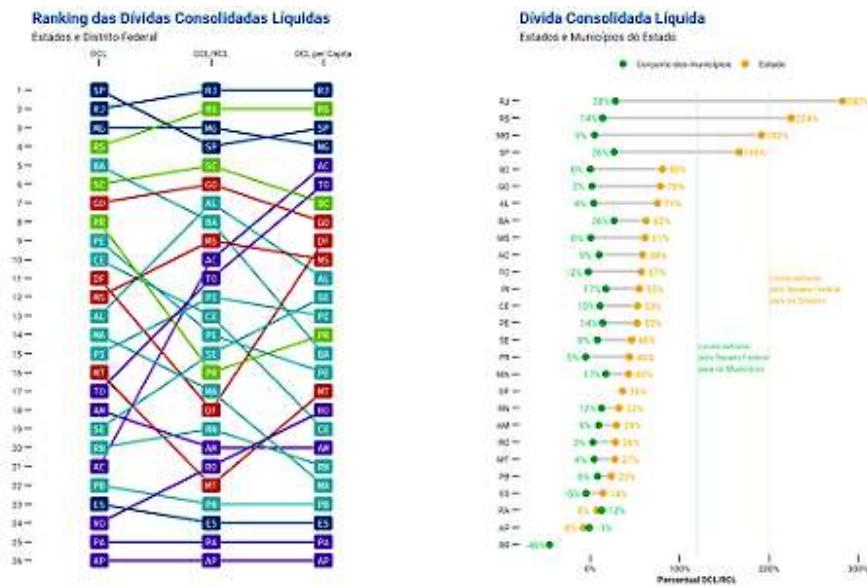
Além disso, é possível acrescentar interatividade à página com o uso de Javascript, cujos códigos também podem ser incluídos como chunks. Por fim, graças ao pacote r2d3, pode-se ainda incluir chunks de D3.js, a principal biblioteca em Javascript para construção de visualizações interativas e sem formatos pré-determinados.

O Projeto

Neste projeto, utilizamos as demonstrações financeiras declaradas ao Tesouro Nacional pelo governo federal, pelos governos estaduais e por 93% dos 5.570 municípios brasileiros. Com base nesses dados, construímos com

R Markdown uma página Web apresentando as informações das dívidas públicas dos governos brasileiros numa estrutura narrativa e envolvente. Para isso utilizamos diversos recursos.

Toda a importação, preparação, processamento, agrupamento e tratamento dos dados foi feita com R, utilizando extensivamente os pacotes do tidyverse. Para os gráficos estáticos, utilizamos ggplot2.



Para encontrar um equilíbrio entre simplicidade e rigor conceitual no texto, deixamos o texto principal mais simples, mas incluímos a possibilidade de o usuário clicar em termos mais técnicos para que a página exiba quadros com explicações adicionais.

Isso é possível com uma combinação do uso da sintaxe do Pandoc para criar elementos customizados, chunks de CSS (para a formatação dos elementos) e Javascript (para acrescentar interatividade).

Além disso, como mencionado e indicado acima, como experimento fizemos uso do pacote r2d3 para construir gráficos em D3 diretamente no R Markdown. Uma vantagem de usar essa abordagem é permitir a utilização de gráficos interativos no próprio formato HTML, sem a necessidade de um servidor Shiny.

Finalmente, utilizamos o chunk de CSS para definir o layout geral da página, fontes, cores, espaçamentos, margens etc.

Comunicando Finanças Públicas de uma forma mais amigável para a sociedade

Assim, para poder combinar “Finanças Públicas” com “histórias de dados envolventes”, ou seja, para tornar esse assunto menos árido e mais interessante, decidimos fazer uso do potencial das linguagens dos padrões da Web, reunidas no ambiente ao qual já estamos tão bem familiarizados para realizar nossas análises de dados.

Dessa forma, é possível, num único documento de texto simples .Rmd, escrever textos, importar e manipular dados com R, gerar visualizações complexas e atraentes com ggplot2, construir elementos visuais com markdown e HTML, formatar a aparência desses elementos com CSS, incluir interatividade com Javascript, embutir visualizações interativas e completamente customizadas com D3.

Além da versatilidade, a vantagem do R Markdown é a de integrar todo o fluxo desde a importação dos dados originais até a criação da página Web final. Assim, nas atualizações futuras dos dados, basta “costurar” novamente o arquivo .Rmd para se obter a página Web atualizada.

Em nossa apresentação, pretendemos mostrar os detalhes da implementação e compartilhar aprendizados e dicas que podem inspirar e contribuir para a construção de história de dados visualmente atraentes com R Markdown. O código está disponível em: <https://github.com/tchiluanda/DC>, e a aplicação está on-line em: <https://tchiluanda.github.io/DC/>.

Sesión R en producción, computación y flujos de trabajo

R en producción: aprendizajes, retos y mejores prácticas

Ángel Escalante , Nancy Morales

Abstract Con más de tres años de experiencia trabajando como desarrolladores en R, compartimos las experiencias, retos y aprendizajes a los que nos hemos enfrentado en la industria al usar R como un medio para el procesamiento de datos, análisis y generación de reportes. Asimismo, proponemos un conjunto de puntos, a los que denominamos *imprescindibles*, tales como: apegarse a un esquema de trabajo de desarrollo (p. ej. SCRUM), elegir una versión de software adecuada a las necesidades como Microsoft R Open o R, emplear librerías que optimicen el uso de memoria y permitan crear paquetes escalables e implementar testing (unitario y de regresión) riguroso para mayor confiabilidad. Estos puntos nos han ayudado a mantener modelos estadísticos y comunicar resultados orientados a clientes en un esquema automatizado y confiable. De igual forma, dicho esquema nos ha permitido tener tiempos de respuesta rápidos ante bugs en producción y así reducir el impacto en costos que estos puedan tener.

Palabras clave: R - producción - paquetes - metodología

Introducción

Las organizaciones hoy en día necesitan sistemas que ejecuten continuamente código para apoyar la toma de decisiones de los clientes tanto internos como externos. Esto implica asegurar un funcionamiento correcto y eficiente de dicho sistema para cumplir con la demanda de los usuarios. Por lo tanto, características como contar con código legible, estable y escalable, uso de recursos y tiempo de ejecución (entre otros) son fundamentales en ambientes de producción. Estas características se traducen en evitar costos por bugs inesperados y generar satisfacción para los usuarios. Las compañías han volteado a ver las herramientas Open Source como claves para su desarrollo. Hoy en día, grandes compañías transnacionales como Netflix, AT&T, PayPal, entre otras, consideran software open source como pieza fundamental en su funcionamiento. R es un lenguaje de código abierto para estadística, visualizaciones y machine learning. Con R puedes generar desde una tabla resumen de datos hasta un dashboard que permita a los usuarios/clientes revisar y analizar su información en tiempo real. Sin embargo, al no ser un lenguaje multipropósito, ¿Es R un lenguaje que puede ser usado en Producción?

Después de haber trabajado en el sector privado y dar mantenimiento a paquetes en R que ejecutan cerca de 3,500 análisis mensuales dentro de la organización para entrega de reportes, modelos e incluso dashboards de análisis, proporcionaremos nuestras experiencias, retos y aprendizajes sobre nuestro uso de R en una organización de alta demanda.

Flujo de trabajo como desarrollador

Existen metodologías de trabajo en desarrollo de software denominadas “*metodologías Ágiles*”, las cuáles son esenciales conocer para trabajar en un equipo de desarrollo y, sobre todo, a distancia. Trabajamos bajo la metodología ágil SCRUM, la cual está diseñada para proyectos en entornos de alta demanda, de inmediatez, que busca la innovación y la productividad. Esto implica, organización y coordinación entre los desarrolladores con la persona que tiene la responsabilidad de definir los alcances y expectativas de nuevos desarrollos, actualizar desarrollos previos y definir proyectos de innovación. En el flujo de trabajo, se establecen olas de trabajo (mejor conocidos como sprints) que son períodos de 3 semanas, cada una con sus etapas de desarrollo: definición de features y/o bugs del sprint, desarrollo (escribir código), testing, presentación de resultados y retroalimentación del sprint.

Lo imprescindible para R en producción

En nuestra experiencia, hemos encontrado puntos que encontramos imprescindibles para ser un buen desarrollador en R en un ambiente de producción:

- **Adoptar una metodología de trabajo ágil.** Apoyarse de herramientas y paquetes que te permitan seguir el flujo de la metodología lo más fácil posible. Esto implica, auxiliarnos de herramientas de código colaborativo y control de versiones, Git, Azure DevOps, Amazon Web Services, Google Cloud o GitHub.
- **Establecer una versión de R** en la cuál correrán los análisis, aplicaciones o producto el final que se quiera entregar. Esto hará que los paquetes funcionen de forma esperada y reducirá el riesgo de enfrentar bugs por actualización de versiones.

- **Uso de paquetes de desarrollo** especializados como **devtools**, **usethis** y **testthat**. Así como seguir las mejores prácticas en desarrollo de paquetes propuestos por Jenny Bryan y Hadley Wickham en su libro R Packages (Hadley, 2015)¹.
- **Código legible, claro y documentado.** A veces, es mejor tener más líneas de código entendibles, que un código de pocas líneas, pero críptico (Dustin & Trevor, 2011)².
- **Crear un ecosistema de paquetes** intercomunicados y con propósitos específicos. Esto ayudará a simplificar todo el flujo del análisis que se quiere entregar. Si el propósito de la solución involucra recibir, analizar y presentar datos, tener un paquete dedicado a cada uno de estos pasos es lo deseable.
- **Memoria y uso de recursos.** Los recursos siempre son limitados, escribir código que sea eficiente en términos de memoria es fundamental. Revisaremos algunos paquetes, como **data.table**, que pueden ser de ayuda cuando se manejan volúmenes de datos grandes.
- **Regression testing.** Estamos trabajando con un paquete que entrega un reporte, después de un mes de desarrollo, ¿este reporte no ha sido involuntariamente alterado? ¿el nuevo código interfiere el anterior? Además del unit test, el cual es obligatorio si queremos evitarnos problemas de malfuncionamiento del paquete, el propósito del regression test toma el resultado de un producto final que definido como aceptable y comparamos con el resultado después de haber hecho diversos cambios. Esto asegurará que el código sigue produciendo los resultados esperados.

Conclusiones

Los ambientes de producción no son exclusivo de lenguajes multi-propósito como Python, JavaScript, etc. R tiene todo lo necesario para poder ser usado como lenguaje backend en aplicaciones web, herramienta de análisis principal y de reporteo en ciencia de datos entre muchas cosas más. Quienes venimos de escuelas de matemáticas quizá estemos más familiarizados con R que con otros lenguajes comunes en la industria de software y conceptos técnicos del desarrollo. Sin embargo, es la comunidad de R y los múltiples recursos de fácil acceso son los que permiten adentrarse en el mundo de DevOps e iniciar un camino como desarrollador en R en específico. El camino no es sencillo y habrá que aprender bajo prueba y error, pero con el tiempo entender estas capacidades, adoptar una metodología y aprender de otros desarrolladores, permitirá ayudar a la organización sin la necesidad de salir de R.

Referencias

- ¹ Wickham, Hadley. 2015. "R Packages". USA: O'Reilly
² Boswell, Dustin, and Foucher, Trevor. 2011. "The Art of Readable Code: Simple and Practical Techniques for Writing Better Code." USA: O'Reilly

Analogsea: Using R for Big Data Analytics

Introduction

Analogsea is a community project created by and for statisticians from the R community. The package's goal is to provide analytical capacity for Lesser Developed Countries and users that don't have access to large computational power. This package provides an interface to DigitalOcean, a cloud service such as AWS, Linode and others, for R users, but at a much lower cost and providing more gradual scalability options.

Even when DigitalOcean provides an API, computing across clusters composed by virtual servers is not straightforward, since vanilla Ubuntu servers ('droplets' in DO jargon) require additional configurations to install R itself and numerical libraries such as OpenBLAS.

Cutting edge research methods are really computationally intensive. As an example, on my laptop, using all CPU cores, a bayesian model can take up to 45 minutes to run one set of models, and with all the different model specifications and robustness checks, it takes hours to run the complete analysis.

By scaling computing hardware, this can be reduced to less than five minutes, but not all universities and research centers in Latin America and Asia Pacific have access to servers, and analogsea provides the option to rent a server and access to large computational resources to fit large scale models even on a tight budget.

However, because R packages require additional configuration, and the tradeoff between setup times and Kubernetes configuration versus just doing it all manually is not favorable. This is why analogsea also implements readily available server side solutions to use R out-of-the-box.

Fitting models in the cloud

Here I'll provide an elementary example that benefits from distributed computing, by fitting an elemental regression for a large taxi fares dataset.

The model to fit shall be of the form

$$\text{Total Amount}_i = \beta_0 + \beta_1 \text{Trip Distance}_i + \varepsilon_i$$

The error is of the form $\varepsilon \sim N(0, \sigma^2)$ and regardless of the simplicity of this OLS model, it can be fitted on most common laptops since it's a ~14 million rows dataset for a single month, and therefore fitting a regression for a complete year would be unfeasible.

You can have a look at the NYC Taxi dataset, which is stored in DigitalOcean Spaces with S3 compatible features. You can copy and explore the data locally like this.

```

space <- arrow::S3FileSystem$create(
  anonymous = TRUE,
  scheme = "https",
  endpoint_override = "sfo3.digitaloceanspaces.com"
)
# just 2009/01 for a quick exploration
try(dir.create("~/nyc-taxi/2009/01", recursive = T))
arrow::copy_files(space$path("nyc-taxi/2009/01"), "~/nyc-taxi/2009/01")

```

One interesting exercise is to explore how trip distance and total amount are related. You can fit a regression to do that! But you should explore the data first, and a good way to do so is to visualize how the observations are distributed.

```

library(arrow)
d <- open_dataset(
  "~/nyc-taxi",
  partitioning = c("year", "month")
)
dreg <- d %>%
  select(trip_distance, total_amount) %>%
  collect()
dreg
# # A tibble: 14,092,413 x 2
#   trip_distance total_amount
#       <dbl>        <dbl>
# 1 2.63          9.40
# 2 4.55         14.6
# 3 10.4          28.4
# 4 5             18.5
# 5 0.400         3.70
# # ... with 14,092,408 more rows

```

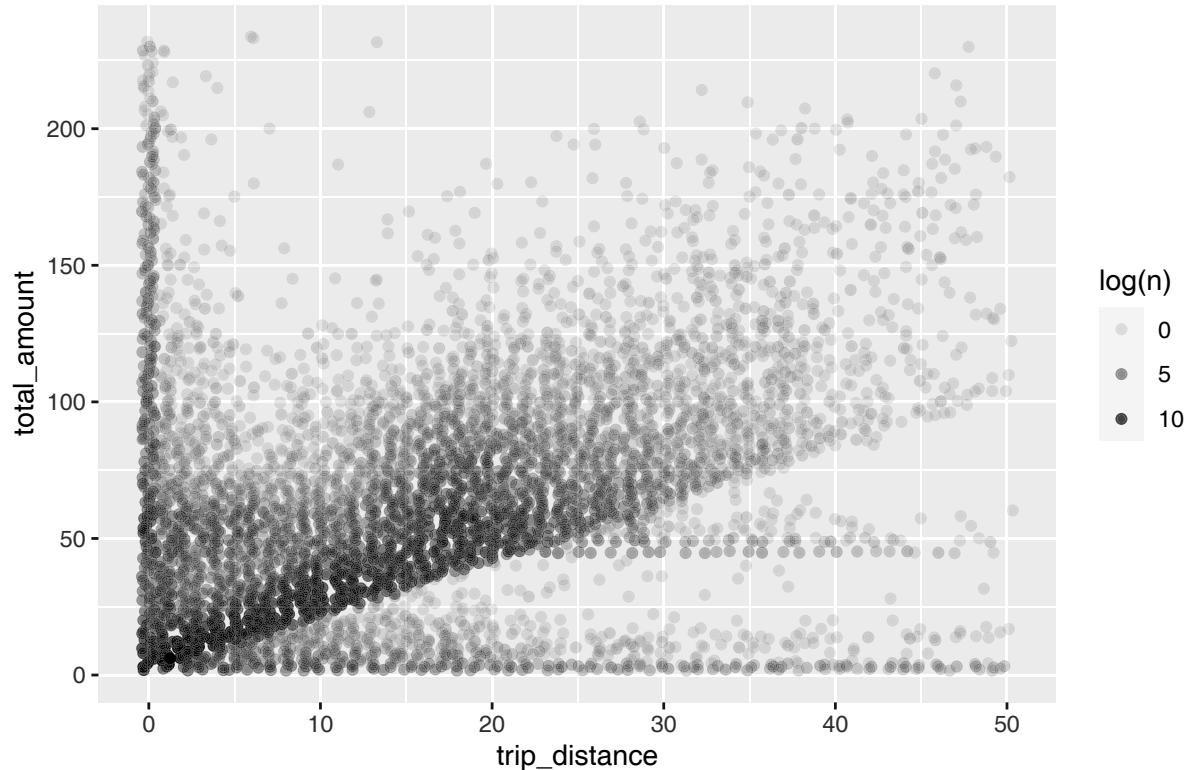
As this is a ~14 million rows for a single month, it will be hard to plot without at least some aggregation. There are repeated rows so you can count to get the frequencies for the pairs (distance, amount), and even round the numbers before doing aggregation since the idea is to explore if there's a trend.

```

dplot <- dreg %>%
  mutate_if(is.numeric, function(x) round(x,0)) %>%
  group_by(trip_distance, total_amount) %>%
  count()
dplot

```

Exploring distance versus total amount



There is not an obvious trend in this data, but it's still possible to obtain a trend line.

```
ggplot(dplot, aes(x = trip_distance, y = total_amount, alpha = log(n))) +
  geom_jitter() +
  labs(title = "Exploring distance versus total amount")
```

You can fit a model with Gaussian errors such as the next model with intercept.

```
summary(lm(total_amount ~ trip_distance, data = dreg))
# Call:
#   lm(formula = total_amount ~ trip_distance, data = dreg)
#
# Residuals:
#   Min     1Q     Median      3Q     Max 
# -122.736 -1.242    -0.542    0.509  228.125 
# 
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)    
# (Intercept) 4.0253576  0.0013786   2920 <2e-16 ***
# trip_distance 2.4388382  0.0003532   6905 <2e-16 ***
# ---      
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Residual standard error: 3.909 on 14092411 degrees of freedom
# Multiple R-squared:  0.7718, Adjusted R-squared:  0.7718
# F-statistic: 4.767e+07 on 1 and 14092411 DF, p-value: < 2.2e-16
```

Let's say we are interested in fitting one model per month and then average the obtain coefficients to smooth a potential seasonal effect (e.g. such as holidays, Christmas, etc).

To do this you can create 1,2,...,N droplets and assign different months to each droplet and read the data from S3, to then get the regression results to your laptop.

You can start by loading the analogsea package and create two droplets that already include arrow. The next code assumes that you already registered your SSH key at digitalocean.com.

```
library(analogsea)
s <- "c-8" # 4 dedicated CPUs + 8GB memory, run sizes()
droplet1 <- droplet_create("RDemo1", region = "sfo3", size = s,
                           image = "rstudio-20-04", wait = F)
droplet2 <- droplet_create("RDemo2", region = "sfo3", size = s,
                           image = "rstudio-20-04", wait = T)
```

What if you need to install additional R packages in the droplets? Well, you can control that remotely after updating the droplet status and install, for example, eflm to speed-up the regressions.

```
droplet1 <- droplet(droplet1$id)
droplet2 <- droplet(droplet2$id)
install_r_package(droplet1, "eflm")
install_r_package(droplet2, "eflm")
```

Before proceeding, you need the droplet's IPs to create a cluster that will use each droplet as a CPU core from your laptop.

```
ip1 <- droplet(droplet1$id)$networks$v4[[2]]$ip_address
ip2 <- droplet(droplet2$id)$networks$v4[[2]]$ip_address
ips <- c(ip1, ip2)
```

Now you can create a cluster, and you need to specify which SSH key to use.

```
library(future)
ssh_private_key_file <- "~/.ssh/id_rsa"
cl <- makeClusterPSOCK(
  ips,
  user = "root",
  rshopts = c(
    "-o", "StrictHostKeyChecking=no",
    "-o", "IdentitiesOnly=yes",
    "-i", ssh_private_key_file
  ),
  dryrun = FALSE
)
```

The next step is to create a working plan (i.e. 1 droplet = 1 ‘CPU core’) which allows to run processes in parallel and a function to specify which data to read and how to fit the models in the droplets.

```
plan(cluster, workers = cl)
fit_model <- function(y, m) {
  message(paste(y,m))
  suppressMessages(library(arrow))
  suppressMessages(library(dplyr))
  space <- S3FileSystem$create(
    anonymous = TRUE,
    scheme = "https",
    endpoint_override = "sfo3.digitaloceanspaces.com"
  )
  d <- open_dataset(
    space$path("nyc-taxi"),
    partitioning = c("year", "month")
  )
  d <- d %>%
    filter(year == y, month == m) %>%
    select(total_amount, trip_distance) %>%
    collect()
  fit <- try(eflm::elm(total_amount ~ trip_distance, data = d, model = F))
  if (class(fit) == "lm") fit <- fit$coefficients
  rm(d); gc()
  return(fit)
}
```

The fit_model function is a function that you should iterate over months and years. The furrr package is very efficient for these tasks and works well with the working plan defined previously, and the iteration can be run in a very similar way to purrr.

```
library(furrr)
fitted_models <- future_map2(
  c(rep(2009, 12), rep(2010, 12)), rep(1:12, 2), ~fit_model(.x, .y))

fitted_models[[1]]
# (Intercept) trip_distance
# 4.025358     2.438838
```

In the previous code, even when the iteration was made with two years there are some clear efficiencies. On the one hand, the data is read from S3 without leaving DigitalOcean network, and any problem with home connections is therefore eliminated because you only have to upload a text with the function to the droplets and then you get the estimated coefficients without moving gigabytes of data to your laptop. On the other, and the computation is run in parallel so I could create more droplets to reduce the reading and fitting times even more.

Using the results from the cloud

If you want to average the estimated coefficients, you need to subset and discard the unusable months.

```
intercept <- c()
slope <- c()
for (i in seq_along(fitted_models)) {
  intercept[i] <- ifelse(is.numeric(fitted_models[[i]]),
                         fitted_models[[i]][1], NA)
  slope[i] <- ifelse(is.numeric(fitted_models[[i]]),
                      fitted_models[[i]][2], NA)
}
avg_coefficients <- c(mean(intercept, na.rm = T), mean(slope, na.rm = T))
names(avg_coefficients) <- c("intercept", "slope")
avg_coefficients
# intercept      slope
# 4.406586   2.533299
```

Now you can say that, on average, if you travelled 5 miles by taxi, you'd expect to pay $4.4 + 2.5 \times 5 \approx 17$ USD, which is just an estimate depending on traffic conditions, finding many red lights, etc.

Don't forget to delete the droplets when you are done using them.

```
droplet_delete(droplet1)
droplet_delete(droplet2)
```

Final remarks on cloud estimation

Arrow and S3 provide a powerful combination for big data analysis. You can even create a droplet to control other droplets if you fear that your internet connection is not that good like to transfer complete regression summaries.

Think about the possibility of fitting many other models: generalized linear models, random forest, non-linear regression, posterior bootstrap, or basically any output that R understands.

IBM Cloud Functions com R

Thiago Pires (IBM)

Introdução

A computação sem servidor é um recurso útil para executar na nuvem. Ele combina economia financeira, tempo reduzido de gerenciamento, facilidade na configuração e implementação. Isso significa que um cientista de dados pode trabalhar com mais rapidez ao construir um pipeline de aprendizado de máquina e ainda fornecer um serviço de previsão. A plataforma Function-as-a-Service (FaaS) na IBM Cloud é um recurso para executar código sob demanda com solicitações de API baseadas em HTTP. O IBM Cloud Functions é baseado no projeto de código aberto Apache OpenWhisk.

Hoje é possível com qualquer linguagem de programação criar IBM Cloud Functions. Para algumas linguagens, como Python, Node, Go e outras, a plataforma oferece um tempo de execução padrão e a criação da função sem servidor com menos etapas. Muitos cientistas de dados acreditam que não é possível desenvolver ou pode ser bastante complexo para linguagens de programação sem esse suporte na IBM Cloud. Este artigo mostrará como é possível criar uma função IBM Cloud com a linguagem R.

Predição usando Tidymodels

Foi ajustado um modelo com o banco de dados do Titanic usando a biblioteca tidymodels para, mais tarde, ser capaz de implementar o modelo no IBM Cloud Function.

Uma regressão logística foi utilizada para classificar os passageiros em sobreviveram ou não sobreviveram ao desastre do Titanic. As variáveis selecionadas foram Sexo e Pclass. Os detalhes sobre a preparação de dados para obter train_data são omitidos neste artigo.

```
library(tidymodels)
library(magrittr)
lr_mod <- logistic_reg() %>% set_engine("glm")
lr_fit <- lr_mod %>% fit(Survived ~ Sex + Pclass, data = train_data)
```

Após o ajuste, foi salvo o modelo como arquivo yaml no diretório local. Aqui é necessário usar o pacote yaml e o tidypredict para parse em um arquivo yaml.

```
yaml::write_yaml(tidypredict::parse_model(lr_fit), "R/my_model.yml")
```

Configuração e Deploy

Para executar uma função com uma linguagem que não é suportada pelo IBM Cloud Functions, você precisará configurar um arquivo exec. Na nuvem, a função será executada em um contêiner em Docker cuja imagem é openwhisk/dockerskeleton. O apk é um gerenciamento de pacote Alpine Linux, então você pode adicionar dependências do Linux em seu contêiner. Neste projeto, além das dependências R e do sistema, também é necessário um pacote para manipular com json (por exemplo, jsonlite). Porque, na função, tanto a entrada quanto a saída devem estar neste formato. A estrutura principal do exec é: instalar as dependências do sistema e pacotes R, obter entrada e salvar como arquivo json e, finalmente, executar um script R como executável.

Precisa configurar um arquivo exec como esse:

```
#!/bin/bash#
run R script
chmod +x script.R # turn executable
```

```
echo "$@" > input.json # set input
./script.R # run script
```

Um Dockerfile assim:

```
FROM openwhisk/dockerskeleton
RUN apk update && apk add R R-dev R-doc build-base
RUN R -e "install.packages(c('jsonlite', 'tidypredict', 'yaml'),
repos = 'http://cran.rstudio.com/')"
```

Por último é necessário um script.R para carregar o modelo e calcular as previsões usando tidymodels:

```
#!/usr/bin/env Rscript

# carregar modelo
loaded_model <- tidypredict::as_parsed_model(yaml::read_yaml("my_model.yml"))
# input
input <- jsonlite::fromJSON("input.json", flatten = FALSE)
# calcular predição
pred <- tidypredict::tidypredict_to_column(as.data.frame(input), loaded_model)
# output
jsonlite::stream_out(pred, verbose = FALSE)
```

Para deploy do modelo são necessários os seguintes passos:

- build e push para o Docker Hub.

```
docker build th1460/titanic .
docker push th1460/titanic
```

- logar na IBM Cloud e empacotar os arquivos exec, script.R e my_model.yml.

```
ibmcloud login
ibmcloud target --cf
zip -r titanic.zip exec script.R my_model.yml
```

- Crie a função declarando --web true para transformá-la em uma API.

```
ibmcloud fn action create titanic titanic.zip --docker th1460/titanic --web true
```

Para solicitar uma previsão de API, você pode fazer um POST usando curl ou a função do pacote httr. O e podem ser encontrados com ibmcloud fn action get --url

```
input <- list(Sex = "male", Pclass = "3rd")
"https://<APIHOST>/api/v1/web/<NAMESPACE>/default/titanic.json" %>%
httr::POST(., body = input, encode = "json") %>% httr::content() %>%
.[c("Sex", "Pclass", "fit")] %>% jsonlite::toJSON(pretty = TRUE, auto_unbox = TRUE)
```

Resultados

Após a requisição, a saída mostra os parâmetros (Sex e Pclass) e a probabilidade de sobreviver no desastre do Titanic (fit). Neste exemplo, a solicitação leva 964 ms para obter os resultados.

```
{"Sex": "male", "Pclass": "3rd", "fit": 0.0979}
```

Conclusão

A IBM Cloud Function é um recurso incrível para qualquer linguagem de programação. É fácil de configurar, implantar e escalar. Se este recurso corresponder às necessidades do seu projeto (disponibilidade, frequência de solicitações, etc). Deve ser uma escolha interessante executar uma função sem servidor com R em produção.

Para instalação e configuração pode seguir estes passos em <https://cloud.ibm.com/functions/>. A IBM Cloud fornece a possibilidade de utilizar 5.000.000 execuções no mês com uma conta lite como mostrado em <https://cloud.ibm.com/functions/learn/pricing>.

Referência

- [1] T. Pires. *_IBM Cloud Functions com R_*. Ed. by medium.com. [Online; posted 09-June-2021]. jun. 2021. <URL:
<https://medium.com/ibm-data-ai/ibm-cloud-functions-with-r-e92deec355e2/>>.

Más velocidad y menos colapsos: preprocesamiento de archivos con utilidades del sistema operativo

La creciente disponibilidad de datos masivos fácilmente puede sobrepasar nuestra paciencia, o la capacidad de procesamiento de nuestros computadores. Sin embargo, es posible reducir los tiempos de ejecución y el uso de memoria en la etapa inicial de un análisis, filtrando y procesando archivos de texto delimitado antes de importarlos a R, usando herramientas especializadas - simples pero de alta eficiencia.

Desde la versión 1.2.0 de R, es posible ejecutar comandos del sistema operativo desde R, tendiendo disponibles los resultados de salida de estos comandos para leerlos e incorporarlos al flujo de trabajo. Una de estas conexiones, creada con la función nativa 'pipe', permite ejecutar utilidades de terminal dentro de otras funciones y acceder a sus resultados. Para trabajar con archivos de texto delimitado, 'awk' es una utilidad de consola y un lenguaje de análisis semántico. Aunque data de los años setenta, hoy en día sigue siendo una opción rápida y eficiente para trabajar con archivos de texto de gran tamaño.

Este trabajo muestra como podemos especificar condiciones para descartar registros, de manera similar a la que muchos ya conocemos gracias al paquete 'dplyr'. Por ejemplo: filtrar valores en una columna mayores o menores a algún valor particular, descartar valores NA, o seleccionar columnas de interés usando awk durante el paso de lectura de archivos con las funciones `read.csv` de R Base o `read_csv` del paquete readr. Se presentan ejemplos, con datos espaciales de biodiversidad y con estadísticas deportivas, iterando múltiples archivos, y en archivos individuales con millones de registros. El tiempo para filtrar e importar archivos tiende a reducirse a la mitad comparado con el mismo filtrado de registros dentro de R, con una importante mejoría en el uso de memoria de hasta 700%, que en computadores portátiles actuales puede ser la diferencia entre perder o no la sesión ante un colapso del sistema por falta de memoria.

Uso de R como *front-end* en un datawarehouse de gestión académica universitaria

Daniel Alessandrini, Pablo Martínez, Óscar Montañés, Juan Manuel Serralta

Palabras clave: datawarehouse, enseñanza universitaria, visualización, puntos críticos

Introducción

Un *datawarehouse* es una base de datos que conjuga información extraída de diversas fuentes, las cuales son integradas y transformadas en nuevas estructuras que se adaptan de la mejor manera a las tareas analíticas en una organización, mejorando aspectos que los datos originales tienen, como p.ej. estar optimizados para su almacenamiento pero no para realizar informes.

En el ámbito de la enseñanza universitaria, durante los últimos años han ganado notoriedad este tipo de iniciativas, dada la gran cantidad de información disponible pero, al mismo tiempo, la diversidad de fuentes y la calidad de la información almacenada suele tener serias limitantes para una masiva implementación.

En la gestión académica los denominados “puntos críticos” o “asignaturas de alta complejidad” están cobrando mayor importancia a medida que se obtiene más información de los procesos educativos. Estas son unidades curriculares (UC) a lo largo de las carreras donde los estudiantes tienen problemas para seguir avanzando, donde aumentan los niveles de reprobación y se generan “cuellos de botella” que retrasan buena parte de la escolaridad. Esto puede deberse a distintos factores, desde la complejidad de una UC dado el bagaje que tienen los alumnos que la hacen hasta ese momento, hasta temas de diseño curricular que muchas veces pasan totalmente desapercibidos y que, con información precisa, son mejor detectados.

Forma de trabajo

Como parte de una estrategia sinérgica en la FIIng, se tomó este problema operativo como tema central de una propuesta de proyecto final de carrera de Ingeniería en Computación. De esa forma, el equipo de estudiantes guiado por su tutor, tomó como cliente para su proyecto a la Unidad de Enseñanza, aportando cada uno saberes y experiencias puntuales (el desarrollo de las herramientas de principio a fin por los estudiantes, dadas las correctas definiciones, informes existentes y comentarios de UEFI), generando una propuesta flexible y potencialmente extensible a otras fuentes de datos actuales o futuras.

R es una pieza clave en este engranaje: a través de diferentes aplicaciones Shiny, se generan tanto interacciones entre diferentes fuentes de datos así como también informes dinámicos sobre indicadores de egreso y desempeño, particularmente útiles para comprender mejor los complejos fenómenos como el de los puntos críticos.

Resultados

Durante la presentación se mostrarán los distintos reportes generados y se hará énfasis en la utilidad de esta aplicación para ayudar a considerar a una actividad académica como “de alta complejidad”. Se busca desde este tipo de trabajos aportar luz a estos fenómenos que impactan fuertemente en la actividad académica, además de generar sinergia a través de la suma de los distintos saberes en juego.

Sesión Desarrollo de paquetes y modelos

missMSPC: un paquete de herramientas gráficas para aplicar MSPC con datos faltantes

Julia Inés Fernández , Diego Marfetán Molina , José Alberto Pagura , Marta Beatriz Quaglino

Palabras clave: Control de Calidad - MSPC - Datos Faltantes - Componentes Principales

El control estadístico de procesos es una estrategia ampliamente utilizada en el contexto de control y mejora de la calidad. De acuerdo a la International Organization for Standardization (ISO), calidad es “el grado en el que un conjunto de características inherentes a un objeto (producto, servicio, proceso, persona, organización, sistema o recurso) cumple con los requisitos” (ISO 2005). La calidad constituye un factor decisivo al momento de seleccionar productos o servicios, tanto para transacciones entre organizaciones como en ventas al consumidor final. Actualmente es de suma importancia implementar técnicas de control de calidad, ya que proporcionan ventajas económicas y competitivas.

Generar productos o servicios de alta calidad no es un asunto trivial o sencillo. La constante evolución tecnológica que está transformando industrias y organizaciones, en áreas tan diversas como electrónica, metalúrgica, química, biotecnología, etc., hace posible registrar cada vez más información sobre un proceso, incluso de manera automática, y permite tener una visión más completa del mismo. Una de las consecuencias de estos avances es que la calidad no depende de un único atributo, sino que se mide a través de un **conjunto de características**.

Una de las posibles estrategias a aplicar dentro de este contexto es el Control Estadístico Multivariado de Procesos (MSPC). Entre otras cosas, el MSPC permite monitorear un proceso a lo largo del tiempo, observar si su comportamiento se ajusta al patrón esperado, y detectar eventos especiales en tiempo real que señalen que el proceso se encuentra fuera de control. La principal herramienta del MSPC son los **gráficos de control**, los cuales representan una estadística a lo largo del tiempo y la contrastan con valores límites determinados en función de su distribución de probabilidad, lo que permite determinar el estado del proceso. Existen dos etapas en la utilización de gráficos de control en MSPC, llamadas Fase I y Fase II. En la Fase I se definen los límites de control, los cuales serán utilizados para monitorear el proceso en tiempo real durante la Fase II.

La posibilidad de recolectar datos a gran escala sobre estos procesos se traduce en una baja relación de señal-ruido (*low signal-to-noise ratio*), fuertes estructuras de correlación entre las variables (multicolinealidad) y matrices de covariancia no invertibles, las cuales imposibilitan el cálculo de las estadísticas necesarias para construir gráficos de control multivariados convencionales. En estos casos se aplican métodos basados en **variables latentes**, como Análisis de Componentes Principales (PCA) y Mínimos Cuadrados Parciales (PLS). En primer lugar se proyectan los datos hacia un subespacio de menor dimensión, para luego realizar el control del proceso sobre las variables latentes identificadas como relevantes. En este trabajo se considera el uso simultáneo de los gráficos T^2 de Hotelling y SPE (*squared prediction error*, también llamada estadística Q) sobre variables latentes para el monitoreo del proceso.

Un problema común frente a grandes bases de datos es la ocurrencia de **valores faltantes** en las observaciones a controlar, los cuales pueden surgir debido a fallas en sensores, formularios incompletos, errores de registro, etc. Se han propuesto numerosas estrategias para imputar estos valores cuando el control se realiza aplicando modelos de variables latentes. Dos de las técnicas más populares son *Known Data Regression* (KDR) y *Trimmed Score Regression* (TSR) (Arteaga and Ferrer 2002), (Folch-Fortuny, Arteaga, and Ferrer 2015). El método KDR es equivalente a *Conditional Mean Replacement* (CMR) (Nelson, Taylor, and MacGregor 1996). Hasta donde los autores de este trabajo pudieron constatar, ninguna de estas técnicas se encuentra implementada actualmente en lenguaje R.

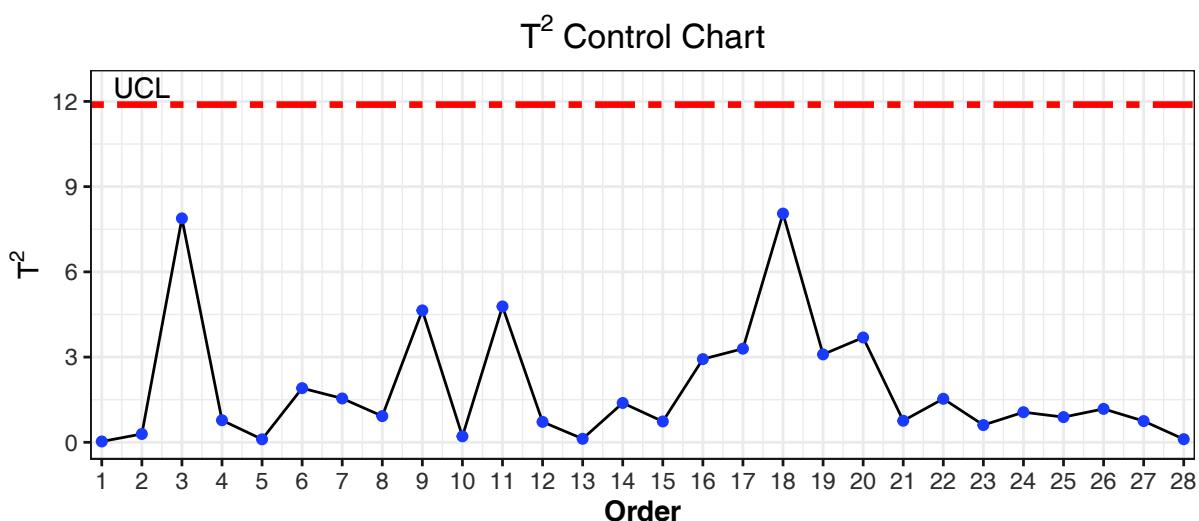
El paquete propuesto missMSPC permite construir los gráficos de control multivariados T^2 y SPE a partir de un modelo PCA en Fase II en ambas situaciones: datos completos o incompletos. Ante la presencia de valores faltantes, se permite realizar la imputación a través de alguno de los métodos propuestos (KDR/CMR o bien TSR) mediante la función `score_imp()`. Consideramos que la importancia de este paquete reside en el hecho de permitir que cualquier interesado/a en realizar control estadístico de procesos tenga acceso a herramientas

avanzadas mediante un software libre. A futuro se planifica implementar el método de control en base a PLS, la estimación del modelo PCA en Fase I cuando se trabaja con datos incompletos, y la utilización de técnicas de validación cruzada para elegir el número óptimo de componentes a incorporar en el modelo PCA. Además, se pretenden incorporar nuevos gráficos de control de mejor performance en presencia de datos faltantes.

A continuación se presenta un ejemplo de aplicación, donde se utilizan los conjuntos de datos *bimetal1* y *bimetal2* del paquete MSQC (Santos-Fernández 2013). El modelo PCA se ajusta sobre los datos *bimetal1*, y luego se aplica el método CMR a una versión de *bimetal2* en la que se generaron aleatoriamente un 25 % de valores faltantes. Finalmente, se construye el gráfico T^2 de Hotelling:

```
bm1 <- scale(bimetal1) #data(bimetal1) del paquete MSQC
bm2 <- scale(bimetal2) #data(bimetal1) del paquete MSQC
set.seed(1974) #Semilla
datos <- mice::ampmpute(bm2, mech = "MCAR", prop = 0.25)$amp #Agrego valores faltantes
acp <- princomp(bm1) #Ajuste modelo CPA
autov <- acp$sdev^2 #Vector de autovalores
pesos <- matrix(as.numeric(acp$loadings), ncol = length(autov)) #Matriz de Pesos

#Imputación de Scores y Gráfico T2
ajuste <- score_imp(datos = datos, pesos = pesos, autov = autov, A = 2, metodo = "CMR")
graficoT2(x = ajuste, alfa = 0.01)
```



Referencias

- 10 Arteaga, F., and A. Ferrer. 2002. "Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples." *Journal of Chemometrics* 16 (810): 408–18. <https://doi.org/10.1002/cem.750>.
- Folch-Fortuny, A., F. Arteaga, and A. Ferrer. 2015. "PCA Model Building with Missing Data: New Proposals and a Comparative Study." *Chemometrics and Intelligent Laboratory Systems* 146: 77–88. <https://doi.org/https://doi.org/10.1016/j.chemolab.2015.05.006>.
- ISO. 2005. "ISO 9000:2005, Quality Management Systems - Fundamentals and Vocabulary." International Organization for Standardization.
- Nelson, Philip R. C., Paul A. Taylor, and John F. MacGregor. 1996. "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations." *Chemometrics and Intelligent Laboratory Systems* 35 (1): 45–65. [https://doi.org/https://doi.org/10.1016/S0169-7439\(96\)00007-X](https://doi.org/https://doi.org/10.1016/S0169-7439(96)00007-X).
- Santos-Fernández, Edgar. 2013. *Multivariate Statistical Quality Control Using r*. Vol. 14. Springer. <http://www.springer.com/statistics/computational+statistics/book/978-1-4614-5452-6>.

QR: Un paquete para la factorización QR sin rotación.

Juan Claramunt González

Abstract El paquete QR contiene la función QR(). Esta función permite realizar una factorización QR de cualquier matriz real sin rotación en las columnas. Esto se contrapone a la función qr() disponible en base R, que realiza la factorización QR utilizando rotación de columnas. De este modo, la función QR se asegura que el producto de las matrices Q y R da como resultado la matriz factorizada, a diferencia de la función qr().

Palabras clave: QR - factorización - sin rotación

Traduciendo un método de Matlab a R nos dimos cuenta de que la función qr() de base R no retornaba el resultado esperado. El producto de las matrices Q y R obtenidas con qr() no era igual a la matriz inicial. El motivo es que qr() llama a la rutinas DQRDC(2) de LINPACK y DGEQP3 o ZGEQP3 de LAPACK. Estas rutinas llevan a cabo la descomposición QR de una matriz utilizando rotación de columnas y esto da lugar al problema. Afortunadamente, en LAPACK también encontramos rutinas que llevan a cabo la factorización QR sin rotación. En nuestro paquete hemos incluido la función QR() que llama a la rutina DGEQRF, que a su vez factoriza QR sin rotación.

Veamos a continuación como utilizar el nuevo paquete y su comparación con las funciones qr(), qr.Q() y qr.R() de base R.

Primero, definamos una matriz aleatoria para factorizar. El paquete QR acepta matrices reales de cualquier tamaño. Una vez definida la matriz, ya podemos utilizar la función QR().

```
#Definimos la matriz
A<-matrix(runif(25,min = -100, max = 100), 5, 5)

#Aplicamos la funcion QR y observamos el resultado
QRres<-QR(A)
QRres$Q
## [1]      [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] -0.42309448  0.234923652  0.6929345 -0.08446852  0.52773935
## [2,] -0.05026825 -0.784271484  0.3661381 -0.43580225 -0.24168291
## [3,] -0.29397211 -0.381323169  0.1177431  0.86461649 -0.08214574
## [4,] -0.84749158  0.001992615 -0.4634462 -0.23454854 -0.10935555
## [5,]  0.11748442 -0.429322475 -0.3964074 -0.01915120  0.80272909
QRres$R
##      [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] -106.2881 -78.51503 -44.49275  12.876767  57.37270
## [2,]    0.0000 112.95345 -44.35744   3.440147 -32.31233
## [3,]    0.0000  0.00000 88.80717 100.827403 29.61217
## [4,]    0.0000  0.00000  0.00000 88.449295 -49.37597
## [5,]    0.0000  0.00000  0.00000  0.000000 69.72084
```

La nueva función QR realiza la factorización QR y también retorna Q y R a diferencia de las funciones de base R, en las que es necesario realizar primero la factorización y después reconstruir las matrices Q y R.

Veamos ahora como lo haríamos usando base R y comparemos los resultados:

```
qrres<-qr(A)
qrQ<-qr.Q(qrres)
qrR<-qr.R(qrres)
#Comprobamos si las matrices son iguales
all.equal(qrQ,QRres$Q)
## [1] TRUE
all.equal(qrR,QRres$R)
## [1] TRUE
```

En este caso vemos que los resultados son iguales, pues la función qr() no ha requerido la rotación de las columnas, sin embargo, este no es siempre el caso.

En el siguiente ejemplo definimos una matriz para que qr() use rotación. En este caso, si comparamos los métodos, podemos ver que los resultados no son iguales, y sólo la factorización obtenida usando QR() da como resultado la matriz original, X, cuando se realiza el producto Q*R.

```
X <- cbind(1, rep(1:0, each = 3), rep(0:1, each = 3),
            rep(c(1,0,0), 2), rep(c(0,1,0), 2), rep(c(0,0,1), 2))
QRresX<-QR(X)
qrresX<-qr(X)
qrX_Q<-qr.Q(qrresX)
qrX_R<-qr.R(qrresX)

#Comprobamos si las matrices son iguales
all.equal(qrX_Q,QRresX$Q)
## [1] "Mean relative difference: 0.7358707"
all.equal(qrX_R,QRresX$R)
## [1] "Mean relative difference: 1.131095"

#Q obtenida con QR()
QRresX$Q
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] -0.4082483 -0.4082483 7.925939e-01 -0.04580286 -0.1308926  0.1386750
## [2,] -0.4082483 -0.4082483 -5.661385e-01 -0.13740858 -0.3926777  0.4160251
## [3,] -0.4082483 -0.4082483 -2.264554e-01  0.18321144  0.5235703 -0.5547002
## [4,] -0.4082483  0.4082483  1.110223e-16 -0.79391626  0.1308926 -0.1386750
## [5,] -0.4082483  0.4082483  2.498002e-16  0.39695813 -0.5796671 -0.4160251
## [6,] -0.4082483  0.4082483  2.775558e-16  0.39695813  0.4487746  0.5547002

#Q obtenida con qr()
qrX_Q
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] -0.4082483 -0.4082483 0.5773503 -2.775558e-17  0.5745653  0.05663922
## [2,] -0.4082483 -0.4082483 -0.2886751 -5.000000e-01 -0.3363337  0.46926857
## [3,] -0.4082483 -0.4082483 -0.2886751  5.000000e-01 -0.2382317 -0.52590779
## [4,] -0.4082483  0.4082483  0.5773503  1.029760e-16 -0.5745653 -0.05663922
## [5,] -0.4082483  0.4082483 -0.2886751 -5.000000e-01  0.3363337 -0.46926857
## [6,] -0.4082483  0.4082483 -0.2886751  5.000000e-01  0.2382317  0.52590779

#Comprobamos si el producto de Q y R es igual a la matriz original, X.
all.equal(X,QRresX$Q%*%QRresX$R)
## [1] TRUE
all.equal(X,qrX_Q%*%qrX_R)
## [1] "Mean relative difference: 1"
```

Por lo tanto, únicamente QR() es capaz de llevar a cabo la factorización QR sin necesidad de rotación. De esta manera, las matrices Q y R resultantes siempre tendrán como producto la matriz original a factorizar.

El proyecto, a parte de ser interesante por tratar de solucionar el problema inicial, incluye el hecho de extender R usando una librería externa de Fortran sin necesidad de tener conocimientos en esta lengua. Esto se realiza gracias a la función .C() que nos permite utilizar código compilado de C. En este código de C hacemos una llamada a la rutina deseada de Fortran usando la función F77_CALL disponible en C. En este punto, únicamente debemos prestar atención a los argumentos de entrada y salida. Para ello, es de gran ayuda netlib.org, que contiene una guía con todas las rutinas disponibles en LAPACK así como sus archivos de ayuda (E. Anderson and Sorensen 1999). En dichos archivos podemos observar los argumentos de entrada y salida, sus tipos de datos y como deben ser utilizados.

Extender R reutilizando código de otras lenguas nos permite ampliar las capacidades de nuestra lengua de un modo óptimo y con menor esfuerzo.

El paquete está disponible en GitHub ([jclarumunt/QR](https://github.com/jclarumunt/QR)) y próximamente lo estará en CRAN.

10 E. Anderson, C. Bischof, Z. Bai, and D. Sorensen. 1999. *LAPACK Users' Guide*. Third Edition. SIAM. http://www.netlib.org/lapack/lug/lapack_lug.html.

Estimación de un modelo computacional mediante computación Bayesiana Aproximada

Juan Ignacio Baccino Costa , Mauro Loprete , Alvaro Valiño , Daniel Ciganda

Abstract El primer objetivo de este trabajo es estimar los parámetros de un modelo computacional del comportamiento reproductivo, en ausencia de una expresión analítica para la función de verosimilitud. Para esto, se utilizan técnicas de Computación Bayesiana Aproximada (ABC) y se analiza la incertidumbre asociada a las predicciones del modelo. Se trabaja con el modelo Comfert, un modelo de microsimulación que modela las trayectorias reproductivas de una cohorte de mujeres en un régimen de fecundidad natural, es decir, en ausencia de intentos dirigidos a prevenir nacimientos. Con estas trayectorias simuladas se obtienen las tasas específicas de fecundidad por edad para dicha cohorte y se utiliza esta información para ajustar el modelo a las tasas observadas en una población histórica. Los datos utilizados provienen de una cohorte de Hutteritas, una comunidad anabaptista frecuentemente estudiada en demografía por su rechazo del uso de métodos anticonceptivos y el nivel elevado de su fecundidad. Por otro lado, como objetivo secundario, se plantea el estudio sobre la incertidumbre en la predicción del modelo anteriormente descrito. Para ello, se realizó primero una estimación puntual considerada óptima (aquella que minimiza el Error Cuadrático Medio) y luego se construyeron diferentes intervalos de credibilidad. Si bien todos se encuentran construidos al 95 % de credibilidad, la incertidumbre varía notablemente en función de los niveles de tolerancia considerados en el algoritmo. Los resultados obtenidos ilustran la utilidad del enfoque bayesiano en la realización de inferencia estadística sobre modelos computacionales, al igual que una relación positiva entre el nivel de incertidumbre y el incremento en la tolerancia del algoritmo.

Palabras clave: Modelos Computacionales - Computación Bayesiana Aproximada - Fecundidad - Error Cuadrático Medio - Intervalos de credibilidad

Introducción

El presente trabajo tiene como principal objetivo utilizar el enfoque ABC para estimar un modelo demográfico de microsimulación para el que no es posible derivar una expresión de la función de verosimilitud, pero desde el que es posible simular datos que pueden compararse con los datos provenientes de una población histórica. Por otro lado, como objetivo secundario, se plantea el estudio sobre la incertidumbre en la predicción del modelo anteriormente mencionado.

El modelo con el que se trabajó, *Comfert* (Ciganda and Todd 2019), simula las trayectorias reproductivas de una cohorte de mujeres en un régimen de fecundidad natural (población en la que no existe control explícito de la natalidad).

Para ajustar el modelo se utilizan una serie de tasas específicas de fecundidad por edad de una cohorte de mujeres de la población Hutterita, una comunidad religiosa Anabaptista con un estilo de vida tradicional que incluye el rechazo al uso de métodos anticonceptivos.

Se entiende como proceso reproductivo a la secuencia de nacimientos de una mujer y edades de la madre al nacimiento.

A partir de esta secuencia es posible calcular los indicadores del nivel de la fecundidad más frecuentemente utilizados, como las tasas específicas de fecundidad por edad, definidas como:

$$nF_x [0, T] = \frac{nB_x [0, T]}{nL_x [0, T]}$$

siendo el numerador la cantidad de nacimientos de mujeres de edades entre x y $x + n$ sobre la población promedio (en T años) de este grupo de mujeres. Como se puede ver, esta tasa mide la cantidad promedio de hijos por mujer a la edad x .

Técnicas utilizadas

- **Modelo de microsimulación**
- **Computación Bayesiana Aproximada ABC**

El modelo de microsimulación, *Comfert* está enteramente programado en R, al ser un modelo de simulación de tiempos discretos, utilizamos la librería *lubridate*, lo cual nos permite trabajar con los tiempos de los eventos principales (Nacimiento, Edad de Unión, Tiempo de no susceptibilidad, etc) de una trayectoria individual, dado el tipo de modelos se necesita una gran eficiencia velocidad en la ejecución de las simulaciones, es por esto, que se hace ayuda de el paquete *data.table*(Dowle and Srinivasan 2021).

En base de las características del modelo construido, sus parámetros son variables aleatorias con diferentes distribuciones que determinan el comportamiento de los individuos y es **imposible** encontrar una expresión analítica de la función de verosimilitud, es que se recurren a técnicas de Computación Bayesiana Aproximada (ABC).

Es por esto, que para implementar el algoritmo, fue necesario utilizar librerías como *mlgcp*(Dancik and Dorman 2008) para poder implementar procesos de simulación gaussiana para explorar el espacio paramétrico de nuestro meta-modelo.

Comenzando con distribuciones a priori de losparámetros a estimar el modelo y luego mediante sucesivas simulaciones del modelo ejecutadas en paralelo, mediante la librería *parallel*(R Core Team 2021) del modelo **Comfert** aproximamos la distribución a posteriori marginal de los mismos que mejor ajustan a los observados.

A continuación, a modo de resumen los principales parámetros del modelo *Comfert* y el rol de los mismos en el algoritmo ABC.

Parámetro	Descripción	Rol ABC	Valor
α	Edad de inflexión del declive de la fecundidad	A estimar	30-40
κ	Tasa de declive de la fecundabilidad	A estimar	0.20-0.4
shape	Heterogeneidad de la Fecundidad (entre mujeres)	Fijo	5
rate	Heterogeneidad de la Fecundidad (entre mujeres)	Fijo	38
μ	Edad media de la unión	A estimar	15-20
σ	Desvío de la edad media de la unión	A estimar	0.1-0.30
nsp	Período de no susceptibilidad	Fijo	6
srb	Ratio por sexo al nacimiento (en términos de hombres)	Fijo	0.515

Cuadro 1: Resumen

Por último, a la hora de estudiar la incertidumbre en la predicción del modelo, se procedió a computar el ECM en cada combinación de α y κ seleccionando como candidatos de valores pertenecientes a la estimación de la distribución a posteriori aquellos que tienen error menor o igual al percentil que acumula un 10 %, 50 %, y 75 % de la probabilidad de la distribución del error respectivamente. A continuación se muestra una visualización de los diferentes valores mediante visualizaciones en *ggplot2*(Wickham 2016)

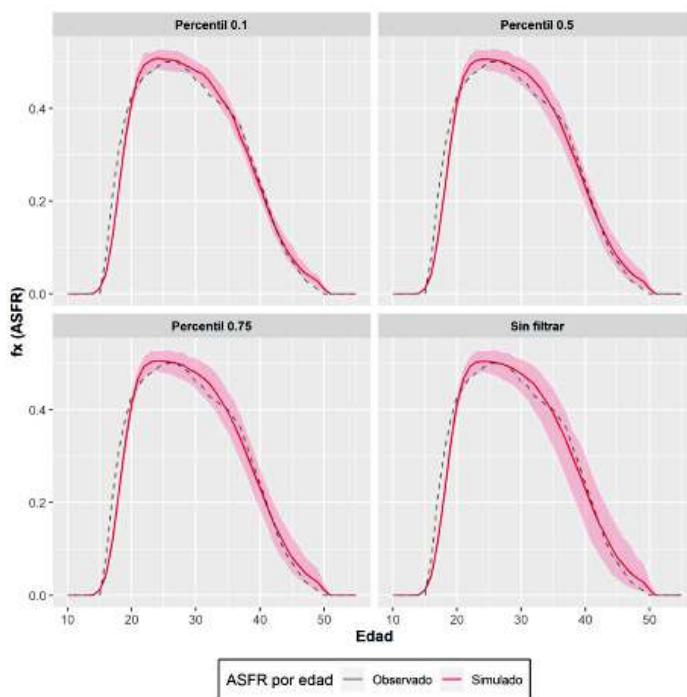


Figura 1: Gráfico de líneas de las ASFR, valores observados, mediana e intervalo de confianza al 95 %. Se observa que la amplitud de los intervalos de confianza aumenta conforme se incrementa el valor de epsilon seleccionado y la edad de la madre.

Referencias

- 10 Carnell, Rob. 2020. *Lhs: Latin Hypercube Samples*. <https://CRAN.R-project.org/package=lhs>.
- Ciganda, Daniel, and Nicolas Todd. 2019. “The limits to fertility recuperation.” MPIDR Working Papers, no. WP-2019-024. <https://doi.org/10.4054/MPIDR-WP-2019-024>.
- Dancik, Garrett M, and Karin S Dorman. 2008. “mlegp: Statistical Analysis for Computer Models of Biological Systems Using r.” *Bioinformatics* 24 (17): 1967.
- Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of ‘Data.frame’*. <https://CRAN.R-project.org/package=data.table>.
- Grolemund, Garrett, and Hadley Wickham. 2011. “Dates and Times Made Easy with lubridate.” *Journal of Statistical Software* 40 (3): 1–25. <https://www.jstatsoft.org/v40/i03/>.
- Gutmann, Michael U., and Jukka Corander. 2015. “Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models.” <http://arxiv.org/abs/1501.03291>.
- Henry, Louis. 1961. “Some Data on Natural Fertility.” *Eugenics Quarterly* 8 (2): 81–91. <https://doi.org/10.1080/19485565.1961.9987465>.
- Kostaki, Anastasia, and Peristera Paraskevi. 2007. “Modeling Fertility in Modern Populations.” *Demographic Research* 16: 141–94. <https://doi.org/10.4054/demres.2007.16.6>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Schmertmann, Carl. 2003. “A System of Model Fertility Schedules with Graphically Intuitive Parameters.” *Demographic Research* 9: 81–110. <https://doi.org/10.4054/demres.2003.9.5>.
- Sheps, Mindel C. 1965. “An Analysis of Reproductive Patterns in an American Isolate.” *Population Studies* 19 (1): 65. <https://doi.org/10.2307/2173165>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.

Juan Ignacio Baccino Costa
Universidad de la República, FCEA, IESTA
nacho.baccino.3323@gmail.com

Mauro Loprete
Universidad de la República, FCEA, IESTA

Alvaro Valiño
Universidad de la República, FCEA, IESTA

Daniel Ciganda
Max Planck Institute for Demographic Research

Feature and variable selection in complex data classification

Manuel Oviedo de la Fuente, University of A Coruña, CITIC (manuel.oviedo@udc.es)
Manuel Febrero Bande University of Santiago de Compostela

Introduction

This study addresses the classification of complex data such as spectrometric curves, hyperspectral images and 3D point cloud. The study also focuses on the procedures for feature and variable selection through the recursive use of distance correlation (DC). For this, the functional data analysis (FDA) framework will be used through the R package fda.usc.

Feature and variable selection in logistic regression

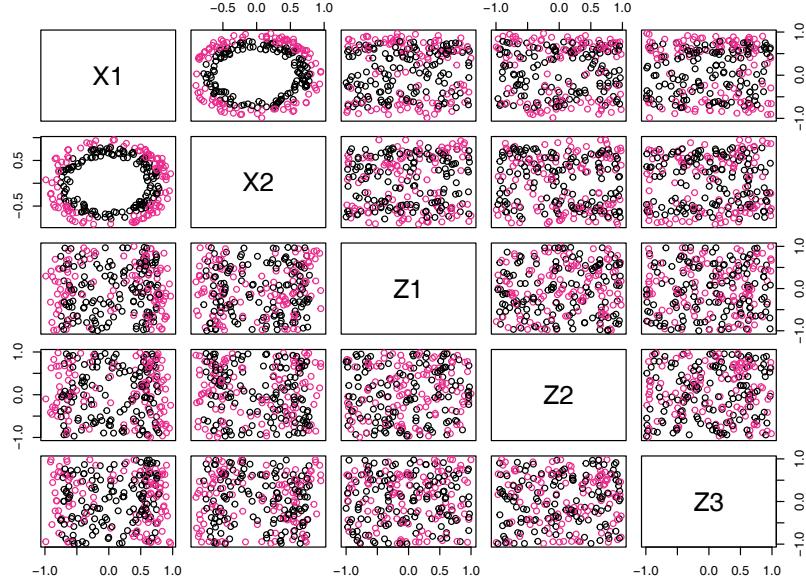
Febrero-Bande et al. (2019) consider the problem of variable selection in regression models in the case of functional variables that may be mixed with other type of variables (scalar, multivariate, directional, etc.). Their proposal begins with a simple null model and sequentially selects a new variable to be incorporated into the model based on the use of DC proposed by (Székely et al., 2007), $Y_i = \alpha + \sum_{j=1}^J f_j(X_i^{(j)}) + \varepsilon_i, \quad i = 1, \dots, N$

We are interested in an automatic regression procedure capable of dealing with a large number of covariates of different nature. We adapt the variable selection procedure proposed by Febrero-Bande et al. (2019) from regression to classification.

Simulation Example

Multi class classification is implemented by training multiple logistic additive regression classifiers (one vs all scheme) using incoming function `classif.gsam.vs()`.

```
Nt=250; Np=100; nB=100;Nvar = 19 ; Xdat = simul2d(Nt,Np,Nvar) # data generation  
Xtrain <- Xdat$Xtrain; Xtest <- Xdat$Xtest; pairs(Xtrain[,2:6],col=Xtrain$grupo)
```



```

gsam <- classif.gsam.vs(lpdata(Xtrain[,1:(Nvar-1)]), "grupo")
gsam$i.predictor # Variable selected 1, otherwise 0

##   X1   X2   Z1   Z2   Z3   Z4   Z5   Z6   Z7   Z8   Z9   Z10  Z11  Z12  Z13  Z14  Z15
##   1    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
pred <- predict.classif(gsam,lpdata(Xtest[,1:(Nvar-1)]))
table(Xtest$grupo,pred) # Confusion matrix

##      pred
##      1 2
##  1 41 1
##  2  0 58
mean(Xtest$grupo==pred) # Accuracy
## [1] 0.99

```

Complex data classification

- **Hyperspectral image classification.** In this second case study, we also review different models classification algorithms for the prediction of the future class of pixel in a hyperspectral image that have in common that make use of FDA.
- **3D point cloud classification.** Finally, we will reproduce the examples of Oviedo-de la Fuente et al. (2021) to select optimum scales in multiscale classification problems with machine learning. A maximum of three scales for each feature was sufficient to obtain the best results in the classification, measured in terms of precision, recall and F1-index.

Funding and Financial Support:

CITIC funding Consellería de Economía, Empleo e Industria and FEDER Galicia 2014-2020) and MODES group by the Xunta de Galicia (ED431C-2020-14 and ED431G 2019/01).

References

- Frbrero-Bande, M., González-Manteiga, W., and Oviedo de la Fuente, M. (2019). Variable selection in functional additive regression models. *Computational Statistics*, 34(2):469–487.
- Oviedo-de la Fuente, M., Cabo, C., Ordóñez, C., and Roca-Pardiñas, J. (2021). A distance correlation approach for optimum multiscale selection in 3d point cloud classification. *Mathematics*, 9(12):1328.
- Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794.

Optimizando @RStatsJobsBot: un modelo de aprendizaje automático para clasificar tweets de ofertas de empleo.

Martin Rodriguez Nuñez^a, Juan Cruz Rodriguez^b

^aCONICET, IMBIV, Universidad Nacional de Córdoba, Argentina

^bFAMAF, Universidad Nacional de Córdoba, Argentina

Abstract

Keywords: Minería de texto, Bot, Twitter, Aprendizaje supervisado

@RStatsJobsBot es el bot de Twitter con más seguidores en el ámbito de la oferta de puestos de trabajo vinculados con R. Su propósito es compartir todos los tweets cuyo objeto sea publicitar vacantes laborales relacionados con el lenguaje de programación R, así facilitando la ardua tarea de buscar un nuevo empleo. Desde su creación, el bot implementó un conjunto de reglas de decisión creadas por su desarrollador para discriminar los tweets que eran verdaderas ofertas de trabajo de las que no lo eran. A pesar de contar con un algoritmo de decisión, el bot cometió muchos errores de clasificación que lo llevaron a publicar contenido erróneo. Desde sus inicios, todos los tweets fueron guardados y curados manualmente por su desarrollador. En la actualidad, existe una base de datos de 6783 tweets de los cuales 249 son verdaderas ofertas de empleo.

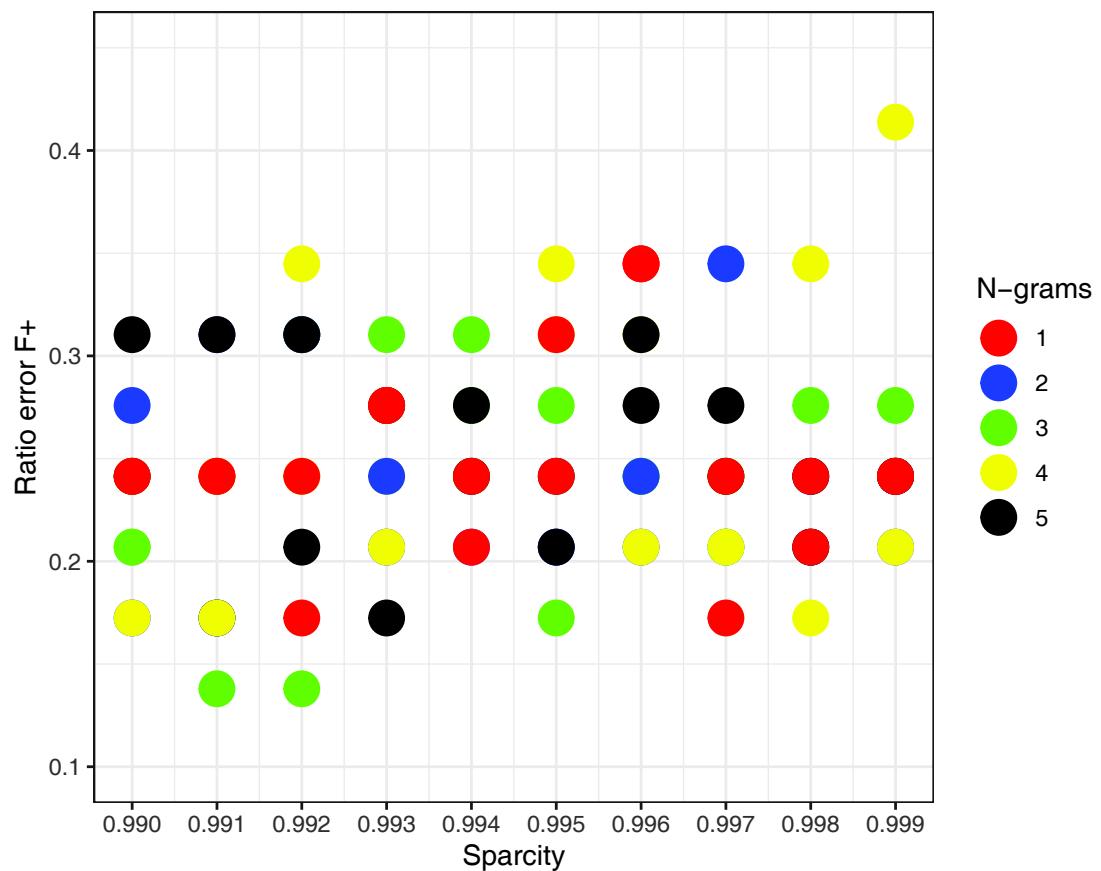
Previo al desarrollo de modelo predictivo, es necesario implementar técnicas de limpieza de texto sobre los diferentes tweets con el objetivo de eliminar toda aquello que aporte ruido y de transformar el texto de forma tal de maximizar el aprovechamiento de la información para el modelo predictivo. Para cumplir este objetivo es necesario convertir los datos a un **Corpus** y aplicar distintas técnicas de limpieza y transformación, las más importantes involucradas fueron: remover símbolos y puntuación, eliminar emails y páginas web añadiendo como variable respuesta el número de cada una de ellas en cada tweet, convertir a minúsculas, quitar stop words y realizar stemming, llevando las palabras a su word stem.

Para desarrollar el modelo predictivo es necesario convertir el texto procesado en una *Document Term Matrix* (DTM). A la hora de llevar a cabo el ajuste del modelo surgen dos incógnitas. La primera de ellas se encuentra referida a cuál es el mejor algoritmo de clasificación que ajusta a los datos y la otra se encuentra relacionada a la determinación de los hiperparámetros que fijan

Email addresses: martinrnu@gmail.com (Martin Rodriguez Nuñez), jcrodriguez@unc.edu.ar (Juan Cruz Rodriguez)

la dimensión de la **DTM**. Estos hiper parámetros son el número máximo de **N-grams** a considerar y la **sparsity** que se tomara. Para determinarlos se ajustaron distintos algoritmos de clasificación a todas las combinaciones posibles de estos hiper parámetros, analizando cuál era el que lograba el mejor ajuste. Para el entrenamiento de los distintos algoritmos se procedió a implementar la herramienta de **AutoML** perteneciente al paquete de **h2o**, tomando como métrica a optimizar el **mean_per_class_error**. Para el procedimiento de ajuste se optó por dividir los datos en un 75/10/15 para entrenamiento, testeo y validación. El ajuste de los diferentes algoritmos predictivos se realizó por medio de **10 fold cross validation** y la selección de los mejores modelos se realizó a través del desempeño en la base de datos de **testeo**. El problema de clasificación presentado tiene una variable respuesta con dos posibles categorías: propuesta de trabajo falsa (false) y propuesta de trabajo verdadera (true). El objetivo principal fue minimizar la **ratio de error para los falsos negativos** (propuestas de trabajo verdaderas clasificadas como falsas), además también se buscó minimizar la **ratio de error para los falsos positivos** (propuestas de trabajo falsas clasificadas como verdaderas). Los resultados arrojados por el procedimiento empleado posicionaron el algoritmo de **Gradient Boosting Machine (GBM)** como el mejor algoritmo de clasificación para esta variable respuesta. La combinación de hiper parámetros que minimizan los *false positives* en el set de datos de testeo fue de considerar hasta un máximo de 3 grams con una sparsity de 0.992 y la que minimizó los *false negatives* fue de 2 grams como máximo con una sparsity de 0.997. Para maximizar la capacidad predictiva se planteó unificar ambos modelos en un **ensemble**, con lo cual la clasificación se realiza en dos pasos, primero se clasifica con el modelo que mejor identifica los *true positives* (menor error en *false negatives*) y luego se filtra con el que mejor identifica *true negatives* (menor error en *false positives*).

Ratio error F+ vs Sparsity para distintos N–grams en los GBM



El objetivo de esta *flash talk* es mostrar el análisis de minería de textos y aprendizaje automático (Silge 2017, Hvitfeldt 2021) que llevó a la definición de un modelo de aprendizaje automático que discrimina qué tweets que son ofertas de trabajo verdaderas y cuáles no.

Referencias

- Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media, Inc.
- Hvitfeldt, E. & Silge, J. (2021) Supervised Machine Learning for Text Analysis in R. CRC press.

Como implementar algunos modelos de imputación múltiple para datos de panel

Anónimo

Abstract Los datos faltantes son todo un reto en los análisis estadísticos porque los resultados que arrojan tienen limitaciones. La imputación, entendida como el proceso de reemplazar los datos faltantes con un valor estimado, es un problema regular en los proyectos de investigación. Existen muchos modelos y paquetes destinada para este proceso, sin embargo, la selección del modelo de imputación adecuado al tipo de datos disponibles es trascendental para la fiabilidad del resultado. En este estudio se trabaja con una tabla de datos cruzada que involucran series de tiempo (datos panel) para 33 países y 17 variables (Índice de Gini anual para período 2000-2016), con un 24% de datos faltantes. Con el objetivo de imputar estos datos, se utilizó un modelo de imputación múltiple propuesto por Honaker y King (2010) y se agregaron algunas restricciones al sistema. Los principales resultados obtenidos, conducen a la siguiente interrogante Se puede confiar en la imputación? Todos los archivos necesarios para reproducir los resultados presentados están disponibles en: <https://gitlab.com/iesta.fcea.udelar/como-implementar-algunos-modelos-de-imputacion-multiple-para-datos-de-panel>.

Palabras clave: Datos faltantes - Datos panel - Imputación

Referencias

- Arellano, Manuél, and Olympia Bover. 1990. "La Econometría de Datos Panel." *Investigaciones Económicas* XIV (1): 3–45. <https://www.fundacionsepi.es/investigacion/revistas/paperArchive/Ene1990/v14i1a1.pdf>.
- Bell, Melanie L, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. 2014. "Handling Missing Data in RCTs; a Review of the Top Medical Journals," 8.
- Honaker, H., G. King, and M. Blackwell. 2018. "AMELIA II."
- Honaker, James, and Gary King. 2010. "What to Do About Missing Values in TimeSeries CrossSection Data." *America Journal of Political Science* 54 (2): 561–81. <https://doi.org/10.1111/j.1540-5907.2010.00447.x>.
- Kossen, Tabea, Michelle Livne, Vince I Madai, Ivana Galinovic, Dietmar Frey, and Jochen B Fiebach. 2019. "A Framework for Testing Different Imputation Methods for Tabular Datasets." *bioRxiv*, January, 773762. <https://doi.org/10.1101/773762>.
- Leite, W., and N. Beretvas. 2010. "The Performance of Multiple Imputation for Likert-Type Items with Missing Data." *Journal of Modern Applied Statistical Methods* 9: 64–74. <https://doi.org/10.22237/jmasm/127268620>.
- Madley-Dowd, P., R. Hughes, K. Tilling, and J. Heron. 2019. "The Proportion of Missing Data Should Not Be Used to Guide Decisions on Multiple Imputation." *Journal of Clinical Epidemiology* 110: 63–73. <https://doi.org/10.1016/j.jclinepi.2019.02.016>.
- Medina, F., and M. Galván. 2007. "Imputación de Datos: Teoría Y Práctica." Publicación de las Naciones Unidas: CEPAL. https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf.
- Muñoz-Rosas, Juan Francisco, and Encarnación Álvarez-Verdejo. 2009. "Métodos de Imputación Para El Tratamiento de Datos Faltantes: Aplicación Mediante R/Splus." *Revista de Métodos Cuantitativos Para La Economía Y La Empresa*, 3–30. <http://www.upo.es/RevMetCuant/art25.pdf>.
- Murray, Jared S. 2018. "Multiple Imputation: A Review of Practical and Theoretical Findings." *Statistical Science* 33 (2): 142–59. <https://doi.org/10.1214/18-STS644>.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. USA: John Wiley & Sons, Inc. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316696.fmatter>.
- Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.
- Takahashi, M. 2017. "Multiple Ratio Imputation by the EMB Algorithm: Theory and Simulation." *Journal of Modern*

Applied Statistical Methods 16 (1): 630–56. <https://doi.org/10.22237/jmasm/1493598840>.

Van Buuren, S., and K Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software, Articles* 45 (3): 1–67. <https://doi.org/10.18637/jss.v045.i03>.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Segunda. Florida, USA: Taylor & Francis. <https://stefvanbuuren.name/fimd/>.

Wood, Angela M, Ian R White, and Simon G Thompson. 2004. "Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals." *Clinical Trials: Journal of the Society for Clinical Trials* 1 (4): 368–76. <https://doi.org/10.1191/1740774504cn032oa>.

Yamaguchi, Yusuke, Mai Ueno, Kazushi Maruo, and Masahiko Gosho. 2020. "Multiple Imputation for Longitudinal Data in the Presence of Heteroscedasticity Between Treatment Groups." *Journal of Biopharmaceutical Statistics* 30 (1): 178–96. <https://doi.org/10.1080/10543406.2019.1632878>.

Zhang, Zhongheng. 2016. "Missing Data Imputation: Focusing on Single Imputation." *Annals of Translational Medicine* 4 (1): 9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38>.

Años potenciales de vida perdidos por siniestros de tránsito en Uruguay

Gonzalo De armas , Mauro Loprete , Ramón Álvarez-Vaz

Abstract Los Años Potenciales de Vida Perdidos (APVP) son un indicador demográfico que permite cuantificar cuantos años, en promedio, se han dejado de vivir por un conjunto de causas de muerte, en particular. Se calculan como la sumaatoria, para todos los fallecimientos, de la diferencia de años entre la edad del fallecido y una cierta edad límite L , que en la presente investigación se especifica como la esperanza de vida a la edad x . Los siniestros de tránsito en el Uruguay son una de las causas externas con mayor cantidad de decesos y una de las primeras en el grupo de edad que comprende los 15 a 45 años. Los resultados obtenidos se visualizan mediante una interfaz gráfica desarrollada mediante el paquete shiny en R. Puede concluirse que las muertes y APVP por siniestros de tránsito tienen un perfil masculino, joven, donde las muertes entre conductores de motocicletas dominan en el aporte al total de APVP. Si se compara entre los distintos escenarios de siniestros, las muertes entre quienes viajaban en una motocicleta tienen una mediana de edad inferior a los 30 años, mientras que las muertes entre los peatones se caracterizan por una mediana superior a los 60 años.

Palabras clave: APVP - Shiny - Siniestros de Tránsito

Introducción

En el contexto de la primera transición demográfica, se ha registrado una disminución sostenida de la mortalidad, que frecuentemente se desagregan en: transmisibles, no transmisibles y externas.

En los países donde la mortalidad, ha llegado a niveles mínimos, las causas externas son una causa de muerte de alta prevalencia en los jóvenes, dentro de las causas externas, se pueden caracterizar como:

- Intencionales
- No intencionales
- Intención indeterminada

Los siniestros de tránsito, por lo tanto, al ser una causa de muerte externa y no intencional, tienen un mayor potencial para disminuir su cuantía en con políticas públicas apropiadas.

Objetivo

Con motivo entonces de poder analizar el problema descrito, se plantea la pregunta: ¿Cuál ha sido la evolución de los años potenciales de vida perdidos por mortalidad asociada a siniestros de tránsito en Uruguay entre 2012-2018?

Para contestar esta pregunta se plantea como objetivo el calcular los APVP generados por siniestros de tránsito en el Uruguay, para el período 2013-2018, desagregando por el rol del fallecido (peatón, conductor o acompañante) y el tipo de vehículo.

En base a este objetivo, se utilizan los datos abiertos proporcionados por UNASEV (Unidad Nacional de Seguridad Vial 2013-2019)^{1}}

Metodología

Años de vida perdidos

Es un indicador utilizado en Epidemiología y la Salud Pública, surgido de la Demografía, es un buen indicador de **mortalidad prematura**. Los *Años Potenciales de Vida Perdidos* (APVP), utilizado por primera vez por Dempsey (Dempsey 1947) y desarrollado por Arriaga (Arriaga 1996), describen la suma algebraica de los años de vida que, potencialmente, hubiesen vivido los individuos que fallecen por una cierta causa considerando una cierta edad límite L de supervivencia.

Siendo L una edad fijada para indicar el límite, donde se establece que edades de muerte son prematarias y a partir de que momento no se considera una perdida potencial.

¹Disponibles en : https://catalogodatos.gub.uy/dataset/unasev-fallecidos_siniestros_tránsito

Sin embargo, por considerarse que los siniestros de tránsito son, para toda edad, una causa de muerte evitable es que se prefiere no limitar las muertes cuantificadas por los APVP aplicando la siguiente fórmula:

$$APVP = \sum_{x=0}^{\omega-1} e_x d_x \quad (1)$$

La ecuación (1) considera la esperanza e_x de vida que corresponde a una persona que llega con vida a la edad x y fallece por la causa de estudio a esa edad, estos serían los años potenciales de vida perdidos efectivamente para esa persona.

Resultados

Los resultados de este trabajo son procesados mediante programación en el lenguaje R (Team 2020), en base a datos abiertos publicados por la [UNASEV](#). Se propone presentar los mismos a través de una aplicación [{Shiny}](#) (Rstudio 2013), en ella, se pueden aplicar diferentes filtros, obtener georreferencias de los siniestros de tránsito mediante mapas interactivos construidos con [{leaflet}](#) (Cheng, Karambelkar, and Xie 2021), una pestaña que muestra la evolución de este indicador en el tiempo y la descomposición de un año en concreto por edad y medio de tránsito, con visualizaciones de [{ggplot2}](#) (Wickham et al. 2020) y temas de [{hrbrthemes}](#) (Rudis 2020), como puede ser la siguiente :

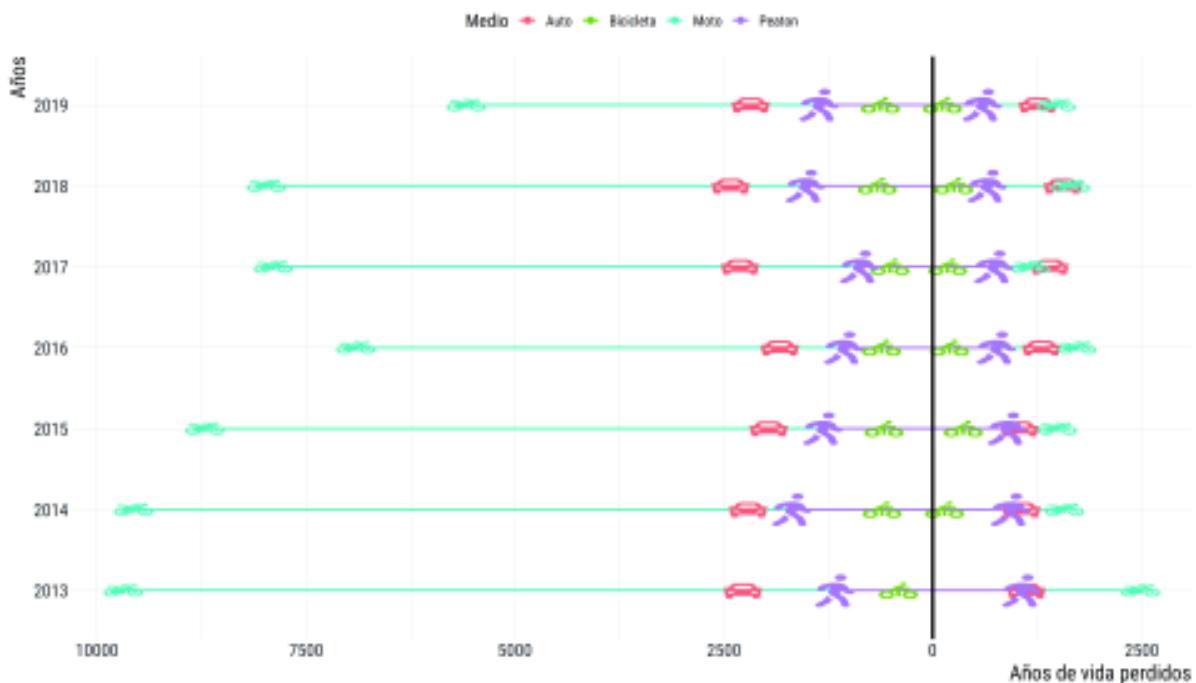


Figura 1: Aquí se puede ver la diferencia por sexo, mayor magnitud en hombres, además de que el medio de tránsito que aporta más a los APVP es del de las motocicletas.

Además, se presentan análisis univariados de los determinantes de los APVP, una sección que muestra la proporción de los accidentes mortales por medio de tránsito, según Departamento y género, otra que muestra la distribución de las edades de los fallecidos en el accidente para un año en específico y separado por medio de tránsito. Todo el código de la aplicación se encuentra en un repositorio de [Gitlab](#) y se implementó el proyecto en plataformas como [Gitpod](#) y [RStudio Cloud](#), además de utilizar contenedores Docker para garantizar la reproducibilidad de la aplicación, que se encuentra disponible [aquí](#)

Referencias

- 10 Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for Ggplot2*. <https://github.com/jrnold/ggthemes>.
- Arriaga, E. 1996. "Los años de Vida Perdidos: Su Utilización Para Medir El Nivel y Cambio de Mortalidad."
- Bache, Stefan Milton, and Hadley Wickham. 2020. *Magrittr: A Forward-Pipe Operator for r*. <https://CRAN.R-project.org/package=magrittr>.
- Baio, Gianluca. 2013. *Bayesian Methods in Health Economics*. Champman; Hall CRC - Biostatistics Series.
- Chang, Winston, and Barbara Borges Ribeiro. 2018. *Shinydashboard: Create Dashboards with Shiny*. <http://rstudio.github.io/shinydashboard/>.
- Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *Shiny: Web Application Framework for r*. <https://shiny.rstudio.com/>.
- Cheng, Joe, Bhaskar Karambelkar, and Yihui Xie. 2021. *Leaflet: Create Interactive Web Maps with the JavaScript Leaflet Library*. <https://rstudio.github.io/leaflet/>.
- Dávila-Cervantes, C., and A. Pardo. 2016. "Análisis de La Tendencia e Impacto de La Mortalidad Por Causas Externas: México, 2000-2013." *Salud Colectiva* 12: 251.
- Dempsey. 1947. "M. Decline in Tuberculosis : The Death Rate Fails to Tell the Entire Storyd."
- Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of 'Data.frame'*. <https://CRAN.R-project.org/package=data.table>.
- Frenk, J. Bobadilla, J. Stern, C. Frejka, T. Lozano, and R. Elements. 1991. "For a Theory of Transition in Health." *Salud pùblica de México* 33: 448-62.
- Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1-25. <https://www.jstatsoft.org/v40/i03/>.
- Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. <https://CRAN.R-project.org/package=purrr>.
- Karambelkar, Bhaskar, and Barret Schloerke. 2018. *Leaflet.extras: Extra Functionality for Leaflet Package*. <https://CRAN.R-project.org/package=leaflet.extras>.
- La, OPS. n.d. "Salud En Las Américas" 1998.
- Murray, C. J. L., A. D. Lopez, W. H. Organization, and W. and Bank. 1996. "Y of Public Health, h." In *Global Burden of Disease : A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020 :summary and Edited by Christopher j*, edited by S. The. alan d. lopez: l. murray.
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- Müller, Kirill, and Hadley Wickham. 2021. *Tibble: Simple Data Frames*. <https://CRAN.R-project.org/package=tibble>.
- Omran, A. The epidemiologic transition. 1971. "A Theory of the Epidemiology of Population Changes." *The Milbank Memorial Fund Quarterly, Vol 49*.
- Ooms, Jeroen. 2021. *Rsvg: Render SVG Images into PDF, PNG, PostScript, or Bitmap Arrays*. <https://github.com/jeroen/rsvg#readme>.
- Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439-46. <https://doi.org/10.32614/RJ-2018-009>.
- . 2021. *Sf: Simple Features for r*. <https://CRAN.R-project.org/package=sf>.
- Perrier, Victor, Fanny Meyer, and David Granjon. 2021. *shinyWidgets: Custom Inputs Widgets for Shiny*. <https://github.com/dreamRs/shinyWidgets>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rinker, Tyler W., and Dason Kurkiewicz. 2018. *pacman: Package Management for R*. Buffalo, New York. <http://github.com/trinker/pacman>.
- Rinker, Tyler, and Dason Kurkiewicz. 2019. *Pacman: Package Management Tool*. <https://github.com/trinker/pacman>.
- Rstudio, Inc. 2013. *Shiny*. Easy web application in R.
- Rudis, Bob. 2020. *Hrbrthemes: Additional Themes, Theme Components and Utilities for Ggplot2*. <http://github.com/hrbrmstr/hrbrthemes>.
- Salomon, J., and C. Murray. 2002. "The Epidemiologic Transition Revisited: Compositional Models for Causes of Death by Age and Sex." *Population and Development Review* 28: 205-28.

- Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2021. *Lubridate: Make Dealing with Dates a Little Easier*. <https://CRAN.R-project.org/package=lubridate>.
- Team, R Core. 2020. “R Foundation for Statistical Computing.”
- Unidad Nacional de Seguridad Vial. 2013-2019. “Fallecidos En Siniestros de Tránsito Por año 2013 - 2019.”
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- . 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. <https://CRAN.R-project.org/package=stringr>.
- . 2021a. *Forcats: Tools for Working with Categorical Variables (Factors)*. <https://CRAN.R-project.org/package=forcats>.
- . 2021b. *Tidyr: Tidy Messy Data*. <https://CRAN.R-project.org/package=tidyr>.
- . 2021c. *Tidyverse: Easily Install and Load the Tidyverse*. <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. <https://CRAN.R-project.org/package=readxl>.
- Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, and Jim Hester. 2020. *Readr: Read Rectangular Text Data*. <https://CRAN.R-project.org/package=readr>.
- Yu, Guangchuang. 2020. *Ggimage: Use Image in Ggplot2*. <https://CRAN.R-project.org/package=ggimage>.

Gonzalo De armas
Universidad de la Repùblica,FCEA,IESTA
gonzalo.dearmas@iest.edu.uy

Mauro Loprete
Universidad de la Repùblica,FCEA,IESTA
mauroloprete1@gmail.com

Ramón Álvarez-Vaz
Universidad de la Repùblica,FCEA,IESTA
ramon@iest.edu.uy

Sesión Desarrollo de paquetes y aplicaciones en ciencia

{agromet}: un paquete para el análisis de datos meteorológicos

Natalia Gattinoni, Paola Corrales, Elio Campitelli, Yanina Bellini y Gabriel Rodriguez

Palabras claves: índices agrometeorológicos - representación geoespacial - seguimiento climático

Uno de los factores que influye sobre la producción agropecuaria es la variabilidad de las condiciones meteorológicas. En este sentido, es primordial hacer un seguimiento y análisis de las distintas variables meteorológicas (lluvias, temperaturas, radiación, etc) que permitan evaluar su efecto a largo del periodo de la producción agrícola y pecuaria.

Con este enfoque se ha desarrollado el paquete *{agromet}* que incluye una serie de funciones que pueden ser utilizadas de manera habitual para el cálculo de índices y estadísticos a partir de datos meteorológicos. Los datos de entrada se trabajan bajo la filosofía de datos *tidy* o datos ordenados [cita 1]. De esta forma, las funciones del paquete son genéricas, pudiendo ser aplicadas a cualquier set de datos tabulares sin importar su origen, orden o nombre de las columnas. De todas formas, el paquete incorpora herramientas para leer datos en un formato ascii particular, considerando como ejemplo de datos de entrada un archivo perteneciente al Observatorio Agrometeorológico de INTA Castelar-Bs.As. Argentina.

A continuación, se detallan algunas de las funciones destacadas:

- *umbrales()* permite contar la cantidad de observaciones que cumplen una determinada condición. Esta función podría ser utilizada para contabilizar la cantidad de días con temperaturas mínimas inferiores a 0°C durante el periodo crítico del cultivo de trigo o la cantidad de días con temperaturas máximas superiores a los 38°C para el cultivo de soja. Estos son dos ejemplos de seguimientos de temperaturas extremas que pueden generar daños en el cultivo.
- *dias_promedios()* devuelve el primer y último día del año promedio de ocurrencia de un evento. Esta función puede utilizarse para el análisis de datos históricos de heladas (por ejemplo, temperaturas mínimas inferiores o iguales a 0°C), pudiendo determinar la fecha media de primera y última helada, y de esta forma obtener valores estadísticos de esta adversidad.
- *olas()* identifica periodos de persistencia de un evento definido a partir de una condición lógica, por ejemplo días consecutivos con temperaturas mínimas inferiores a 0°C para análisis de heladas o con temperatura máxima superior a un umbral para el análisis de olas de calor.
- *ith()* esta función permite calcular el Índice de Temperatura y Humedad (ITH) ampliamente utilizado para el seguimiento del estrés calórico o confort del ganado lechero. El cálculo se realiza a nivel diario y considerando la ecuación Earl C. Thom (1959) [cita 2].
- *decil()* y *anomalía_porcentual()* permiten el cálculo de estos estadísticos para distintos intervalos de tiempo y pueden resultar útiles para el seguimiento de las precipitaciones mensuales permitiendo identificar meses en los cuales las mismas fueron superiores o inferiores a los valores medios históricos de una zona de interés.

Otras funciones como *spi()* y *spei()* permiten calcular el índice estandarizado de precipitación (SPI) mundialmente utilizado para el seguimientos de los periodos con excesos y déficit hídricos. Se adaptaron funciones de otros paquetes para ser compatibles con el manejo de datos tidy.

Además de las funciones para calcular índices, el paquete cuenta con funciones para trabajar con series de datos:

- *completar_serie()* esta función permite completar la serie de datos temporales definiendo alguna resolución disponible. Esta función es útil, por ejemplo, para completar los días ausentes o el registro ausente en una serie de datos diarios, completando dicho registro con NA (Not Available) en todas las variables.

Si se cuenta con la posibilidad de utilizar información meteorológica de distintas localidades de Argentina, {agromet} permite realizar una representación geoespacial de dicha información. Para ello se encuentran disponibles cartografías a nivel nacional, provincial y de países limítrofes (<https://www.naturalearthdata.com/>) y a nivel departamental (<https://www.ign.gob.ar>), cuya función es *mapa_argentina*. Para el gráfico de contornos se disponen de distintas escalas ya organizadas según intervalos de las variables y escala de colores (por ejemplo, *escala_temp_min()*).

Entre los autores del paquete se encuentran profesionales del Instituto Nacional de Tecnología Agropecuaria (INTA) es así que se incluye una función de graficado de datos georeferenciados *mapear()* con el estilo y logo propios de la institución. El paquete da la posibilidad de elaborar un reporte final en RMarkdown, el cual incluye un espacio para mapa y tabla de valores destacados. Este tipo de informe estandariza las salidas y cálculos de información producida por INTA relacionadas con la agrometeorología.

Si bien algunas de las funcionalidades incluidas en {agromet} ya existen en otros paquetes de R, su elaboración permitió unificarlas para que funcionen con los datos en una estructura genérica y generar funciones específicas como la creación de mapas y reportes con el estilo institucional del INTA. Así también, es un ejemplo más de cómo utilizar sistemas de desarrollo colaborativo utilizando herramientas de control de versiones como GitHub.

El repositorio se encuentra disponible en <https://github.com/AgRoMeteorologiaINTA/agromet>. {agromet} es de código abierto, está disponible gratuitamente, y en continuo desarrollo permitiendo mejorar las funciones que ya contiene como poder incluir nuevas según las necesidades de los usuarios. La licencia de uso es Licencia pública general de GNU y el proyecto cuenta con un código de conducta y una guía para quienes quieran contribuir.

Referencias:

1. Wickham, H. Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:<http://dx.doi.org/10.18637/jss.v059.i10> (2014).
2. SURFER, versión 9.0, Golden Software. Sitio web: <https://www.goldensoftware.com/>
3. Earl C. Thom. The Discomfort Index. Weatherwise 12, 57-59, (1959).

Natalia Gattinoni

INTA, Instituto de Clima y Agua
gattinoni.natalia@inta.gob.ar

pacote **Rocc**: gestão e análise de dados de ocorrências de espécies

O pacote **Rocc**(<https://libre.github.io/Rocc/index.html>) disponibiliza ferramentas para criar fluxos de trabalho reproduutíveis para tratar dados abertos de ocorrência de espécies. Atualmente existem diversas iniciativas de compilação de bases de dados digitais de biodiversidade, como GBIF no âmbito global e speciesLink no Brasil. Um grande volume de dados está disponível de forma aberta na rede, porém, converter a informação disponível em um dado a ser utilizado em um estudo exige uma combinação de rotinas automatizadas e conhecimento contextual para tomada de decisões. Este pacote se propõe a auxiliar a geração de rotinas automatizadas de obtenção e checagem de dados de biodiversidade usando ferramentas como speciesLink (<https://specieslink.net/>), Flora do Brasil (<http://floradobrasil.jbrj.gov.br/reflora/listaBrasil>) e GADM (<https://gadm.org/data.html>). O objetivo do pacote é automatizar tarefas repetidas e facilitar a tomada de decisão por parte de especialistas e não exclui a necessidade de uma avaliação contextual dos dados. O pacote permite estabelecer um fluxo de trabalho desde a obtenção do dado até a checagem de nomenclatura taxonômica e geográfica e também permite que cada uma das funções seja utilizada de forma independente.

O pacote **Rocc** possui três grandes eixos: obtenção de dados de ocorrência, checagem taxonômica e obtenção de dados geográficos. No eixo de obtenção de dados, o pacote possui uma função para fazer download de dados da base do speciesLink, `rspecieslink()`. Essa função se conecta com a API do speciesLink (<https://api.splink.org.br/>) e permite fazer o download de dados a partir de nomes científicos de espécies ou de municípios brasileiros e na América Latina. Na página do pacote, há um artigo com exemplos de diferentes usos da função *Downloading speciesLink data from R* (https://libre.github.io/Rocc/articles/articles/using_rspeciesLink.html). O pacote possui também a função `rgbif2()`, uma função para fazer o download de dados de ocorrência do GBIF (<https://www.gbif.org/>), que encapsula uma função do pacote **rgbif** (Chamberlain et al 2021) para criar uma saída compatível com o fluxo de trabalho. Uma vez tendo o download de dados das duas fontes, é possível combiná-los usando a função `bind_dwc()`. Isso permite à pessoa usuária obter e combinar dados de diferentes fontes que vêm de origem seguindo padrões diferentes. Os dados são formatados em uma tabela seguindo os campos do formato DarwinCore, o formato padrão para dados de biodiversidade.

Além disso, o pacote permite comunicação direta com a base de dados do projeto Flora do Brasil (Flora do Brasil 2020), que é uma iniciativa de catalogação das plantas do Brasil desenvolvida por um comitê de especialistas da área de botânica. **Rocc** possui uma função para fazer o download da base de dados completa da Flora do Brasil, `update_flora()`. É possível também fazer a busca de listas de espécies seguindo critérios de domínio fitogeográfico onde a espécie ocorre, endemismo, forma de vida ou tipo de vegetação usando a função `search_flora()`. O artigo *Searching for species names in FB2020* (https://libre.github.io/Rocc/articles/articles/searching_flora.html) contém exemplos de como usar essas funções.

Como o dado de biodiversidade corresponde a uma espécie em um dado local, é bastante usual passar os dados por uma checagem geográfica. Como existem diversos pacotes de R desenvolvidos para esse fim, como o **coordinateCleaner** (Zizka et al 2017) por exemplo, o pacote **Rocc** se propõe a efetuar o download de shapefiles e gazetteers que irão auxiliar no processo de checagem geográfica. No eixo de obtenção de dados geográficos o pacote **Rocc** possui três funções: `getGADM()`, `getGAZ()` e `getWDPA()`. A função `getGADM()` faz o download de shapefiles de unidades administrativas para países do globo a partir de GADM, <https://gadm.org/data.html> e retorna objetos de classe sf ou sp. A função `getGAZ()` faz o download do gazetteer do DIVA-GIS (<http://www.diva-gis.org/gData>). A função `getWDPA()` faz download de shapefiles da base de dados de áreas protegidas da IUCN para cada país (<https://www.protectedplanet.net>). Diferentemente das funções do pacote que se apoiam em bases como speciesLink e Flora do Brasil, esse eixo do pacote tem escopo global e pode ser aplicado em diferentes contextos. O artigo *Geographic data download with Rocc* (https://libre.github.io/Rocc/articles/articles/data_download.html) contém exemplos de uso das funções desse eixo.

A partir dos três eixos do pacote de obtenção de dados de biodiversidade, checagem taxonômica e obtenção de dados geográficos, é possível criar fluxos de trabalho reproduzíveis totalmente baseados em dados abertos para obtenção e limpeza de dados de biodiversidade. Na página do pacote existem artigos para guiar o desenvolvimento de fluxos de trabalho. Com isso, espera-se facilitar o acesso a diferentes bases e permitir autonomia às pessoas usuárias de criarem seus próprios fluxos de trabalho adaptados a sua necessidade. O pacote já vem sendo utilizado em projetos acadêmicos e consultorias de listagens de espécies regionais. Além disso, também tem sido utilizado em aulas de limpeza de dados de biodiversidade. Esperamos com isso, permitir que as pessoas usuárias possam focar seu trabalho mais em seu contexto e no significado biológico e menos em tarefas automatizadas de download e limpeza que já são feitas pelo pacote **Rocc**.

Palavras-chave: DIVA-GIS; flora do Brasil; GADM; GBIF; gazetteer; informática da biodiversidade; reproduzibilidade; speciesLink

Referências

Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K (2021). *rgbif: Interface to the Global Biodiversity Information Facility API*. R package version 3.5.0, <URL: <https://CRAN.R-project.org/package=rgbif>>.

Flora do Brasil 2020. Jardim Botânico do Rio de Janeiro. Available at: <<http://floradobrasil.jbrj.gov.br/>>. Accessed on: 28 Jul. 2021

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svanteson S, Wengstrom N, Zizka V, Antonelli A (2019). “CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases.” *Methods in Ecology and Evolution*, -7. doi: 10.1111/2041-210X.13152 (URL: <https://doi.org/10.1111/2041-210X.13152>), R package version 2.0-18, <URL: <https://github.com/ropensci/CoordinateCleaner>>.

INTEROPERABILIDAD Y GRANDES VOLÚMENES DE DATOS

Como potenciar el diseño de políticas públicas basada en evidencia

Objetivos Generales

Relevar diferentes fuentes de datos para el estudio de caso de interoperabilidad, que proporcionen evidencia para el diseño de las políticas públicas orientadas al sector apícola

Objetivos Específicos

- Realizar análisis exploratorios individuales de datos en las diferentes dependencias del Estado Nacional. (Coord Apícola - Ministerio de Agricultura, Ganadería y Pesca, Dirección de Seguimiento -Secretaria de Trabajo, Ministerio de Trabajo, Empleo y Seguridad Social Ministerio de Desarrollo Productivo, INCAA - Ministerio de Cultura, Superintendencia Riesgo de Trabajo)
- Encontrar fuentes de datos que puedan ser interoperables y que aporten mayor conocimiento del sector apícola, referidos a aspectos cuantitativos y cualitativos.
- Procesar datos y generar información relevante que permitan elaborar tableros de gestión para la toma de decisión.
- Formular una propuesta de mejora en el diseño de política pública a partir de la interoperabilidad de las bases de datos seleccionadas.

Antecedentes:

INTEROPERABILIDAD

Luego del 2000, en el ámbito de la Administración Pública Nacional se comienzan a discutir los términos de Interoperabilidad y Gobernanza que vienen impulsados por una nueva forma de concebir la acción del Estado y el uso de las TIC, ya no como una herramienta sino como una ventaja estratégica para el desarrollo.

Y nuestro país no es ajeno a esta nueva corriente y con la iniciativa de la Oficina Nacional de Tecnología Informática (ONTI) y el impulso del Foro Permanente de Responsables Informáticos en el año 2008 se plasman en una norma estos conceptos que dan origen al Componente de Interoperabilidad para el Gobierno Electrónico en el ámbito de la Oficina Nacional de Tecnologías de Información a través de la Resolución 99/2008 de la SECRETARIA DE GABINETE Y GESTION PUBLICA

DATOS ABIERTOS

A partir del paradigma de Gobierno Abierto apoyado fuertemente sobre 4 ejes fundamentales como Transparencia, Participación Ciudadana, Rendición de cuentas y Datos abiertos se comienza a transitar un nuevo camino.

En nuestro país con la publicación del decreto 117/2016 se reglamentan plazos tanto para la publicación de los primeros conjuntos de datos del portal de datos abiertos denominado <http://datos.gob.ar> como los tiempos para detallar los activos de datos de los distintos organismos de la Nación y su cronograma de publicación. Este instrumento complementa el Régimen de Acceso a la Información Pública creado por la Ley 27275 en el año 2016.

Con estas herramientas se genera una nueva concepción para la gestión de la información y los datos y de esta manera, la información comienza a ser una evidencia al momento de diseñar y concebir las Políticas Públicas.

En base al estudio de estos dos antecedentes creemos que la complementación de ambos puede proporcionar mejoras a la hora de analizar y diseñar políticas públicas, por eso nos propusimos hacer un ejercicio

Actividades y metodología:

- Elección del Registro Nacional de Productores Apícolas del área Coord Apicultura - MAGyP.
- Realizar análisis exploratorio e Investigación en los diferentes organismos con fuentes de datos que se relacionen con la temática.
 - Los integrantes buscan datos que se interrelacionan con la fuente de datos seleccionada, en los distintos organismos oficiales.
- Depurar y analizar datos, utilizando métodos de cleansing con el lenguaje R para relacionar y generar información que aporte mejor conocimiento de la problemática. Se incluirá en el análisis, las fuentes de información existentes en las áreas mencionadas en los objetivos del presente documento.
- Identificar a partir del análisis de las fuentes de datos los actores indirectos que pueden aportar a la mejora del diseño de política pública.
 - buscar información del sector apícola (aserraderos, envases metálicos, implementos apícolas, etc)
- Analizar datos disponibles y realizar la anonimización en caso de tratarse de datos que revistan carácter privado
 - Si la información tiene carácter privado/no público
- Elaborar informes y/o reportes utilizando QGIS y/o R que evidencien la interoperabilidad de las diferentes fuentes de datos de información analizadas y permitirán aportar mayores herramientas para el diseño de políticas públicas.
- Generar documento con la propuesta de mejora que ofrece la interoperabilidad al diseño de política pública.
- Generar documentos de análisis utilizando RMarkDown

Factibilidad:

A partir de acuerdos entre instituciones/organismos a cargo de la elaboración de la fuentes de datos a utilizar

Conclusión

En base a este ejercicio pudimos mostrar los beneficios que proporciona poner en práctica herramientas que permitan utilizar la información con la que cuenta el Estado y vincularla a partir de criterios de interoperabilidad.

Desde su proceso inicial para el desarrollo de Políticas Públicas partiendo del Diagnóstico, la dinámica de interoperar información relacionada con la política a desarrollar nos proporciona mayor conocimiento de la temática a abordar, definiendo de manera clara la línea de base para eventuales evaluaciones

También pueden aplicarse las conclusiones obtenidas del cruce de información para focalizar políticas ya desarrolladas.

O en su defecto utilizar el ejercicio de interoperación como herramienta para las métricas de evaluación de las políticas.

Pero por sobretodas las cosas permite diseñar políticas que cuenten con más de un enfoque de las realidades a desarrollar, entendiendo al Estado como una unidad y sumando las potencialidades de cada una de las carteras de gobierno.

Resultados

Los resultados alcanzados utilizando paquetes de R existentes y proponiendo nuevos desarrollos se encuentran detallados en el trabajo completo en https://rpubs.com/julietacoll/trabajo_final_UNAB

Aplicación de R para analizar la perspectiva del consumidor sobre la consistencia de alimentos

Anónimo

Palabras clave: consistencia - consumidores - Rstudio

Los estudios con consumidores de tipo cualitativo permiten comprender las percepciones del consumidor y decisiones de compra de productos alimenticios (Gambaro 2018). El término “consistencia” suele usarse en Argentina cuando se habla de textura de un alimento, pero es un término no utilizado como descriptor en el análisis sensorial de la textura. Por ello se propuso como objetivo utilizar las herramientas de R para explorar la perspectiva del consumidor sobre el concepto de consistencia de un alimento.

Mediante un muestreo no probabilístico del tipo bola de nieve se realizó un cuestionario online semiestructurado mediante Google Forms, a través de diferentes redes sociales (Facebook, Instagram, Whatsapp). Las participantes tuvieron que definir con sus palabras “consistencia de un alimento,” “alimento consistente,” “alimento muy consistente” y “alimento poco consistente.” Los términos empleados en las definiciones se agruparon en categorías por similitud de respuesta. Para el análisis de los datos se utilizó RStudio (versión 4.0.3) (Team 2020). Se utilizó la librería tm (Feinerer and Hurnik 2020) para eliminar, mayúsculas, signos de puntuación y palabras que no eran de interés (artículos, conectores, etc.) y para analizar la frecuencia absoluta de los términos usados para definir los conceptos. Para la visualización de los mismos se utilizó la librería ggplot2 (Wickham 2016). Para explorar las asociaciones de las categorías obtenidas con los conceptos solicitados se realizó un análisis de correspondencia simple utilizando la librería FactoMineR (Lê, Josse, and Husson 2008).

Participaron 390 consumidores (337 eran mujeres, y 345 tenían entre 18-50 años) utilizando 130 términos diferentes para definir consistencia, siendo los 10 más frecuentes (Figura 1): textura (12,7 %), nutritivo (7,6 %), firmeza (6,7 %), dureza (5,3 %), solidez (5,3 %), densidad (3,3 %), composición (3,1 %), calidad (2,4 %), alimento (2,1 %) y blando (2,1 %). Los términos fueron agrupados en 22 categorías, de las cuales las más frecuentes fueron: nutrición y salud (16 %), textura en general (13 %), estado físico (10 %), apariencia y textura visual (9 %), vida útil (7,7 %), firmeza (7,4 %), dureza (6,9 %) y características físicas/químicas (6,9 %).

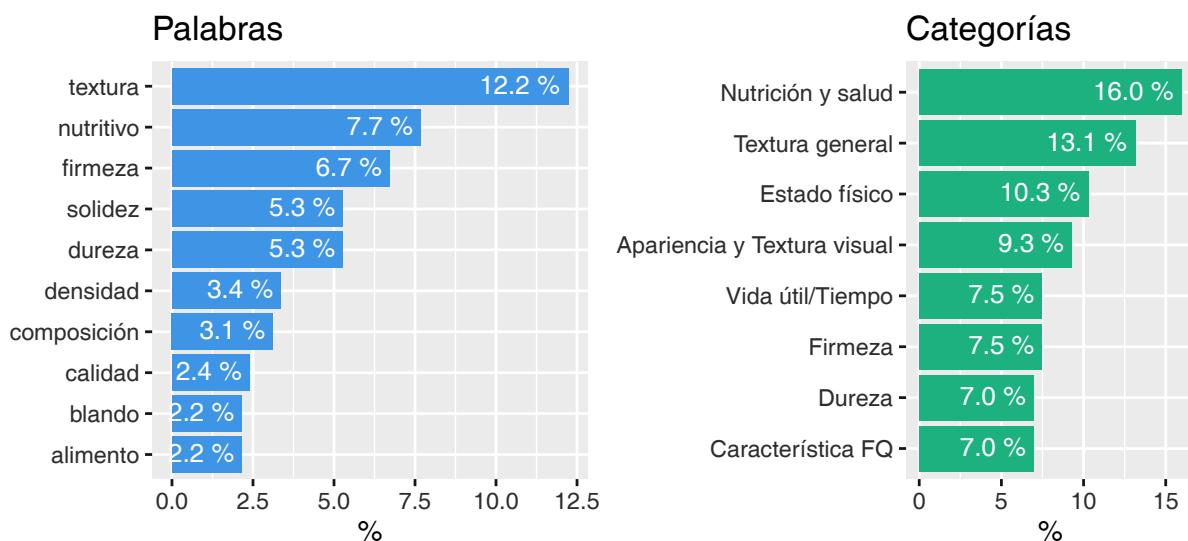


Figura 1. Palabras y categorías más usadas para definir consistencia de un alimento.

Del análisis de correspondencia se obtuvieron dos dimensiones con un 77 % de inercia (Figura 2). Se observó que los consumidores asociaron la definición de consistencia con términos relacionados a la textura, nutrición y salud, al sabor y términos hedónicos. Además, definieron los alimentos muy consistentes y consistentes como

alimentos de muy alto y alto valor nutritivo, utilizando términos relacionados a la textura auditiva y manual, dureza y dando como ejemplos alimentos de textura sólida o blanda. Al contrario, han asociado alimentos poco consistentes con alimentos de bajo valor nutritivo, utilizando términos relacionados a la blandezza y fluidez y exemplificando con alimentos de textura semisólida o líquidos.

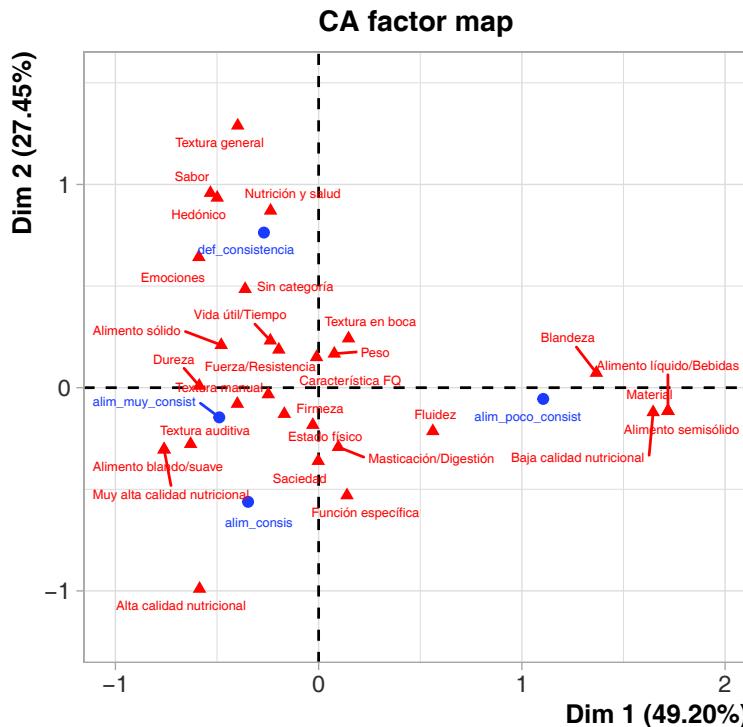


Figura 2. Biplot de categorías de palabras utilizadas para definir consistencia, alimento consistente, poco consistente y muy consistente.

Como conclusión, los consumidores asociaron la “consistencia” de un alimento con su valor nutricional y con distintas características relacionadas a la textura. La aplicación de R para estudios con consumidores resultó práctica y efectiva, siendo fácilmente reproducible para futuras investigaciones en el área.

Referencias bibliográficas

- 10 Feinerer, Ingo, and Kurt Hurnik. 2020. “Tm: Text Mining Package. R Package Version 0.7-8.” <https://cran.r-project.org/package=tm>.
- Gambaro, Adriana. 2018. “Projective Techniques to Study Consumer Perception of Food.” *Current Opinion in Food Science* 21 (June): 46–50. <https://doi.org/10.1016/j.cofs.2018.05.004>.
- Lê, Sébastien, Julie Josse, and François Husson. 2008. “FactoMineR : An R Package for Multivariate Analysis.” *Journal of Statistical Software* 25 (1). <https://doi.org/10.18637/jss.v025.i01>.
- Team, R Core. 2020. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Use R! Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-24277-4>.

Conjuntos conformacionales de Proteínas Intrínsecamente Desordenadas moldean las velocidades de evolución dando origen a patrones conformacionales

Julia Marchetti^{1,*}, Nicolas Palopoli^{1,*}, Alexander Miguel Monzon², Diego Javier Zea³, Maria Silvina Fornasari¹, Silvio C.E. Tosatto² and Gustavo Parisi¹.

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina

2 Department of Biomedical Sciences, University of Padua, Padua, Italy

3 Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Paris, France

* Estos autores contribuyeron de igual manera a la elaboración del trabajo.

Introducción

Las proteínas intrínsecamente desordenadas (Intrinsically Disordered Proteins, IDPs) son cadenas de aminoácidos que carecen de una estructura estable en condiciones fisiológicas. Esta característica las diferencia de las proteínas más tradicionales y las hace particularmente desafiantes para la biología estructural y la biología evolutiva. En este trabajo se propone usar un enfoque evolutivo para la descripción de la extensa diversidad conformacional presente en cada IDP y comprender cuáles son los aspectos estructurales que impactan de manera directa sobre la velocidad de evolución en cada una de sus posiciones.

Los experimentos de determinación estructural de las proteínas proveen información sobre las coordenadas espaciales de cada uno de los átomos que las conforman. La ventaja de la Resonancia Magnética Nuclear (NMR, por sus siglas en inglés) sobre otras técnicas similares es que permite obtener información estructural de múltiples modelos en simultáneo, que representan en su conjunto el comportamiento de la proteína en solución, cercano a su comportamiento natural en la célula.

Metodología y herramientas

Para este análisis se recolectó información estructural de *ensembles* o conjuntos de conformaciones alternativas de 311 proteínas intrínsecamente desordenadas, provenientes de experimentos de NMR. Para cada uno de sus modelos estructurales se determinó el carácter de desorden intrínseco posición a posición. También se estimó la tasa sitio-específica de intercambio por otros aminoácidos a lo largo de la evolución, inferida a partir de alineamientos de proteínas vinculadas evolutivamente, con el programa Rate4Site (1). Por lo tanto en este trabajo se contó con información sitio-específica y modelo-específico, tanto estructural como evolutiva, dando lugar a un total de 790,128 datos para analizar.

Utilizamos paquetes estándar de R para llevar a cabo los análisis necesarios para derivar información científica relevante sobre nuestro conjunto de datos. En particular y para cada conformación presente en el conjunto conformacional de una proteína dada, consideramos cada uno de sus aminoácidos por separado y exploramos la relación entre los contactos establecidos con otros aminoácidos de la proteína, su carácter de desorden intrínseco y su velocidad de evolución.

Para el análisis estadístico se utilizaron funciones del paquete tidyverse (2), mientras que la visualización de datos y la generación de gráficos para publicación se realizó con ggplot2 (3) , ggpublisher (4), patchwork (5) y ggttext (6). El reporte general del trabajo se realizó con R Markdown. Junto con el repositorio abierto en gitlab (<https://gitlab.com/sbgung/publications/palopoli-marchetti-2020-rates>) esto permitió asegurar la reproducibilidad del análisis, pilar de la ciencia moderna y requisito para su publicación (7)

Resultados

Encontramos que las IDPs presentan una gran heterogeneidad en su velocidad de evolución sitio-específica, asociadas a diferentes restricciones estructurales impuestas por los contactos establecidos entre residuos. La correlación entre el establecimiento de contactos y las velocidades de evolución mejora cuando se tiene en cuenta información estructural derivada de múltiples confórmeros, pero no del conjunto redundante de estructuras disponibles. El análisis simultáneo de distintas proteínas permitió identificar perfiles de velocidad característicos y compartidos entre ellas, que se vinculan con la alternancia de regiones ordenadas y desordenadas en las proteínas.

Bibliografía

1. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*. 2002;18 Suppl 1:S71-7.
2. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. *JOSS*. 2019 Nov 21;4(43):1686.
3. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
4. Alboukadel Kassambara (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0.<https://CRAN.R-project.org/package=ggpubr>
5. Thomas Lin Pedersen (2020). *patchwork: The Composer of Plots*. R package version 1.1.1. <https://CRAN.R-project.org/package=patchwork>
6. Claus O. Wilke (2020). *ggttext: Improved Text Rendering Support for 'ggplot2'*. R package version 0.1.1. <https://CRAN.R-project.org/package=ggttext>
7. Palopoli N, Marchetti J, Monzon AM, Zea DJ, Tosatto SCE, Fornasari MS, et al. Intrinsically disordered protein ensembles shape evolutionary rates revealing conformational patterns. *J Mol Biol*. 2021 Feb 5;433(3):166751.