

Actas de LATINR 2020

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

Editoras: Yanina Bellini Saibene, Florencia D'Andrea,
Riva Quiroga, Natalia da Silva



Diseño Gráfico

Dis. Gráf. Francisco Etchart

Enero de 2022

Equipo

Chairs

Yanina Bellini Saibene

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

Florencia D'Andrea

- CONICET - Instituto Nacional de Tecnología Agropecuaria
- R-Ladies BsAs, Argentina + Global Team

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

Encargados Comité Científico

Carmen Le Foulon

- Instituto de Ciencia Política, Universidad Católica de Chile

Ricardo Olea

- Departamento de Estadística, Universidad Católica de Chile

Comité Científico

Marcela Alfaro

- Universidad de Costa Rica

Ignacio Alvarez-Castro

- IESTA-Universidad de la República

Steve Arndt

- University of Iowa, Estados Unidos

Mathias Bourel

- IMERL-Universidad de la República

Marina Cock

- Universidad Nacional de La Pampa - CONICET

Mario Cortina Borja

- University College London

Maria Inés Fariello

- IMERL-Universidad de la República

Ileana Frasier

- INTA - CONICET

Adriana Gili

- Universidad Nacional de La Pampa

Juan José Goyeneche

- IESTA-Universidad de la República

Leonardo Moreno

- IESTA-Universidad de la República

Priscilla Minotti

- Universidad Nacional de San Martín

Germán Rosati

- IDAES-UNSAM / CONICET, Argentina

Marcelo Soria

- CONICET-Universidad de Buenos Aires

Walter Sosa Escudero

- CONICET-Universidad de San Andrés

Vianey Leos

- North Carolina State University

Jairo Cugliari

- ERIC-Université Lumière Lyon 2

Lucia Rodriguez Planes

- Universidad Nacional de Tierra del Fuego -
CONICET

Laura Ación

- LatinR Inaugural Local Chair / UBA-CONICET,
Argentina

Sara R. Mara

- Researcher at International Institute for
Sustainability (IIS-Rio), Rio de Janeiro, Brazil

Juan Cruz Rodriguez

- FAMAF - Universidad Nacional de Córdoba

Carmen Le Foulon

- Centro de Estudios Públicos (CEP)

Zulema Bazurto Blacio

- Universidad de Guayaquil

Comité Organizador**Elio Campitelli**

- Universidad de Buenos Aires, Argentina
- R en Buenos Aires, Argentina

Paola Corrales

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

Patricio Cofré

- Metrics Arts, Chile

Roxana Noelia Villafañe

- R-Ladies Resistencia Corrientes, Argentina
- R en NEA, Argentina
- LEMyP (IQUIBA-NEA) CONICET

Patricia Loto

- R-Ladies Resistencia Corrientes, Argentina
- Universidad Nacional del Nordeste

Gabriela Sandoval

- R-Ladies Santiago, Chile

Daniela Vazquez

- R-Ladies Montevideo, Uruguay
- GURU

Francisco Etchart

- INTA Anguil

Fernando Hannaka

- Blue Economic Solutions

María Nanton

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

Andrea Gómez Vargas

- Universidad de Buenos Aires
- R Ladies Buenos Aires

Beatriz Milz

- Universidad de São Paulo
- R-Ladies São Paulo

Cintia Callamullo León

- Fundación Dr. Manuel Sadosky -
Universidad de Buenos Aires
- R en Buenos Aires

Maine Fariello

- Universidad de la República
- R-Ladies Montevideo, Uruguay
- URU

Paula Pereda

- Secretaría de Transformación Productiva y
Competitividad
- R-Ladies Montevideo, Uruguay
- GURU

Haydee Svab

- ASK-AR Data Consultancy
- RLadies São Paulo

Prólogo

La tercera edición de LatinR estaba planificada para ser llevada a cabo en Montevideo, Uruguay. Como todas las actividades del 2020 la conferencia tuvo que realizarse de forma virtual debido a pandemia provocada por el COVID-19. Se decidió no cobrar registro para participar de la conferencia.

En esta edición hubo dos charlas invitadas: *Learning without a net* por Alison Presmanes Hill y *Blogueo con R Markdown* por Maëlle Salmon.

En días previos y posteriores a la conferencia se realizaron nueve tutoriales que fueron hospedados por capítulos de R-Ladies y otros grupos de usuarios de R de la región. Seis tutoriales se impartieron en español y tres en portugués:

- Expresiones regulares para la limpieza y transformación de datos por Riva Quiroga y Stephanie Orellana y ofrecido de manera conjunta por RLadies Talca, RLadies Galápagos, RLadies Ciudad de México, RLadies Valparaíso y RLadies Santiago.
- Aplicaciones web interactivas con Shiny por Florencia D'Andrea, Juan Cruz Rodríguez y Vilma Romero hospedado por RLadies Querétaro y RLadies Concepción.
- Escrevendo manuscritos acadêmicos usando rmarkdown a cargo de Andrea Sánchez - Tapia y Sara Ribeiro Mortara ofrecido de manera conjunta por RLadies Natal y RLadies Rio de Janeiro.
- Importando datos desde hojas de cálculo a cargo de Luis D. Verde Arregoitia hospedado de manera conjunta por capítulos de RLadies Puebla, Xalapa, Barranquilla y el grupo de R de Rosario.
- Introducción al ABC para enseñar on-line a cargo de Yanina Bellini Saibene, Mariela Rajngewerc organizado por R-Ladies General Pico, R-Ladies Santa Rosa y MetaDocencia.
- Generando tutoriales interactivos con el

paquete learnr a cargo de Yanina Bellini Saibene y Paola Corrales también organizado por R-Ladies General Pico, R-Ladies Santa Rosa y MetaDocencia.

- Introducción a Machine Learning con Tidymodels a cargo de Roxana Noelia Villafañe, Ana Laura Diedrichs y Patricia Loto ofrecido de manera conjunta por los capítulos RLadies Cuernavaca, R-Ladies Resistencia-Corrientes y RLadies Mendoza.
- Shiny: zero to hero a cargo de William Amorim, Julio Trecenti y Athos Damiani y hospedado por RLadies Niterói.
- Comunicando seus resultados com R: aprenda a criar apresentações reprodutíveis a cargo de Beatriz Milz y Haydee Svab ofrecido por RLadies Goiania.

También tuvimos un día dedicado a ReproHack, Hackatón de reproducibilidad computacional donde se presentaron las charlas:

- ¿Por qué es importante la reproducibilidad computacional?. Daniela Ballari. Universidad de Azuay, Ecuador.
- Consejos para reproducir un artículo científico. Florencia D'Andrea. INTA-CONICET, Argentina.
- Contribuyendo al código libre en R. Juan Cruz Rodríguez. FAMAF-UNC, Argentina.
- Cómo escribir manuscritos reproducibles. Francisco Rodríguez-Sánchez. Universidad de Sevilla, España.
- Reproducibilidad en torno a una aplicación web Pablo Bernabeu. Universidad de Lancaster, UK.
- Revisar paquetes para una mejor ciencia - Maëlle Salmon
- La travesía de taxlist en ROpenSci - Miguel Alvarez

Debido al cambio de formato, los autores pudieron optar por presentar en esta edición online o esperar a la edición 2021. Aquellos que decidieron presentar en la edición 2020, grabaron un video de su charla con el tiempo correspondiente (15 minutos para charlas regulares y 5 para charlas relámpagos) y lo enviaron previamente a la conferencia. Esos videos se cargaron al canal de YouTube de LatinR y se publicaron días previos a la conferencia en listas agrupados en las sesiones temáticas correspondientes. De esta manera los asistentes a la conferencia pudieron ver las charlas con anticipación y participaron del encuentro en vivo con preguntas y comentarios.

Los días de la conferencia los autores de cada sesión se reunieron de forma sincrónica en un seminario web realizado en la plataforma zoom con streaming a YouTube y contestaron en vivo las preguntas y comentarios de los asistentes. También se repasaron los comentarios que habían dejado en sus videos en el canal de YouTube de la conferencia. En estas actas se publican los 23 trabajos presentados durante la conferencia.

Agradecemos a nuestros sponsors, RStudio, INTA y RConsortium por su apoyo.

Índice

Sesión Minería de textos y web scraping

MinaR los discursos pResidenciales 10
Juan Pablo Ruiz Nicolini, Lucas Enrich y Camila Higa

Leyendo todas las noticias del mundo con R (and friends) 12
Brenda Walter y Antonio Vazquez Brust

Taquigráficos. Análisis de discursos políticos de la Legislatura porteña con R 14
Sofía Santamarina y Manuel Zapico

Avaliação da transparência em atas de reuniões dos Comitês de Bacias Hidrográficas na Macrometrópole Paulista (Brasil) 16
Beatriz Milz

Sesión Desarrollo de paquetes

Latinr en LatinR: automatizando el envío de trabajos a conferencias de R con un paquete de R 19
Elio Campitelli

geouy: Acceso a las geometrías de Uruguay 21
Richard Detomasi

Una librería para procesar datos públicos. Presentación del paquete eph 23
Germán Rosati, Diego Kozlowski, Pablo Tiscornia, Guido Weksler y Natsumi Shokida

{polAr}: Política Argentina Usando R 25
Juan Pablo Ruiz Nicolini

desuctools: paquete con funciones y bases de datos para análisis de encuestas sociales 27
Cristián Ayala y Cristina Marchant

Sesión Shiny y visualización de datos

Diseñando visualizaciones atractivas e informativas: el caso de {ggplot} y {highcharter} 30
Mariana Villamizar Rodríguez, Juan Pablo Marín Díaz y Camila Achuri

Seguimiento diario de la producción científica sobre COVID-19 31
Juan Pablo Sokil

Creando componentes de Javascript personalizados para Shiny: Cómo se creó el paquete {shinyinvoer} 33
David Daza, Camila Achuri y Juan Pablo Marín Díaz

Cómo usar R con tidyverse para la lucha anti-corrupción 34
Juliana Galvis, Camila Achuri y Juan Pablo Marín Díaz

"#TuitómetroNacional" 35
Juan Pablo Ruiz Nicolini y Camila Higa

Sesión Aplicaciones con datos educativos y políticas sociales

Aplicaciones de R al estudio de enseñanza universitaria: el caso de la Facultad de Ingeniería 38
Daniel Alessandrini, Martín Pratto Burgos y Fernando Fernández Barreiro

R para el monitoreo de políticas sociales 40
Elina Gomez y Sofía Harley

Registros de Moodle para el seguimientos de actividad estudiantil y diseño de estrategias educativas 42

Ignacio Alcántara, Pablo Bobadilla, Claudia Borlido, Paola Cabral, José Passarini y Nicole Rosenstock

Análise preditiva do desempenho dos alunos do curso de pedagogia no âmbito da educação à distância no ENADE 44

Natan Borges y Ranah Costa

Sesión Datos espaciales

Uso de R e imágenes satelitales para la generación de modelos predictivos y caracterización de plantaciones forestales 47

Matías Gaute, Teresa Boca, Hugo Fassola, Ernesto Andenmatten

Optimización de parámetros geoestadísticos para interpolación de lluvia para uso en el sector agropecuario 49

Adrián Cal, Guadalupe Tiscornia

Integración de herramientas SIG, bases de datos, R espacial, Rmarkdown en la automatización de informes periódicos de la información en el Programa de Erradicación de plagas ISCA-MEN, Mendoza, Argentina 51

Teresa Boca, Mariel Vanin, Carlos Flores y Alejandro Asfennato

Sesión Modelos y aplicaciones

Ventajas del análisis de redes para optimizar la vigilancia y control de Brucelosis bovina en la Argentina 54

Alarcón Laura, Marcos Andrea, Grave Emiliano, Hard Jorge, Cipriotti Pablo

Análise de sensibilidade de respostas de contagem 56

Alejandra Tapia Silva

R en Temas de Industria 58

Irma Noemi No, Andrés Redchuk, Luis Alberto Orlandi and Julián Eloy Tornillo

Sesión **Minería de textos y web scraping**

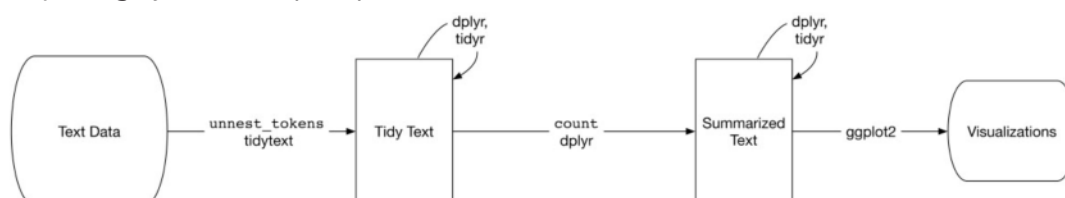
MinaR los discuRsos pResidenciales

Abstract

El primero de marzo de cada año las cámaras de Diputados y Senadores de la Nación Argentina se reúnen en asamblea para dar comienzo al año legislativo y el presidente de turno encabeza el acto con un discurso¹. Éstos suelen girar en torno a los ejes de gobierno o promesas y objetivos del año. Es notorio que estos mensajes tiene un estilo y contenido marcado por quien ejerce el gobierno. En este trabajo analizamos el texto contenido en los discursos presidenciales desde el primero en 1854 por Justo José De Urquiza hasta el último en 2020 por Alberto Fernández.

La minería de texto como estrategia de investigación es de utilidad para un rápido y eficiente análisis exploratorio del gran volumen de información contenida en los discursos presidenciales. Dentro del ecosistema de R este campo ha ido creciendo sostenidamente. Liberías como `tm` y `topicmodels` son herramientas poderosas para el procesamiento, manipulación y modelado de la información contenida en el texto. Siguiendo la filosofía de `tidyverse`, Silge y Robinson (2016) desarrollaron `tidytext` que facilita una primera introducción a esta técnica de investigación y su integración con otras como `ggplot2` para la visualización.

Un flujo de trabajo como el descrito anteriormente puede ilustrarse siguiendo el esquema propuesto por Silge y Robinson (2020):



1 La fuente original de todos los discursos puede consultarse en línea en https://www.hcdn.gob.ar/secparl/dgral_info_parlamentaria/dip/documentos/mensajes_presidenciales.html. Los mismos fueron posteriormente digitalizados mediante proceso de OCR y se encuentran disponibles para descargar desde R a través del paquete `{polAr}` (Ruiz Nicolini 2020).

2 Traducción propia de The tidy text format (Silge and Robinson 2016).

3. Trabajamos con dplyr para calcular frecuencias de palabras, tidytext para identificar las más relevantes comparadas entre discursos (tf-idf) y ggplot2 para las visualizaciones.

Ejemplo: Trayectoria en el discurso

Mediante las técnicas TF-IDF y Principal Component Analysis (PCA) es posible embeber numéricamente los discursos considerando el uso de palabras en cada uno ponderados por su longitud y así visualizar los presidentes de mayor y menor variabilidad discursiva en comparación con los promedios de los demás.

Las limitaciones de técnicas basadas en la frecuencia de palabras independientemente del orden (como son Bag of Words y TF-IDF) están vinculadas con el vocabulario por un lado, y con la semántica por el otro. Así, para trabajar mejor con textos históricos (el más antiguo en este caso tiene 166 años) es recomendable usar técnicas que hagan uso del contexto como puede ser Doc2Vec.

Referencias

- Ruiz Nicolini, Juan Pablo. 2020. "PolAr: Argentina Political Analysis." <https://github.com/electorArg/polAr>.
- Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *Journal of Open Source Software* 1 (3): 37. <https://doi.org/10.21105/joss.00037>.
- ——. 2020. *Text Mining with R*. O'Reilly. <https://www.tidytextmining.com/>.
- Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software, Articles* 59 (10): 1–23. <https://doi.org/10.18637/jss.v059.i10>.



Leyendo todas las noticias del mundo con R (and friends)

Introducción

A pesar del éxito de la vacunación, que en el último siglo ha prevenido la muerte y enfermedad de millones de niños, crece la desconfianza hacia las vacunas alrededor del mundo. Encuestas globales indican que en los países desarrollados está declinando la confianza respecto a la seguridad y valor de las vacunas, a la vez que reaparecen brotes de enfermedades antes controladas, como el sarampión en Europa.

La desinformación sobre los peligros de las vacunas circula en medios masivos y en redes sociales, y la reticencia a vacunar a los menores se transforma en un movimiento, llamado “antivacunas” o antivax. Ante la gravedad de la situación, en 2019 la Organización Mundial de la Salud declaró al movimiento antivacunas como una de las 10 mayores amenazas a la salud pública mundial. La repercusión del mensaje antivacunas depende de la producción y la circulación por medios de comunicación y redes sociales de discurso, a favor o en contra, acerca de la inmunización. Para poder atender el problema de la desinformación, necesitamos entender la dinámica de la circulación de noticias y artículos de opinión, contestando preguntas como:

Qué mensajes circulan sobre las vacunas en los medios locales?

Qué peso tiene el discurso antivacunas en los medios de comunicación?

Cuáles son los principales formadores de opinión sobre de las vacunas?

A leer todos los diarios con R... y sus amigos

Una forma encontrar respuestas sería acce-

diendo a un archivo diario, que cubra los últimos años, con el contenido de todas las noticias publicadas en un país por sus periódicos y portales de noticias. Tal cosa existe: el proyecto GDEL (Global Database of Events, Language, and Tone)¹. GDEL es una inmensa base de datos de

acceso libre sobre la sociedad humana, considerada como la “de mayor tamaño, más completa, y de mayor resolución jamás creada”. Crece cada día gracias a una iniciativa que monitorea las noticias online de una gran cantidad de países en más de 100 idiomas, acumulando las noticias publicadas e identificando las poblaciones, ubicaciones, organizaciones, temas y emociones presentes en cada artículo. Actualizada cada 15 minutos.

Y si: podemos usar R para acceder a la base GDEL y procesar sus registros, utilizando paquetes clave como `dbplyr` y `bigrquery`. Y en combinación con un plantel de invitados open source especializados (como la librería de Python `newspaper3K` para extracción de contenidos periodísticos, o la librería de C `fastText` para clasificación automática de texto) podemos crear un flujo de trabajo automatizado para monitoreo continuo. El sistema de clasificación de texto `fastText` emplea una variante de la estrategia conocida como “bag of words”. Convierte cada palabra una serie de atributos numéricos (un vector). Al analizar un texto, sólo extrae el identificador numérico de cada palabra, sin que sea necesario considerar el orden en que aparecen los términos ni su semántica -de allí el nombre, una simple “bolsa de palabras”. La particularidad del método `fastText` es que incorpora una mejora respecto al bag of words tradicional, al considerar n-gramas (secuen-

¹ <https://www.gdelproject.org/>



Figura 1: Noticias sobre vacunas en medios online argentinos durante 2019

cias de palabras) y no sólo vocablos sueltos al realizar la conversión en vectores.

Con estas herramientas podemos poner en marcha un sistema que, día a día, identifica los artículos del país de interés. Y luego extrae los que mencionan a las vacunas (o el tema que se investigue), recupera su contenido completo y detecta los tópicos presentes utilizando un modelo entrenado por machine learning supervisado.

Resultados

La clasificación automatizada de noticias vía fastText logró un muy alto nivel de precisión, basándose en el etiquetado por revisión humana del tópico central en algunos centenares de artículos. En cuanto al contenido de las noticias, encontramos unos 42 artículos por mes que hablan sobre el fenómeno del antivacunismo; 8 de ellos en alguno de los grandes medios de alcance nacional. El tema es siempre abordado sin atacar la efectividad o valor de la vacunación pública desde la línea editorial. En general se reportan percances asociados al antivacunismo - como el rebrote de enfermedades prevenibles que se creían controladas, o el arresto de militan-

tes antivacunas en algún país- sin insinuar que la postura antivacunas merezca ser considerada. Con una excepción: el aspecto de la farándula, cuando se informa que una u otra persona famosa declara estar en contra de las vacunas, o que en algún show de TV se han invitado a exponentes pro y contra vacunas a exponer en igualdad de condiciones. Esto sugiere que un importante vector para la erosión de la confianza pública en la vacunación podría ser la atención dedicada a personalidades mediáticas que toman la causa del antivacunismo.

Contenido de la presentación

La presentación incluirá una breve introducción a la problemática del antivacunismo, y un repaso al proyecto GDELT. También describirá los métodos utilizados para conectar R a GDELT para extraer noticias de medios locales sobre vacunas, y luego la técnicas de clasificación de contenido de noticias con machine learning. Los minutos finales serán dedicados a presentar resultados preliminares.

TaquiGráficos: Palabras Legislativas. Una propuesta de análisis de las versiones taquigráficas de la Legislatura de la Ciudad Autónoma de Buenos Aires

Con *TaquiGráficos, Palabras Legislativas* deseamos contribuir al análisis de los debates que se llevan a cabo “dentro” de la Legislatura de la Ciudad de Buenos Aires. El objetivo general es producir visualizaciones que recuperen las características de las sesiones en las que se votan cuestiones determinantes para la vida de los y las habitantes de la Ciudad. Consideramos que lo tratado dentro de la Legislatura no ha recibido suficiente atención: los resultados de la votación son publicados, pero no las posiciones esgrimidas por los y las legisladores y sus asesores en las instancias de debate previo en las diferentes comisiones.

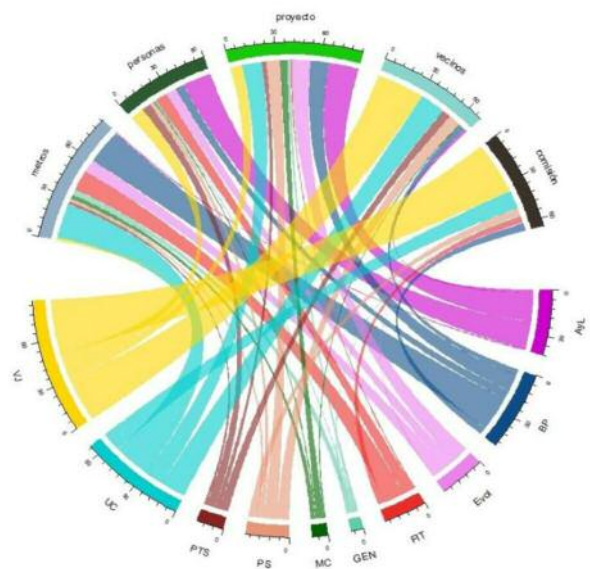
Por tal motivo, en este análisis nos centramos en el debate “dentro del recinto” para recuperar algunas palabras claves y dimensiones importantes de lo discutido en la sesión legislativa. A partir de las versiones taquigráficas de la sesión, publicadas en el sitio web de la Legislatura,¹ elaboramos visualizaciones de las palabras más repetidas por bloque y legislador, así como también construimos dimensiones de los conceptos tratados en las sesiones. La limpieza de datos y las visualizaciones fueron realizadas íntegramente en R, especialmente mediante los paquetes Tidyverse y Ggplot. En este trabajo mostraremos de qué manera obtuvimos los datos, cómo los adaptamos y graficamos, y qué conclusiones desarrollamos a partir de esto.

Para esta presentación en particular proponemos analizar el proceso legislativo que concluyó con la aprobación del Código Urbanístico de la CABA en diciembre de 2018. El Código Urbanístico define las condiciones de edificación y de los aspectos patrimoniales y ambientales, así

como también regula los dominios públicos y privados. Por estos motivos, el análisis del código constituye una herramienta fundamental para conocer las dinámicas de distribución social del territorio y, de esa manera, analizar desigualdades socioeconómicas.

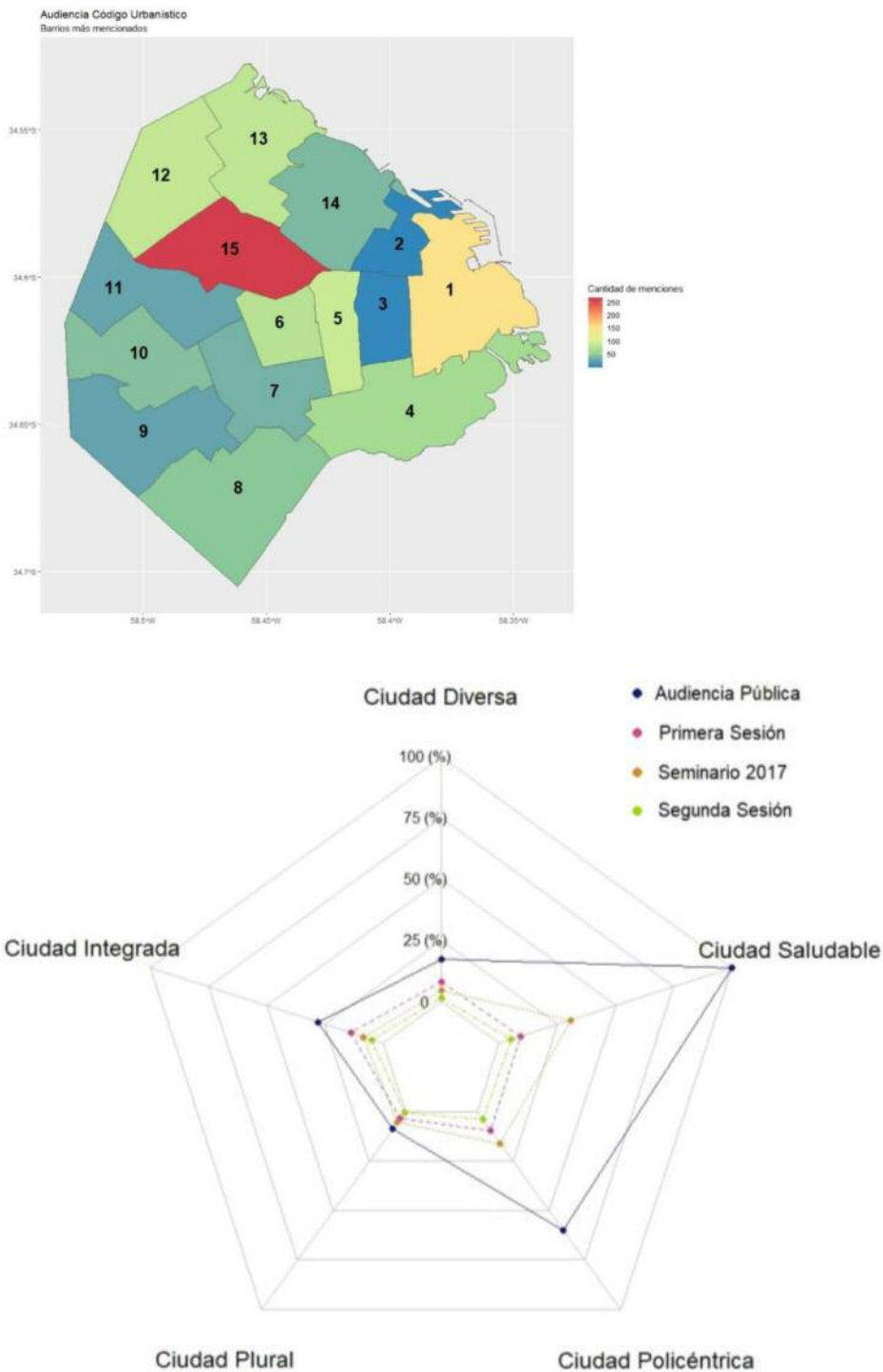
Estas controversias se vieron reflejadas en la aprobación del proyecto, ya que sólo fue votado positivamente por el bloque oficialista (Vamos Juntos). En el presente trabajo, realizamos un seguimiento de la trayectoria del proyecto para conocer sus particularidades, defensas y críticas.

Este trabajo se inscribe en las iniciativas de Legislatura Abierta promovidas en la Ciudad de Buenos Aires, con el objetivo de publicar datos parlamentarios y sumarse a las propuestas de Gobierno Abierto que se han venido desarrollan-



¹ Obtuvimos la versión taquigráfica de la sesión del sitio web de la Legislatura <https://www.legislatura.gov.ar/seccion/versiones-taquigraficas.html>

do en los últimos años. Este enfoque es importante para que se conozcan los trabajos y las dinámicas del lugar donde se discuten, elaboran y aprueban -o rechazan- las leyes.



Evaluation of the transparency in the plenums minutes of the Water Basins in the Sao Paulo Macrometropolis (Brazil)

Abstract¹

The Sao Paulo Macrometropolis (SPMM) is the most important urban agglomeration in Brazil, counting more than 170 municipalities (including the Sao Paulo Metropolitan Region) and more than 33.6 million inhabitants (EMPLASA 2017). Managing the water resources within this region is considered a quite complex issue (DAEE 2013).

The Brazilian National Water Law was created in 1997 (BRASIL 1997), incorporating modern water resources management principles and institutionalizing "Water Basin Committees" (WBC), which includes civil society actors on decision making. Also, the State of Sao Paulo Water Law (Sao Paulo State 1991) separates the state in 22 Water Resource Management Units (WRMU, or UGRHI in the original). Eight WRMU are completely or partially contained in the area of the SPMM: Paraíba do Sul, Litoral Norte, Piracicaba/Capivari/Jundiaí, Alto Tietê, Baixada Santista, Mogi-Guaçu, Tietê/Sorocaba, and Ribeira de Iguape/Litoral Sul.

Each WRMU have a corresponding WBC, and the WBC discuss and make decisions in plenums. The Sao Paulo State Water Law determined that these WBC plenums must be public (Sao Paulo State 1991). Therefore, the minutes of the plenums should be available to the general public. The Integrated Water Resources Management System (IWRMS, or SigRH in the original), makes available information about the Water Resources Management in the State of São Paulo publicly available through its website, including information about each WRMU and corresponding WBC.

The book Principles of Open Government Data (Tauberer 2014) presents criterias that can be used to evaluate if the public data can be considered "opened". Taking the importance of the transparency of information on water resources management into account, the aim of this research is to evaluate if the plenums minutes of the WBC within the SPMM are made publicly available on the website of the SigRH² (Sao Paulo State 2020). In order to evaluate whether the plenums minutes were available on the website, a technique called Web Scraping was used to collect data from websites, using the programming language R (2019). The packages used in all stages of this research was: rvest, purrr, dplyr, tibble, stringr, magrittr, glue, ggplot2 (which are part of the tidyverse (Wickham et al. 2019)), httr, sf, ggspatial. The authors developed functions in order to collect data from the webpages of the each WRMU. With these functions, a tibble was created with the following information: (a) date of data collection, (b) information about the plenum (which WBC, year, name and date) and (c) information about the plenums minute (url of the link, format of file available and status code). The data presented was collected in September 2020.

Figure 1 presents a map with the counting of plenums minutes accessible through the website for each WRMU. The WRMU Paraíba do Sul did not made available any of the plenums minutes, what represents an alert of a lack of transparency. The file format in which a plenums minute is made available is important to evaluate the "Non-proprietary" and "Machine Processable"

¹ Acknowledgments: This research is funded by São Paulo Research Foundation (FAPESP) (process n. 2018/23771-6). This work is part of the activities of the thematic project "Environmental governance in the Sao Paulo Macrometropolis in the context of climate variability" (process n. 2015/03804-9), financed by FAPESP and linked to the FAPESP Research Program on Global Climate Change.

² Portal SigRH - <http://www.sigrh.sp.gov.br/>

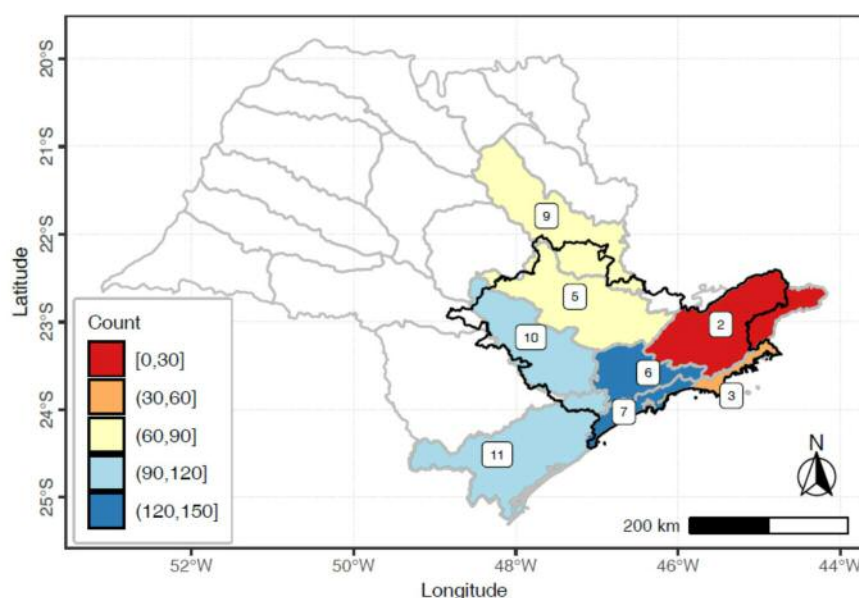


Figure 1: Map of the State of São Paulo and the WRMU that are part of the SPMM. Source: Elaborated by the authors based on the delimitation of the SPMM by DAEE and the packages ggplot2, message=FALSE, warning=FALSE, ggspatial, and sf. Shapefiles of the WRMU by DATAGEO. Shapefiles of the SPMM by LaPlan-UFABC, 2018.

WRMU: 2 – Paraíba do Sul; 3 – Litoral Norte; 5 – Piracicaba/Capivari/Jundiaí; 6 – Alto Tietê; 7 – Baixada Santista; 9 – Mogi-Guaçu; 10 – Tietê/Sorocaba; 11 – Ribeira de Iguape/Litoral Sul

principles (Tauberer 2014). The most common file formats used were PDF (39.9 %), followed by Microsoft Word Document (.doc and .docx) (28.3 %), webpages (.html and .htm) (17.2%), images (.jpg) (13.4 %) and other file formats (1.2 %). Scanned image files are not considered a very suited file format to be processable by machines (Tauberer 2014). Also, 3.8% of the links to the files of the plenums minutes were broken (not available).

The next step is to evaluate more deeply these data based on the Principles of Open Government Data (Tauberer 2014). Moreover, it is important to perform the scrape periodically and to monitor the resulting data over time. Also, another future aim is to expand the data scraping to contemplate other possibilities, such as the information on the composition of the WBC.

References

- BRASIL. 1997. "LEI N. 9.433, DE 8 DE JANEIRO DE 1997." http://www.planalto.gov.br/ccivil_03/Leis/L9433.htm.
- DAEE. 2013. "Macrometrópole - Sumário Executivo - Plano Diretor de Aproveitamento de Recursos Hídricos Para a Macrometrópole Paulista."

http://www.dae.sp.gov.br/index.php?option=com_content&view=article&id=1112:plano-diretor-de-aproveitamento-dos-recursos-hidricos-para-a-macrometropole-paulista.

- EMPLASA. 2017. "Macrometrópole Paulista -." EMPLASA. <https://www.emplasa.sp.gov.br/MMP>.
- R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- São Paulo State. 2020. "Portal SIGRH - Sistema Integrado de Gerenciamento de Recursos Hídricos Do Estado de São Paulo." <http://www.sigrh.sp.gov.br/>.
- —. 1991. "Lei N. 7.663, de 30 de Dezembro de 1991." <http://www.al.sp.gov.br/leis/legislacao-do-estado/>.
- Tauberer, Joshua. 2014. Open Government Data: The Book. 2nd ed. <https://opengovdata.io/2014/8-principles/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.

Sesión **Desarrollo de paquetes**

latinr en LatinR: automatizando el envío de trabajos a conferencias de R con un paquete de R

Una de las grandes ventajas de R es knitr (Xie 2014) y el ecosistema de paquetes al rededor de R Markdown (Xie, Allaire, and Golemund 2018). La posibilidad de tener el código que genera los resultados en el mismo documento que la prosa que los describe implica que los usuarios no necesitan copiar y pegar tablas y figuras generadas manualmente. Esto reduce errores y facilita la reproducibilidad de los reportes. Actualmente existen numerosos paquetes que permiten correr código de R y generar documentos en una infinidad de formatos, desde sitios web y blogs (Xie, Hill, and Thomas 2017) hasta libros (Xie 2016). Roger Peng (Peng 2017) considera que la facilidad de escribir documentos reproducibles usando knitr y R Markdown es la primera de las 4 características que destacan a R como lenguaje y ecosistema.

Pero para muchos formatos, la vida del docu-

mento recién comienza con su creación y a veces el proceso de publicación es dolorosamente manual. En particular, enviar trabajos a congresos y conferencias suele implicar llenar formularios con, en la gran mayoría de los casos, información redundante. ¿Por qué el formulario me exige escribir el título de mi resumen si éste ya está en el propio resumen?

El paquete latinr busca automatizar este proceso lo más posible. Permite escribir el resumen enteramente en R Markdown y enviarlo sin salir de la sesión de R. Todos los datos se ingresan una sola vez en el encabezado YAML y son usados tanto para la generación del PDF como para llenar el formulario de inscripción automáticamente. El funcionamiento interno está fuertemente inspirado en el paquete rticles (Allaire et al. 2020), el cual provee plantillas para generar artículos listos para ser presentados a distintas revistas científicas.

Figura 1: El único formulario que hay que llenar

ficas usando únicamente R Markdown (es decir, sin tener que saber nada de LATEX).

El proceso empieza creando un documento con la planilla provista por el paquete. El paquete provee una interfaz gráfica construida con shiny (Chang et al. 2020; Cheng 2019) que permite llenar los datos del trabajo a ser enviado (Figura 1) y que se invoca con `latinr::latinr_wizard()`. Esta misma interfaz chequea y avisa si hay errores, como datos faltantes o mala cantidad de palabras clave. Alternativamente, se puede crear la plantilla base desde una ventana de RStudio yendo a Archivo Nuevo Archivo R Markdown Desde Plantilla y eligiendo “LatinR submission article”.

Luego de creado el archivo base, se escribe el resumen usando R Markdown.

Cuando el documento está listo, se envía usando `latinr::latinr_submit()`. El comando automáticamente chequea que los metadatos no tengan errores y envía el PDF luego de una inspección final. Por defecto, renderiza el archivo de R Markdown usando la plantilla anonimizada, lo que garantiza que el PDF enviado sea reproducible (al menos en la máquina local) y que se corresponda con los metadatos.

Una preocupación importante fue la de permitir el acceso a personas que no quieran o no sepan usar R Markdown y personas que prefieran hacer el envío con una interfaz gráfica en la web. Como respuesta a esto, el paquete puede usarse para generar el PDF únicamente, o para enviar un PDF previamente creado con cualquier otra herramienta.

Desarrollos futuros

Además de pulir errores y problemas varios de usabilidad (por ejemplo, chequear automáticamente que el PDF cumpla con el límite de páginas), el futuro de `latinr` es adaptarse a otras plataformas. Por ahora el envío de trabajos se hace a través de la plataforma EasyChair y esa conexión es inestable ya que depende de que se mantenga la implementación interna de su formulario web.

La visión es extender `latinr` a herramientas para la gestión de la conferencia. Es decir, recepción de trabajos, distribución a los evaluadores, recepción de evaluaciones y la creación semiautomática de cronogramas. En lo posible haciendo uso de servicios abiertos, gratuitos o de bajo costo.

Referencias

- Allaire, JJ, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, et al. 2020. R Markdown: Article Formats for R Markdown. <https://CRAN.R-project.org/package=rmarkdown>.
- Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. Shiny: Web Application Framework for R. <https://CRAN.R-project.org/package=shiny>.
- Cheng, Joe. 2019. MiniUI: Shiny UI Widgets for Small Screens.
- Peng, Roger. 2017. “How Do You Convince Other People to Use R?” 2017. <https://web.archive.org/web/20190329233513/https://simplystatistics.org/2017/10/30/how-do-you-convince-others-to-use-r/>.
- Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. <http://www.crcpress.com/product/isbn/9781466561595>.
- ———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. <https://github.com/rstudio/bookdown>.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.
- Xie, Yihui, Alison Presmanes Hill, and Amber Thomas. 2017. *Blogdown: Creating Websites with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. <https://github.com/rstudio/blogdown>.

geouy: acceso a las geometrías de Uruguay

Introducción

La librería `geouy` fue creada para permitir a los usuarios de R tener un acceso sencillo a los servicios geográficos oficiales de Uruguay. Este paquete incluye un amplio rango de capas geográficas en formato simple features (`sf`) (E. Pebesma 2018), como ser las unidades estadísticas del Instituto Nacional de Estadística (INE), unidades administrativas (departamentos, barrios, etc.), rutas, peajes, instituciones deportivas, o incluso las grillas de las ortofotos de los vuelos de la Infraestructura de Datos Espaciales (IDE). Se busca ofrecerlas en sus diferentes escalas y variaciones en el tiempo, armonizando sus principales atributos y proyecciones, pero sin dejar de ser fieles a los datos del servicio oficial correspondiente.

El paquete `geouy` surge en el marco de la conjugación de proyectos locales, pero tratando de compatibilizarlo con otras herramientas similares que han aparecido en la región. Este proyecto, comenzó como parte de la librería `ech` en la que continuamos trabajando con Gabriela Mathieu (@calcita), y algunos paquetes de servicios geoespaciales de la región, principalmente: `geobr` y `chilemapas`.

Herramientas

A continuación presentaremos las principales funciones que brinda esta librería.

`load_geouy()`

Esta función permite cargar diversas geometrías desde los servicios oficiales que publican la información. Al ser una librería abierta permite que cualquier usuario complemente los servicios a

medida que vayan siendo requeridos.

Para simplificar la aproximación a la información geográfica a todos los usuarios esta función por defecto traerá estas geometrías en EPSG 32721 que es el Sistema de Coordenadas de Referencia (CRS) correspondiente a Uruguay, pero presenta algunas funciones complementarias para verificar si la capa geográfica con la que se trabaja corresponde a alguna de los CRS.

`which_uy()`

Es un complemento de la función anterior, que permite ahorrar la descarga y unión de atributos con geometrías oficiales para asignar códigos y nombres a las geometrías.

`geocode_ide_uy()`

Esta función permite la geocodificación de casos a partir de direcciones de Uruguay sin coordenadas. Crea la consulta para el servicio de geocodificación de la IDEuy y obtiene el par de coordenadas correspondientes.

`tiles_ide_uy()`

Finalmente presentamos esta función, que se encarga de la descarga de las ortofotos del último vuelo relevado por IDEuy, permitiendo que lo haga en formato `.jpg` (con su correspondiente archivo `.jgw`) o en formato `.tif`, desde el repositorio oficial, recortándolo a la extensión (bounding-box) de un objeto `'sf'`.

Reflexiones

Este camino sobre hombros de gigantes¹, ha permitido a esta librería focalizarse en un peque-

¹ Henry and Wickham (2019); R Core Team (2017); Wickham, Hester, and Ooms (2018); Wickham et al. (2019); J. Hester (2019); Wickham (2019); Bache and Wickham (2014); E. Pebesma (2018); Hijmans (2019); Wickham (2016); Arnold (2019); Santos Baquero (2019); J. Hester and Wickham (2019); E. J. Pebesma and Bivand (2005); Bivand, Pebesma, and Gomez-Rubio (2013)

ño país como Uruguay, que con sus propias particularidades trata de ajustarse a su vez a geobr en su estructura, pero sin perder de vista su origen como complemento de ech.

Se busca seguir creciendo, y se propone incorporar cualquier función de propósitos generales que utilice como base los datos geográficos de Uruguay. Todos los aportes en este sentido son bienvenidos. Y si trabajas con datos geográficos de Uruguay y querés agregar tu función o mas datos, te recomendamos que leas los siguientes consejos de como colaborar.

Referencias

- Arnold, Jeffrey B. 2019. Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'. <https://CRAN.R-project.org/package=ggthemes>.
 - Bache, Stefan Milton, and Hadley Wickham. 2014. Magrittr: A Forward-Pipe Operator for R. <https://CRAN.Rproject.org/package=magrittr>.
 - Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. Applied Spatial Data Analysis with R, Second Edition. Springer, NY. <http://www.asdar-book.org/>.
 - Henry, Lionel, and Hadley Wickham. 2019. Rlang: Functions for Base Types and Core R and 'Tidyverse' Features. <https://CRAN.R-project.org/package=rlang>.
 - Hester, Jim. 2019. Glue: Interpreted String Literals. <https://CRAN.R-project.org/package=glue>.
 - Hester, Jim, and Hadley Wickham. 2019. Fs: Cross-Platform File System Operations Based on 'Libuv'. <https://CRAN.R-project.org/package=fs>.
 - Hijmans, Robert J. 2019. Raster: Geographic Data Analysis and Modeling. <https://CRAN.R-project.org/package=raster>.
 - Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." The R Journal. <https://journal.r-project.org/archive/2018/RJ-2018-009/index.html>.
 - Pebesma, Edzer J., and Roger S. Bivand. 2005. "Classes and Methods for Spatial Data in R." R News 5 (2): 9–13. <https://CRAN.R-project.org/doc/Rnews/>.
 - R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
 - Santos Baquero, Oswaldo. 2019. Ggsn: North Symbols and Scale Bars for Maps Created with 'Ggplot2' or 'Ggmap'. <https://CRAN.R-project.org/package=ggsn>.
 - 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>.
 - Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2019. Dplyr: A Grammar of Data Manipulation. <https://CRAN.R-project.org/package=>
 - Wickham, Hadley, James Hester, and Jeroen Ooms. 2018. Xml2: Parse Xml. <https://CRAN.R-project.org/package=xml2>.
-

Una librería para procesar datos públicos.

Presentación del paquete eph

Introducción

El trabajo con datos producidos por fuentes públicas suele encontrarse con varios problemas: uno de los más comunes es la falta de continuidad en la publicación de bases de datos. En ese sentido, la Encuesta Permanente de Hogares -EPH- del Instituto Nacional de Estadística y Censos (INDEC) en Argentina constituye una excepción. En efecto, se trata de un programa que ha publicado bases de datos usuarias con información desde 1974. No obstante, esto se ha hecho de forma “poco replicable”: desde cambios en los formatos de su publicación (dbase, .txt, .xls, .sav, etc.) hasta renombrado de algunas variables y recodificaciones de sus categorías que los hacen poco prácticos para su uso y procesamiento continuo. La inexistencia de una API para la divulgación de la información producida por INDEC genera un límite a las capacidades de procesamiento de la información, reduciendo el público usuario a i) expertos temáticos con conocimiento respecto a cómo acceder a las fuentes y ii) medios de comunicación que acceden a la información ya procesada bajo la forma de comunicados. Esto limita el potencial valor del enorme trabajo elaborado en el instituto, al desincentivar el uso de usuarios con conocimientos limitados de las fuentes, pero con capacidades de procesamiento de datos, como es la comunidad de usuaries de R. A su vez, ciertos indicadores clave presentados por la EPH cuentan con anexos metodológicos, pero no con implementaciones públicas que le permitan al público hacer uso de la metodología por fuera de los informes elaborados por el instituto.

En el marco del presente proyecto se creó el repositorio HOLATAM que busca darle un marco regional. Allí puede encontrarse una herramienta similar realizada para la encuesta Encuesta de

Caracterización Socioeconómica Nacional (CASEN) de Chile. A su vez, la Encuesta Continua de Hogares de Uruguay (ECH) cuenta con un paquete ech que ha replicado la estructura de alguna de las funciones del paquete eph.

Descripción general de la librería

En este contexto, la librería eph tiene como objetivo facilitar el trabajo de aquellos usuaries de la Encuesta Permanente de Hogares - INDEC que deseen procesar datos de la misma mediante el lenguaje de programación R. La librería presenta las siguientes funcionalidades:

- una sintaxis unificada para la descarga, etiquetado y construcción de datasets con información crosssectional comparables
- una implementación del cálculo de indicadores (pobreza) utilizando la metodología oficial

Algunas de sus funciones son:

- `get_microdata()`: Descarga las bases de microdatos
- `organize_panels()`: Permite armar un pool de datos en panel de la EPH continua,
- `organize_cno()`: Clasifica las ocupaciones según el CNO 2001
- `'organize_caes()`: Clasifica las actividades económicas según CAES Mercosur 1.0 y CAES Mercosur
- `organize_labels()`: Etiqueta las bases siguiendo el último diseño de registro
- `map_agglomerates()`: Mapa de indicadores por aglomerado

El paquete también cuenta con otros sets de datos que pueden ser útiles para el trabajo con la EPH: algunos diccionarios que contienen la codificación de las variables geográficas (como regiones

o aglomerados) o la posición geográfica (centroídes) de los aglomerados en que se releva la encuesta.

Formato de la presentación

- Presentación de los objetivos de la librería (2 min)
 - Mención de algunas de sus funciones más importantes (3 min)
 - Desarrollo de un workflow de ejemplo (10 min)
-

{polAr} Política Argentina Usando R

Abstract

polAr es un paquete pensado para facilitar el flujo de trabajo y el acceso a datos políticos de Argentina. El mismo permite trabajar con resultados electorales, votaciones legislativas o discursos presidenciales. Entre otros, está inspirado en los paquetes eph (Kozłowski et al. 2019) - que facilita el acceso a datos de la Encuesta Permanente de Hogares del Instituto Nacional de Estadísticas y Censos de Argentina- y esaps (Schmidt 2018) -que provee métodos para el cómputo de indicadores de sistemas de partidos y electorales.

1. Datos

Aunque polAr no es un paquete de datos, una de sus principales funciones es facilitar el acceso a los mismos haciendo llamadas a un repositorio independiente¹. El código y diseño de eph fueron centrales para desarrollar esta parte. El primer paso consiste en la curaduría de datos. Así, por ejemplo, para el flujo de trabajo con información electoral, se dio un nuevo formato a los datos partiendo de las fuentes originales y se diseñó una estructura de archivos que nos permitiera consultar resultados de elecciones de un modo sencillo. Partiendo de esos datos se diseñaron funciones como `show_available_elections` (que devuelve una tabla que funciona como índice de elecciones disponibles²) y `get_election_data` (que descarga información).

Flujos similares se diseñaron para cada tópico: con `show_available_bills` y `get_bill_votes` para datos legislativos; y `show_available_speech` y

`get_speech` para discursos presidenciales.

La idea general es disponibilizar la información lo más desagregada y limpia posible para usar del modo más conveniente por cada usuario. Es por ello que se agregan opciones para obtener la data cruda (usando el parámetro `raw = TRUE`).

2. Indicadores

Otra posibilidad es la de calcular indicadores a partir de la información obtenida. Por ejemplo, es posible computar la Competitividad electoral (`compute_competitiveness`); el Número Efectivo de Partidos (`compute_nep`); el reparto de escaños (`compute_seats`); el nivel de Concentración (`compute_concentration`) y la Desproporción electoral (`compute_disproportion`).

3. Visualización

Por último, también podemos hacer uso de funciones que permiten visualizar rápidamente de manera exploratoria los datos. Así, por ejemplo, con `plot_speech()` se puede visualizar la frecuencia de palabras de un discurso presidencial seleccionado en una nube interactiva; con `plot_bill()` la distribución de votos de un proyecto de ley (afirmativo, negativo, nulo o abstención); y, en el caso de datos de elecciones, explorar los resultados electorales mediante tablas (`tabulate_results`), gráficos (`plot_results`) o mapas (`map_results`).

Agradecimientos: polAr (Ruiz Nicolini 2020) contó con la colaboración de Camila Higa, Lucas Enrich e Iván Lewin.

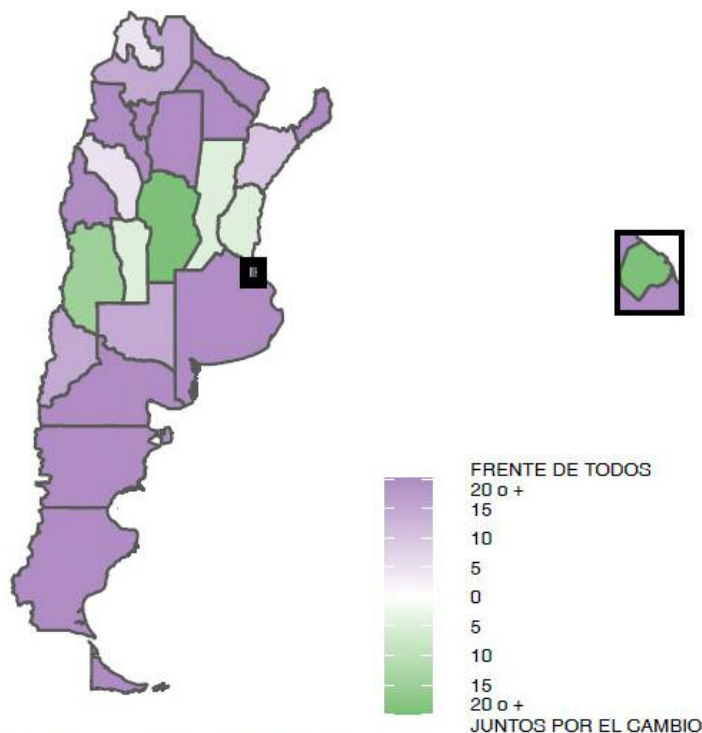
1 El paquete contiene cuatro flujos de trabajos: (i) capas de información geográfica y codificaciones; (ii) discursos presidenciales para el análisis cuantitativo de texto; (iii) registros de votaciones legislativas y (iv) datos de resultados electorales.

2 La fuente original de datos para resultados de elecciones nacionales (2003-2017) provienen del Atlas Electoral de Tow (2020). Los datos de las elecciones de 2019 tienen una estructura diferente de las de años anteriores y fueron reconstruidos de unos paquetes específicos desarrollados por Moracho (2020) para cada turno: P.A.S.O. y Generales. Más detalles disponibles en el repositorio de datos: https://github.com/electorArg/PolAr_Data/.

```
library(polAr)
get_election_data(district = "arg", category = "presi",
  round = "gral", year = 2019) %>%
  map_results()
```

Elección a Presidente de la Nación – General 2019

Puntos Porcentuales de Diferencia



ente: polAr – Política Argentina usando R – <https://electorarg.github.io/polAr>

Referencias

- Kozlowski, Diego, Pablo Tiscornia, GuidoWeksler, Natsumi Shokida, and German Rosati. 2019. Eph: Argentina's Permanent Household Survey Data and Manipulation Utilities. <https://CRAN.R-project.org/package=eph>.
- Moracho, Patricio. 2020. Elecciones.ar.2019: Elecciones Nacionales de Argentina 2019. <http://github.com/pmoracho/elecciones.ar.2019>.
- Ruiz Nicolini, Juan Pablo. 2020. PolAr: Argentina Political Analysis. <https://CRAN.R-project.org/package=polAr>.
- Schmidt, Nicolas. 2018. Esaps: Indicators of Electoral Systems and Party Systems. <https://CRAN.R-project.org/package=esaps>.
- Tow, Andy. 2020. Atlas Electoral. <https://www.andy-tow.com/>.

desuctools: paquete para para análisis de encuestas sociales

La Dirección de Estudios Sociales (DESUC) es una unidad de estudios y servicios profesionales del Instituto de Sociología de la Pontificia Universidad Católica de Chile que trabaja en diversas áreas y materias sociales, de opinión pública, educación, comunicaciones, estudios organizacionales, entre otros temas. Para ello se ha especializado en el levantamiento y análisis de datos cualitativos y cuantitativos. En el ámbito cuantitativo realiza levantamiento de datos presenciales, telefónicos en un call center propio y por Internet.

Contexto

Para el seguimiento, validación, análisis y comunicación de datos DESUC ha ocupado históricamente procesos y herramientas separadas, muchas veces manuales, las cuales podrían aumentar las posibilidades de cometer errores en el proceso. Otras implicancias de este tipo de prácticas fueron señaladas en la charla Reproducible Analysis in the UK Government presentada en LatinR 2019. En particular, en el trabajo en DESUC destacamos tres problemas vinculados a ello:

- El creciente interés por análisis estadísticos sofisticados hacía más habitual el uso de SPSS para obtener frecuencias, y Stata para análisis de regresión y modelamientos. Esto podía generar bases de datos duplicadas con posibles imputaciones y codificaciones de variables divergentes. Por otro lado, para la generación de reportes el uso de Excel era habitual. Debido a ello, para supervisar y apoyar en estas diferentes tareas —además de orden— se necesitaba conocimiento de tres softwares distintos.
- Lo anterior repercute en la velocidad con la que se puede procesar y analizar la información recolectada, sobre todo cuando existen

procesos iterativos de revisión y ajuste de los datos.

- El principal efecto negativo de lo anterior es la lentitud con la que pueden plasmarse innovaciones y mejoras de procesos logrados en un estudio, en un proyecto futuro. Dado esto, es difícil tanto realizar procesos de análisis reproducibles, como también generar un toolbox en donde queden almacenadas ideas que puedan ser reutilizadas posteriormente en otros contextos.

Solución

Para acumular, potenciar y hacer disponible el conocimiento acumulado a lo largo de los distintos proyectos que toca conducir, se decidió crear el paquete de R *desuctools*, el cual contiene funciones o datos que son utilizados habitualmente para la selección de muestras, análisis y visualización de datos de encuestas sociales de la manera idiosincrática en que DESUC lo hace.

El desarrollo de esta herramienta, además de potenciar y mejorar los aspectos recién señalados, ha tenido efectos de segundo orden dentro de la organización que vale la pena destacar. Primero, ha inspirado a detectar patrones habituales del trabajo realizado, y de esta manera, analizar cuales de ellos puede ser paquetizado en una función. Segundo, mostró la necesidad de utilizar sistemas de desarrollo colaborativo de código y control de versiones —GitHub en nuestro caso. Y tercero, permitió seleccionar y unificar los paquetes que habitualmente utilizamos internamente para analizar y visualizar datos y encuestas.

Finalmente, *desuctools* es un paquete que está en continuo desarrollo, tanto mejorando las funciones que tiene, como incluyendo nuevas según las nuevas ideas o necesidades de los proyectos y el equipo.

Figura 1: Gráfico de ejemplo



```
suppressMessages(library(tidyverse))
# install_github('desuc/desuctools')

file <- tempfile()
download.file(url = 'https://github.com/DESUC/30diasdegraficos/raw/master/inputs/12-lollipop-df_bicen_19_30',
  destfile = file)

data <- readRDS(file) %>%
  mutate(across(c(t01_1:t01_2),
    ~desuctools::rec_cat_5a3(., labels = c('Bastante' = 1, 'Algo' = 2, 'Poco' = 3, 'NA/NR' = 9),
      total = TRUE, # Incluye dato total
      .vars = vars(t01_1, t01_2), # Listado de variables de interés
      .segmentos = vars(d07), # Listado de segmentos de interés
      .wt = pond_se) %>% # Ponderador
    mutate(pregunta_lab = desuctools::str_entre_parentesis(pregunta_lab))

desuctools::gg_bar_3_niveles_stack(
  .df = data,
  x = segmento_cat,
  facet_col = pregunta_lab,
  missing = 'NA/NR', y_na = 1.1, x_na = -2.5,
  title = '¿Cuánto temor le producen las siguientes situaciones?',
  font_family = '')
```

Función destacada

Para esta presentación destacaremos la función `tabla_vars_segmentos`, la cual permite generar un `data.frame` tidy con la cantidad y proporción de respuestas para las categorías de un número arbitrario de preguntas para un número arbitrario de segmentos de la población. Las respuestas podrán ser ponderados según una variable auxiliar adicional, además de calcularse porcentajes válidos en el caso que se quiera establecer una o más categorías como missing.

Adicionalmente, se generan columnas adicionales con etiquetas de variables y categorías para obtener una tabla en el formato adecuado para facilitar la generación de tablas y gráficos de resultados.

En el proceso de creación de la figura, se utilizan también funciones adicionales: `desuctools::rec_cat_5a3`, `desuctools::gg_bar_3_niveles_stack` y `desuctools::str_entre_parentesis`.

Sesión **Shiny y visualización de datos**

Designing beautiful and informative visualizations: the case for {ggplot} and {highcharter}

Abstract

Both Highcharter and ggplot provide functionalities to create beautiful visualizations. However, not all developers are familiar with the do's and don'ts of styling their graphics. Design decisions are often made out of instinct and without the necessary knowledge to produce a good balance between usability and beauty. This presentation will describe the key aspects of designing charts in Highcharter and ggplot so as to take full advantage of these libraries' styling features, including color, font and size options.

We often worry more about the type of graphic or the quality of the data than the colors we will represent it with. However, defining a color palette is one of the most advantageous parts of

designing visualizations. A suitable color scheme is critical to ensure proper accessibility and to emphasize certain elements of our graphics. Combining this with an appropriate font selection will help us adequately craft the message we want to send to our audience about our project, our brand or even our data. Considering the possibilities of the Highcharts and ggplot libraries, we will go through the features that will help us create beautiful and usable visualizations, while learning about color and font accessibility. We will also show some live applications of our recommendations to better show the impact of these design decisions.

Seguimiento diario de la producción científica sobre COVID-19

Introducción

En diciembre de 2019 se registró en Wuhan, China, el primer paciente afectado por una neumonía atípica, el 7 de enero de 2020 las autoridades chinas confirmaron que correspondía a un nuevo virus, el 2019-ncov. Tan solo diez días después, apareció la primera publicación científica sobre el tema en la base de datos bibliométrica PubMed¹. Si bien al principio, las publicaciones se concentraron en unos pocos países, al hacerse más claro el peligro de un brote global, la comunidad científica mundial comenzó a trabajar en el tema. El 30 de enero se definió al 2019-ncov como un problema de salud pública, en ese momento ya existían 29 publicaciones firmadas por 20 países. El 11 de marzo se lo clasificó como pandemia, para ese entonces ya había 777 publicaciones sobre el tema, al 30 de agosto existen más de 45 mil publicaciones y 188 países trabajando sobre el tema.

Objetivos

La velocidad con que la temática se expande dentro del campo de la ciencia requiere de un análisis dinámico, por eso, se diseñará un tablero que permita realizar un seguimiento diario de la producción científica sobre covid-19: Conocer su magnitud, evolución, países involucrados, colaboración conjunta y las líneas de investigación.

Materiales y Métodos

Se utilizaron las publicaciones científicas sobre Covid-19 extraídas de la base de datos PubMed². La descarga y procesamiento se realizó a través de la librería RisMed de R. Se desarrolló un código que recopila los datos de forma diaria y los incluye en una base de datos acumulada, de forma automática.

El campo que presentó mayor dificultad para procesar fue el país: este dato se extrae de la afi-

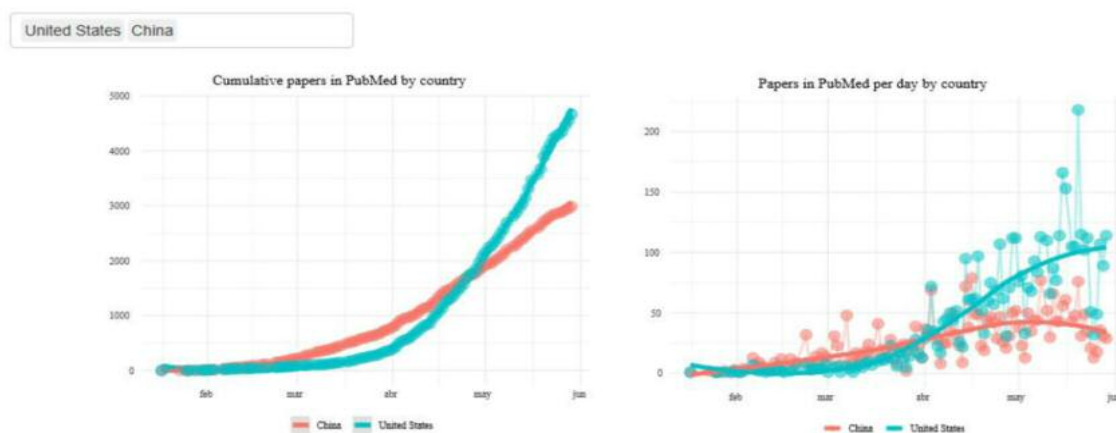


Figura 1: Producción científica sobre COVID-19 por país

1 <https://pubmed.ncbi.nlm.nih.gov/>

2 Estrategia de Búsqueda: "COVID-19"[All Fields] OR "severe acute respiratory syndrome coronavirus 2"[Supplementary Concept] OR "severe acute respiratory syndrome coronavirus 2"[All Fields] OR "2019-nCoV"[All Fields] OR "SARS-CoV-2"[All Fields] OR "2019nCoV"[All Fields] OR (("Wuhan"[All Fields] AND ("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields])) AND 2019/12[PDAT] : 2030[PDAT])



Figura 2: Redes de Colaboración Internacional sobre COVID-19

liación institucional de cada autor; usualmente el texto después de la última coma de cada firma es el país, para los casos donde esto no funcionó se identificaron una serie de expresiones regulares para extraerlo correctamente. Semanalmente se revisa el proceso y se incluyen nuevas expresiones regulares, de ser necesarias.

Resultados

El tablero diseñado con Shiny consta de 5 solapas, cada una, además de las visualizaciones correspondientes, brinda la posibilidad de descargar los datos para realizar análisis propios. La primera solapa muestra totales diarios y acumulados, la segunda solapa permite comparar en forma dinámica la evolución de los distintos países (Figura 1).

La tercera solapa presenta un mapa que permite ver las redes de colaboración³ entre los países (Figura 2), la cuarta solapa presenta un mapa conceptual sobre el contenido de las publicaciones realizado a través de los conceptos más relevantes del resumen de las publicaciones, la quinta solapa es un navegador que permite filtrar las publicaciones por país, acceder a cada una o descargarlas.

El tablero se encuentra disponible en la siguiente dirección: <https://observatorio-cts.shinyapps.io/Covid-19/>. Las bases y el código utilizado se encuentran disponibles en el siguiente repositorio: <https://github.com/juansokil/Covid-19>

³ Se entiende por colaboración al trabajo conjunto de dos (o más autores) de distintos países (definido a partir de su afiliación institucional)

Creando componentes de javascript personalizados para Shiny: Cómo se creó el paquete {shinyinvoer}

Abstract

El nuevo paquete {shinyinvoer} agrega nuevos componentes funcionales a shiny, entre ellos un selector de paletas de color que usa una librería de javascript llamada spectrum, otro que es una caja de búsqueda “predictiva” que permite a los usuarios buscar en listas con miles de registros y otro para personalizar grupos de botones con imágenes. En esta charla se mostrará un paso a paso

para crear componentes de Shiny personalizados con javascript con sus respectivos bindings y opciones, incluyendo cómo enviar información desde R a javascript para actualizar los inputs y cómo recibir información desde javascript en R cuando se usan los componentes en una aplicación de Shiny.

Cómo usar R con tidyverse para la lucha anti-corrupción

Abstract

Esta charla presenta diferentes usos de R en periodismo de datos para la lucha anti-corrupción en América Latina. Partimos de un caso puntual en Colombia donde se muestra cómo se acceden datos abiertos, se filtran y analizan los datos de contrataciones públicas por medio de las librerías del tidyverse con el fin de encontrar particularidades, patrones e irregularidades en los procesos de compras públicas. Además se brindan elementos para presentar los hallazgos como reportes con

análisis de información a través de visualizaciones estáticas con ggplot y dinámicas con htmlwidgets con el fin de difundir los resultados a través de páginas web con enfoque de periodismo de datos utilizando un tema personalizado de Hugo con Blogdown y aplicaciones de Shiny para explorar redes de financiamiento de campañas políticas en el proyecto Monitor Ciudadano de la Corrupción <https://www.monitorciudadano.co/elecciones-contratos/campanas>

#TuitómetroNacional: monitor de la conversación política en Argentina

Abstract

Twitter es la plataforma preferida para el análisis de datos políticos en redes. Hay por lo menos dos razones que lo explican: (a) es donde la mayoría de los dirigentes políticos se expresan (y, por ende, donde el público que consume información política interactúa); y (b) es la red que disponibiliza grandes volúmenes de información.

En la actualidad podemos encontrar varias librerías de distintos lenguajes que facilitan la recopilación de estos datos. Para fines de investigación encontramos dos que pueden ser ejecutadas desde R que se destacan sobre el resto. Una primera opción es trabajar con twint¹ (programa escrito en python que puede correr en R a través de reticulate). Una segunda alternativa, plenamente integrada con la API de la plataforma de microblogging es rtweet (Kearney 2019).

Partiendo de esto desarrollamos una aplicación shiny que nos permite analizar de manera agregada a través de determinados indicadores y visualizaciones las cuentas de las personas que integran el ecosistema político e institucional de Argentina: monitorear la actividad de las cuentas, cómo interactúa el público con la información que publican y las relaciones entre legisladores/as nacionales, gobernadores/as provinciales y miembros del gabinete nacional en Twitter². La herramienta se divide en tres secciones:

1. Métricas: permiten explorar qué y cómo publican contenido las cuentas seleccionadas. Al seleccionar categoría + nombre se obtienen métricas individuales de cada una: mejores publicaciones (considerando frecuencia de likes y RT de

cada publicación) y ranking de usuarios mencionados y hashtags utilizados³.

2. Usuarios en Red: Ernesto Calvo y Andrés Malamud (2014) sostienen que "los motivos para "seguir" o "ser seguido" por un político son variados, y van desde la afinidad partidaria, cultural o territorial hasta el espionaje, pasando por el consumo irónico. Sin embargo, la intuición sugiere que la afinidad prevalece y la evidencia lo confirma: los políticos siguen a más "amigos" que a rivales. Como la decisión de a quién seguir es pública, con ella envían señales de pertenencia y de reciprocidad tanto a los votantes como a sus potenciales aliados".

Siguiendo esa idea identificamos la totalidad de las cuentas seguidas por nuestros usuarios para luego poder agruparlos según "afinidad". Utilizando igraph creamos un grafo para reconstruir las conexiones entre las cuentas y, en base a éstas, a través del método de clusterización random walk identificamos los grupos de pertenencia para cada una de las categorías políticas (vg. Gabinete Nacional o Gobernadores).

A través de visNetwork logramos una visualización interactiva de las conexiones en la que al seleccionar un usuario se visualiza la red que refleja la relación seguidos/seguidores entre miembros de una misma categoría (al 9 de diciembre de 2019, día previo al inicio de una nueva gestión de gobierno en Argentina).

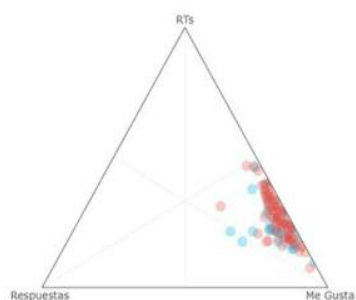
- Cada nodo (círculo) corresponde a la cuenta de un usuario/a y las aristas (flechas) marcan la dirección de quién sigue a quién.
- Los colores de los nodos son en base a los

1 Twint es un proyecto de Python que va más allá de los límites establecidos por la API de Twitter <https://github.com/twintproject>.

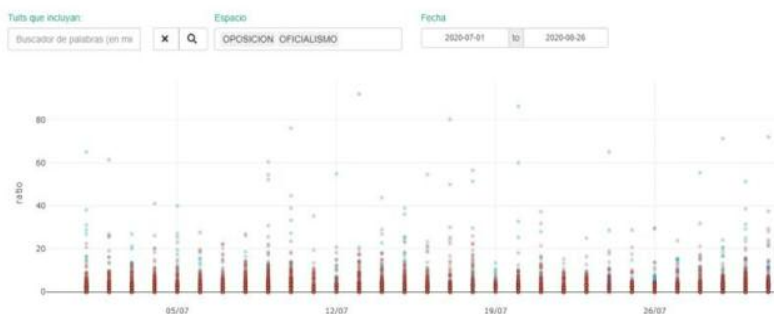
2 Shiny app en línea: <http://tuitometro.mentacomunicacion.com.ar/>

3 Esta sección está basada en el desarrollo de Garrik Aden-Buie <https://github.com/gadenbuie/tweets-of-2019>

Ternario Interactivo



Publicaciones en el tiempo



clusters obtenidos y los colores de las aristas heredan el color de los seguidores en su vínculo con los seguidos. Cuando un par de cuentas se sigue recíprocamente el color refleja una combinación de los dos.

3. Ratio: cuantifica la relación entre las distintas reacciones posibles ante un tuit (RT, Me Gusta, Fav) partiendo de la hipótesis que a mayor cantidad de respuestas en relación a las otras interacciones, el mensaje tuvo peor recepción en el público y está definido como
$$\text{RATIO} = \frac{\text{Respuestas}}{\text{Favs} + \text{RT}}$$

El foco de esta información está puesto en cómo interactúa el público con las publicaciones⁴.

En el Ternario Interactivo cada punto representa a una cuenta y permite visualizar el ratio típico de cada una de ellas. Los distintos ponderadores muestran los puntos en función del nivel de actividad (cantidad de publicaciones) y de rendimiento (promedio de interacciones). Además, se puede filtrar la información por género y espacios políticos (oficialismo y oposición) según el color.

Publicaciones en el tiempo muestra los tuits por día de emisión con su respectivo ratio. Adicionalmente se puede filtrar por palabras clave, por espacio y por fecha de la muestra. Finalmente, al clicar los puntos se accede a la publicación original.

Referencias

- Calvo, Ernesto, and Andrés Malamud. 2014. "Un cafe-cito para Scioli." *El Estadista*, March. <https://www.elestadista.com.ar/?p=4112>.
- Kearney, MichaelW. 2019. "Rtweet: Collecting and Analyzing Twitter Data." *Journal of Open Source Software* (42): 1829. <https://doi.org/10.21105/joss>.
- Pilorget, Juan Pablo, and Luciano Galup. 2018. "La Grieta Graficada." *Perfil*, October. <https://www.perfil.com/noticias/politica/la-grieta-graficada.phtml>.
- Roeder, Oliver, Dhruvil Mehta, and Gus Wezerek. 2017. "The Worst Tweeter in Politics Isn't Trump." *FiveThirtyEight*. October 24, 2017. <https://fivethirtyeight.com/features/the-worst-tweeter-in-politics-isnt-trump/>.

⁴ Este apartado está inspirado en el trabajo de Roeder y otros (2017) sobre el comportamiento de los usuarios con las publicaciones de políticos de EE.UU. en FiveThirtyEight y el posterior análisis sobre el caso argentino de Galup y Pilorget (2018).

Sesión
**Aplicaciones
con datos
educativos y
políticas
sociales**

Aplicaciones de R al estudio de enseñanza universitaria: el caso de la Facultad de Ingeniería, Universidad de la República

Introducción

La Facultad de Ingeniería (FIng) es una de las Facultades de la Universidad de la República (Udelar), ofrece a la fecha 48 carreras entre grado, técnicas y de posgrado, con más de 10000 estudiantes en actividad promedio anual y más de 500 cursos en cada semestre. Estos guarismos conllevan una serie de desafíos en cuanto a la extracción, depuración y mantenimiento de la información que debe ser generada para cumplir con los objetivos establecidos para la Unidad de Enseñanza (UEFI), que se centran en estudiar y mejorar los procesos de enseñanza y aprendizaje, apoyando la formación didáctica de los docentes universitarios así como desarrollando actividades dirigidas a estudiantes en distintos tramos de sus carreras.

Utilizando la taxonomía sugerida por Baker y Yacef, el trabajo en la UEFI se ha desarrollado en diferentes frentes, ya sean trabajos descriptivos, predictivos, de agrupamientos o incluso de descubrimiento utilizando modelos, tanto sea para generar insumos para la gestión educativa (política), como así también mostrando resultados comparativos de distintas metodologías de enseñanza para cursos específicos.

Desde hace unos años, se persigue un equilibrio entre minimizar la información que se muestra en los distintos informes, maximizando el posible impacto que ésta pueda tener en las entidades solicitantes.

R como herramienta integradora

El rol de R es central: ayuda a extraer información desde bases de datos relacionales, permite la depuración y la visualización instantánea, además de generar reportes prácticamente a tiempo real. Se han usado distintas implementaciones, desde

la versión más simple, hasta la actualidad con RStudio, pasando por algunas GUI como JGR o Rcmdr.

A lo largo de este tiempo se han utilizado diferentes paquetes, desde `rmarkdown` y `knitr` -e incluso `shiny` para la generación de informes, así como el `tidyverse` (o un subconjunto de los paquetes que lo integran) para manipular, modificar y graficar datos de distinto tipo, `httr` y `jsonlite` para extraer datos de entornos virtuales de aprendizaje, `RSQLite` para extraer información de bases de datos estudiantiles, por citar los más utilizados.

Resultados

Centraremos la atención en tres formas de presentar la información hasta ahora, con la ayuda de R: informes dinámicos, encuestas de evaluación docente y (futuras) aplicaciones en Shiny. Los scripts y documentos están en el link <https://gist.github.com/dalessandrini>.

Informes dinámicos En los últimos tiempos hemos focalizado el esfuerzo en automatizar la mayor cantidad de informes posible. En primer lugar nos centramos en los datos recurrentes (ingresos, cantidad de estudiantes activos o inactivos, egresos, promedio de duración de carreras, etc.) y a modo de ejemplo presentamos un informe de Autoevaluación de la institución.

Encuestas SEDE Encuestas que forman parte del Sistema de Evaluación Docente. A raíz de la pandemia por COVID-19 la FIng solicitó informatizar a muchas de las encuestas que aún se realizaban en papel, para recolectar información sobre el desarrollo de los cursos en este periodo especial. Hasta el momento se logró coordinar con los Institutos el formato genéri-

co de la encuesta y algunas reglas específicas para que el trabajo de carga y descarga sea automatizable. Con esta estructura en marcha, en el futuro se podrán generar informes automáticos.

Aplicaciones en Shiny Al momento hemos diseñado algunas apps de consulta rápida de datos, aunque por el momento se encuentran en fase experimental ya que no contamos con los permisos específicos para subirlas a un ser-

vidor (se manejan además datos sensibles como nombres, historial académico y otros).

Referencias

- Baker, R.S.J.d; Yacef, K. 2009. "The state of educational data mining in 2009: a review and future visions"
 - Journal of Educational Data Mining 1 (1): 3–17. <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8>.
-

R para el monitoreo de programas sociales

Introducción

La presentación tiene como objetivo dar cuenta de las diferentes aplicaciones del uso de R en la tarea de monitorear la ejecución de políticas sociales, llevada a cabo por la Dirección de Evaluación y Monitoreo (DINEM) del Ministerio de Desarrollo Social (MIDES) en Uruguay. El uso de R a lo largo de todo el proceso de monitoreo

de los diferentes programas que se ejecutan (transferencias monetarias, programas de proximidad, laborales, educativos, entre otros) nos ha permitido un trabajo más eficiente en términos de optimización del tiempo, recursos y reducción de errores, lo cual redundará en una mejora en la toma de decisiones. Así también se hace referencia a un caso práctico (Canastas de Emergencia) que sintetiza las diferentes aplicaciones.

Aplicaciones

Las principales aplicaciones que destacamos y que involucran todo el proceso de monitoreo de los programas sociales son:

1. Gestión y manejo de bases de datos: realización de extracciones automáticas, limpieza de datos, compatibilización de bases provenientes de distintas fuentes.

Paquetes: `foreign`; `dplyr`; `RPostgreSQL`

2. Informes automáticos: desarrollo de repor-

tes e informes automáticos para el seguimiento de los indicadores de procesos y resultados de los programas que se ejecutan, generando plantillas institucionales comunes. Inclusión de código reproducible y dinámico que permite renderizar por parámetros según dimensiones de interés:

- Temporal (anual, mensual, diaria)
- Geográfica (departamental, local, barrial o polígono específico)
- Otros cortes (equipos de trabajo)

Paquetes: `rmarkdown`; `knitr`; `xtable`; `dplyr`; `ggplot2`; `plotrix`

3. Visualización: programación de infografías reproducibles a escala con indicadores de resultado para diferentes programas sociales.

Paquetes: `ggplot2`; `grid`; `gridExtra`; `useful`; `extrafont`

Ejemplo práctico

Canasta de Emergencia Alimentaria (MIDES 2020)

Descripción: Automatización del proceso de selección de beneficiarios/as de las Canastas de Emergencia alimentaria en el marco de la emergencia sanitaria y social por el COVID-19.



1. Extracción de datos: programación de la consulta SQL con RPostgreSQL y odbc del formulario de postulación a la canasta que se encuentra en la Web del MIDES.
2. Limpieza de datos: se realiza la limpieza de los datos con dplyr, eliminando casos duplicados y ya trabajados, así como errores en las postulaciones.
3. Chequeo con otras bases de datos: se realizan chequeos cruzados con otras bases de datos tanto internas como externas (transferencias monetarias, seguridad social), con el objetivo de delimitar la población objetivo de la canasta y definir el estado de la solicitud.
4. Generación de reportes automáticos: se realizan reportes periódicos en rmarkdown que dan cuenta de las personas que han solicitado

el beneficio, los chequeos realizados y sus resultados, así como el estado final de la solicitud. Se presentan tablas con la función kable() y gráficos de ggplot2 para visualizar la evolución temporal de las mismas y su distribución geográfica.

Referencias

- Allaire, Yihui Xie, JJ, and Winston Chang. 2018. "Rmarkdown: Dynamic Documents for R." <https://CRAN.R-project.org/package=rmarkdown>.
 - Wickham, H. 2009. "Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York."
 - Xie, Yihui. 2017. "Knitr: A General-Purpose Package for Dynamic Report Generation in R." <https://yihui.name/knitr/>.
-

Uso de los registros de la plataforma Moodle para el seguimiento de la actividad estudiantil y diseño de estrategias educativas en un curso masivo de la Facultad de Veterinaria, Udelar

El curso de Bioestadística I (BE1) es dictado para los estudiantes de primer año de la carrera de Veterinaria. La matrícula de mismo es de más de 800 estudiantes. La incorporación de sistemas de gestión de aprendizaje en línea mediante el Entorno Virtual de Aprendizaje - EVA (Moodle) ha permitido estructurar los cursos en la plataforma, otra forma de interacción estudiante-docente, ser un repositorio de materiales, seguimiento académico, incorporación de evaluaciones (tareas, cuestionarios, encuestas, etc.), entre otras funciones.

A raíz del surgimiento de la pandemia de Covid-19, la Universidad de la República suspendió las clases e implementó la modalidad no presencial de los cursos a través de EVA. Esto implicó adaptarse a la enseñanza virtual bruscamente. Por lo que la plataforma se transformó en un elemento clave para la gestión del curso y el seguimiento a los estudiantes.

Objetivos

- Implementar un seguimiento continuo de la generación de estudiantes sobre su actividad y su acceso a los distintos recursos disponibles en la plataforma.
- Usar la información disponible para el diseño de estrategias eficaces sobre la gestión del curso.

Moodle posibilita un seguimiento completo de la actividad individual del estudiante, algo que resulta sumamente engorroso en cursos masificados. Como alternativa, el acceso a registros corresponde a la información sobre la actividad ("clic") que hace el usuario en cada momento. Una de las desventajas que presentan estos registros, es la

flexibilidad para completar campos de texto sin criterios estandarizados (ej. nombres, títulos, secciones, recursos). Sin embargo, la potencia y flexibilidad de los paquetes del universo tidyverse y lubridate para procesar texto, editar y ordenar las bases de datos, generar gráficos de alto impacto y procesar fechas, permitieron analizar estos registros complejos en estructura y volumen sin las dificultades de Moodle.

Las figuras que se muestran a continuación ejemplifican algunas de las visualizaciones generadas. El análisis de esta información permitió observar desde, la cantidad de personas que descargaron una clase teórica a quienes leyeron el reglamento de foros (Figura 1 A y B), así como la frecuencia de ingreso a los materiales (Figura 1A y 2A). Las conclusiones extraídas del análisis de la información permitió tomar decisiones que redundaran en un mejor aprovechamiento de los recursos disponibles por parte de los estudiantes y en el diseño de mecanismos de evaluación óptimos. Algunas de estas decisiones estaban relacionadas a la elección del momento en que se publicaban nuevos recursos, de acuerdo al flujo de estudiantes registrado en la plataforma (Figura 2 B); otras decisiones tuvieron que ver con las características de los métodos de evaluación, por ejemplo: la duración y el día más adecuado para la implementación de las evaluaciones virtuales (Figura 2 D y E). Como conclusión, es evidente que la flexibilidad y potencia que ofrece EVA para la enseñanza virtual así como el software R para manejar grandes volúmenes de datos con estructuras no estandarizadas, resultaron ser herramientas claves para el desarrollo de un curso "a distancia", permitiendo un mejor seguimiento del mismo a partir del

estudio del comportamiento de los estudiantes en la plataforma. Para una nueva edición de BE1, ya se contará con información cuantitativa sobre por ejemplo que temas motivan más a los estudiantes

o que preguntas resultan de mayor dificultad en la evaluaciones.

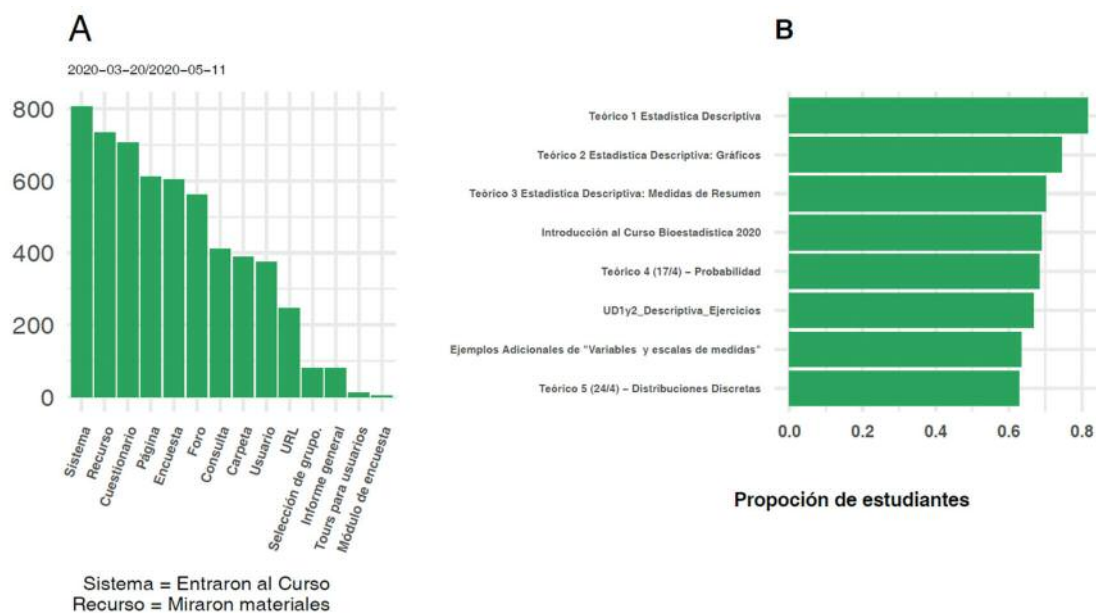


Figura 1: A: Ingreso a componentes del curso. B: Ingreso a materiales.

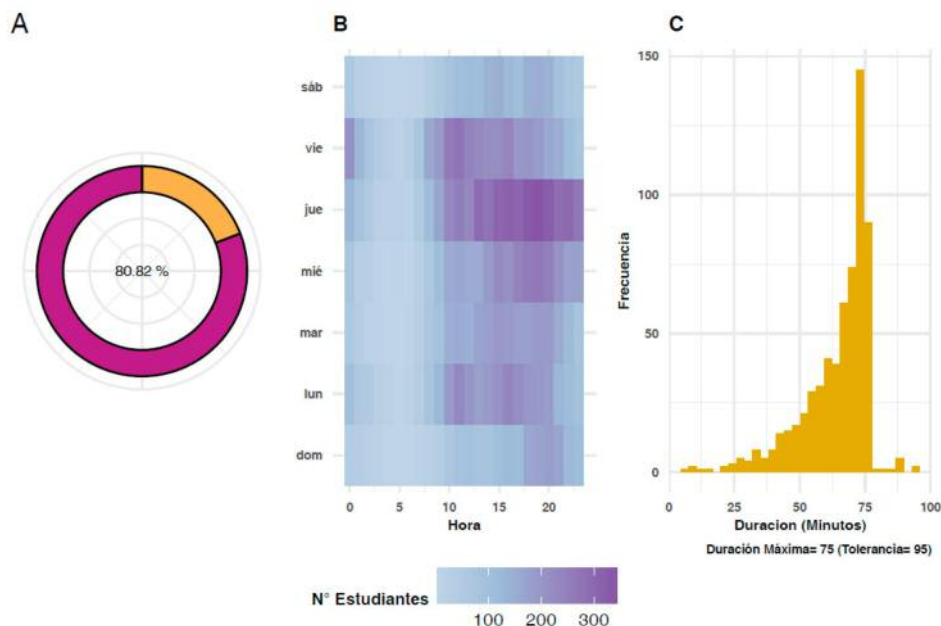


Figura 2: A: Activos en la semana. B: Actividad por día/hora. C: Tiempos del parcial.

Análise preditiva do desempenho dos alunos do curso de pedagogia no âmbito da educação à distância no ENADE

Uma das formas de analisar as instituições de ensino superior e o desempenho dos estudantes no Brasil é através do Exame Nacional de Desempenho dos Estudantes (ENADE). O ENADE busca avaliar o grau de conhecimento dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos de graduação. A partir dos seus resultados é possível tomar decisões inteligentes para a melhoria do processo de ensino-aprendizagem.

O Exame é composto por uma prova, bem como um questionário aplicado aos alunos com o intuito de coletar informações quanto ao seu perfil socioeconômico e aspectos relacionados à sua formação. Portanto, o Enade gera um grande volume de dados que permite estudos visando à melhoria da qualidade do ensino.

A ideia deste trabalho foi identificar fatores que possam influenciar no processo de aprendizagem dos discentes. Mais especificamente, buscam-se relações entre as respostas dadas ao questionário e o resultado obtido na prova. O enfoque é dado aos alunos do curso de Licenciatura em Pedagogia na modalidade de Educação a distância (EaD), oferecido através do Consórcio Centro de Educação Superior à Distância do Estado do Rio de Janeiro (CEDERJ). O Cederj é um consórcio formado por sete universidades públicas do Estado do Rio de Janeiro (UERJ; UENF; UNIRIO; UFRJ; UFF; UFRRJ; IFF) e um centro universitário (CEFET-RJ) em parceria com a Secretaria de Estado de Ciência, Tecnologia e Inovação do Rio de Janeiro, por intermédio da Fundação Centro de Ciências e Educação Superior à Distância do Estado do Rio de Janeiro (CECIERJ), com o objetivo de oferecer cursos de graduação a distância, na modalidade semipresencial para todo o Estado do Rio de Janeiro.

A EaD traz ao ensino superior, alguns desafios e potencialidades a serem explorados, como uma nova forma de se obter conhecimento e qualificação profissional. Sendo assim, como a educação superior tem um papel estratégico na formação, na produção do conhecimento e na informação, a constituição de um sistema de educação superior com modalidade EaD pode contribuir para que os países se adaptem as mudanças de larga escala em curso no mundo todo.

Este trabalho teve como objetivo utilizar a técnica de mineração de dados (TAN et al., 2009) com a utilização de métodos de aprendizado de máquina (MONARD e BARANAUSKAS, 2003). A mineração de dados têm como principal objetivo extrair o máximo de informação de base de dados extensas e, para isso, utiliza-se de técnicas como análise exploratória básica (DA CUNHA e CARVALHAL, 2009), particionamento recursivo (STEINER et al., 2004), análise de agrupamentos (KAUFMAN e ROUSSEUW, 2009), modelagem de regressão (BISHOP, 2006; FOX, 1997), entre outros. Com os avanços tecnológicos de armazenamento de dados, essa área têm ganhado mais visibilidade (HAND, 2006).

O aprendizado de máquina é um ramo de inteligência artificial cujos estudos são feitos para que, com o auxílio de algoritmos, o computador tome decisão com base em informações inseridas neles. Para esse trabalho, o algoritmo utilizado é a regressão de floresta aleatória ou em inglês Random Forest (BREIMAN, 2001), como um modelo de regressão capaz de prever o desempenho do estudante no ENADE através das variáveis socioeconômicas, buscando identificar e analisar o perfil dos estudantes que prestaram a prova.

Dessa maneira, o aprendizado de máquina fornece a base técnica para a mineração de dados

que transforma dados brutos em informações de mais fácil compreensão, como previsões, correlações e relações de causalidade, o que, no processo de análise, auxilia na compreensão e explicação de fenômenos. No que tange os pacotes utilizados no software R, destaca-se o tidyverse, na parte de manipulação de dados, randomForest com o objetivo de obter a regressão de florestas aleatórias, caret e Metrics para obter métricas de avaliação do modelo e, por fim, ggplot2 na visualização gráfica.

A base de dados utilizada para a realização deste estudo foi adquirida no portal do INEP, onde os dados estão disponíveis para o público através de download (INEP, 2017). Os dados escolhidos são oriundos da base de dados do ENADE 2017, que contém dados relativos aos estudantes que realizaram o exame e responderam ao questionário.

Referências

- BISHOP, C. M.. Pattern recognition and machine learning. Springer, 2006.
 - BREIMAN, L.. Random forests. Machine learning, Springer, v. 45, n. 1, p. 5–32, 2001
 - DA CUNHA, S. B.; CARVAJAL, S.. Estatística Básica - a Arte de Trabalhar com Dados. Elsevier Brasil, 2009.
 - FOX, J.. Applied regression analysis, linear models, and related methods. Sage Publications, Inc, 1997.
 - HAND, D. J. Data Mining. Encyclopedia of Environmetrics , v. 2, JohnWiley & Sons, Ltd.,2006.
 - KAUFMAN, L.; ROUSSEEUW, P. J.. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
 - MONARD, M. C.; BARANAUSKAS, J. A.. Conceitos de aprendizado de máquina. In S. O. Rezende, editor, Sistemas Inteligentes - Fundamentos e Aplicações, pages 89–114. Editora Manole, 2003.
 - RAMESH, V. PARKAVI, P. e RAMAR, K.(2013) Predicting Student. Performance: A Statistical and Data Mining Approach. International Journal of Computer Applications. [S.l: s.n.].
 - STEINER, M. T. A. et al.. Data Mining como Suporte à Tomada de Decisões-uma Aplicação no Diagnóstico Médico. XXXVI Simpósio Brasileiro de Pesquisa Operacional, "O impacto da pesquisa operacional nas novas tendências multidisciplinares", v. 23, p. 96-107, 2004.
 - TAN, PANG-NING et al.. Introduction to Data Mining, Pearson, 2009.
-

Sesión **Datos espaciales**

Uso de R e imágenes satelitales para la generación de modelos predictivos y caracterización de plantaciones forestales

La región Mesopotámica Argentina concentra el 80 % de las plantaciones forestales del país. A partir de la técnica de la interpretación visual de imágenes satelitales es posible identificar y delimitar las plantaciones, pero no describir espacialmente información complementaria (edad, altura, área basal, volumen) muy valiosa para caracterizar el estado de las masas forestales. A partir del procesamiento de información espectral de imágenes del satélite Sentinel 2 en la plataforma

Google Earth Engine, información correspondiente al Inventario Forestal del departamento Concordia de la Provincia de Entre Ríos (2017) y modelos de regresión lineal (MRL) generados utilizando R (figura 1), se estimaron distintas variables de estado forestales (figura 2) que permiten caracterizar espacialmente el recurso forestal correspondiente a plantaciones comerciales del género Eucalyptus (figura 3).

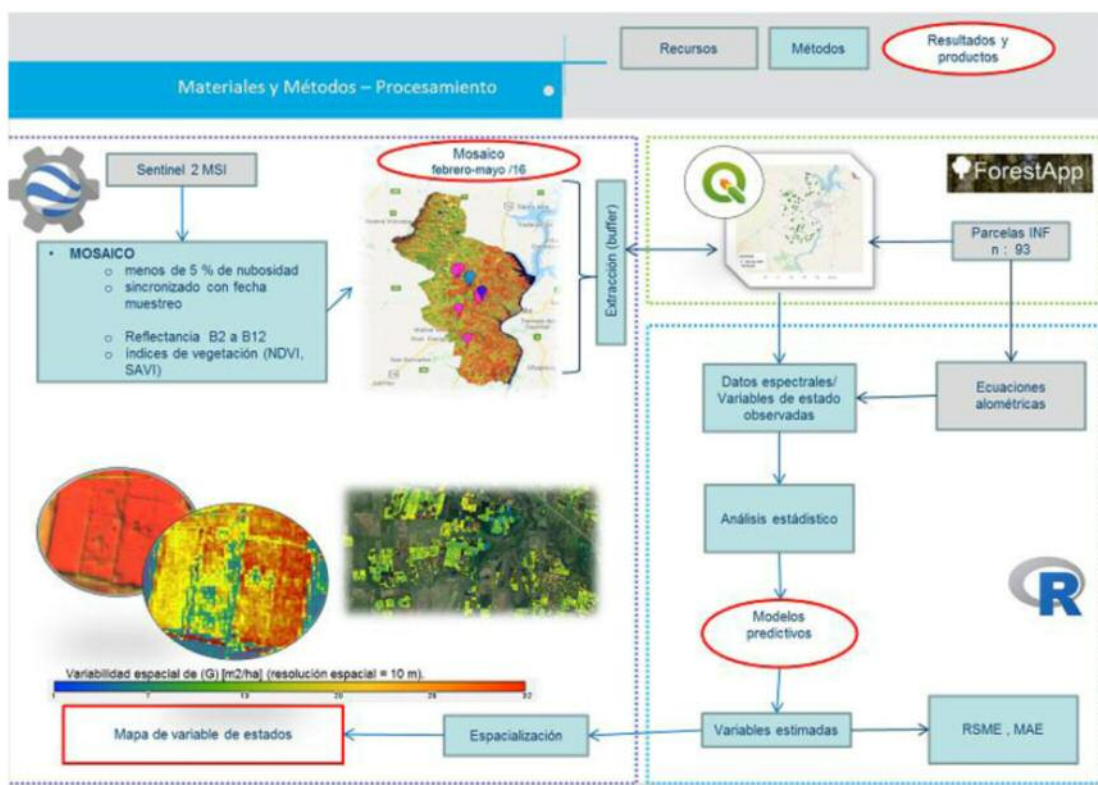


Figura 1. Flujo del procesamiento de los datos.

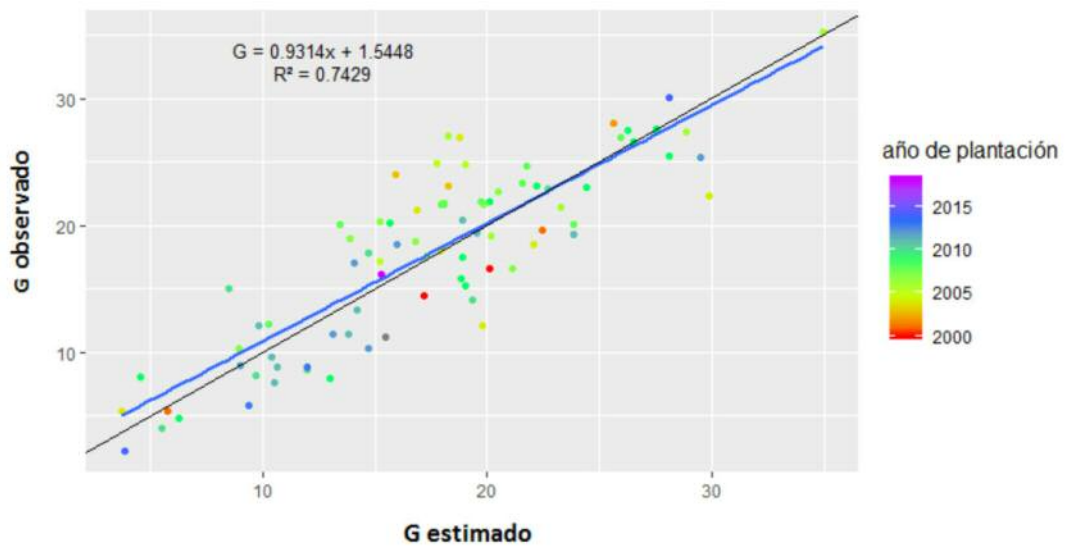


Figura 2. Gráficos de dispersión de los valores observados versus los estimados de área basal (G) [m²/ha] .En líneas sólidas se observa la recta de ajuste de referencia (1:1). En línea azul la recta de ajuste del modelo. En colores los años de plantación registrados en la base de datos.

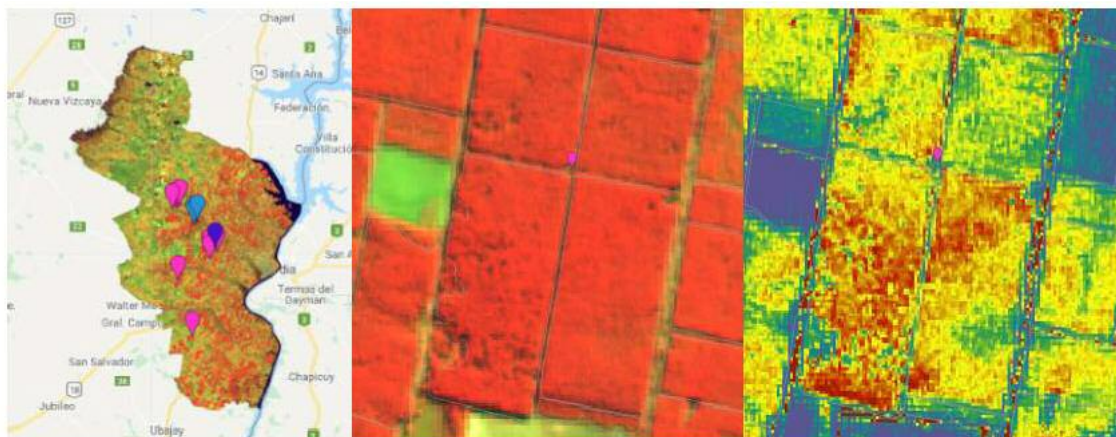


Figura 3. De izquierda a derecha: Mosaico correspondiente al área de estudio. Imagen falso color compuesto a escala de rodal. Mapa del área basal (G), donde se aprecia la variabilidad espacial a escala de rodal (resolución espacial, 10 m).

Optimización de parámetros geoestadísticos para interpolación de lluvia para uso en el sector agropecuario

Introducción

La actividad agropecuaria es altamente dependiente del clima y estado del tiempo, principalmente de la variable lluvia. A su vez, dentro del sector agropecuario, la agricultura es un sector clima dependiente, potencialmente afectado por las consecuencias del clima, siendo la variabilidad climática interanual, la que presenta la principal incidencia relativa comparada con la variabilidad decadal y la de largo plazo (Tiscornia et al., 2016).

Tener mapas interpolados de lluvia es de gran importancia para la toma de decisiones y gestión de riesgos entre otras posibilidades. El objetivo del siguiente trabajo es evaluar diferentes técnicas de optimización de hiperparámetros como grid search, random search y optimización bayesiana para obtener imágenes interpoladas raster de lluvia con el menor error posible y hacerlo de manera automática.

Materiales y métodos

Para el siguiente trabajo se utilizó información de la red pluviométrica del Instituto Nacional de Meteorología (INUMET) de Uruguay para el mes de marzo de 2020. La red pluviométrica de INUMET cuenta con casi 350 pluviómetros distribuidos en todo el país. Del mes de marzo se evaluaron 3 días con lluvias: 11, 15 y 17 de marzo. De esos días se usaron solo los pluviómetros que contaran con datos. En la tabla 1 se ven la cantidad de pluviómetros usados cada día.

Los paquetes R utilizados en este trabajo fue-

Fecha	11/03/2020	15/03/2020	17/03/2020
Pluviómetros	227	223	226

Tabla 1. Cantidad de pluviómetros usados en cada fecha evaluada.

ron gstat, mlrMBO, randomsearch y raster. Es importante aclarar que para el adecuado funcionamiento del paquete mlrMBO es necesario tener instalado el paquete rgenoud.

Para la generación de imágenes raster de precipitación se utilizó la técnica de interpolación conocida como kriging ordinario (Bivand, Pebesma, & Gómez-Rubio, 2013). En la técnica de kriging a partir de puntos georreferenciados con información observada para una variable de interés, en nuestro caso lluvia, se genera una grilla uniforme en donde cada celda cuenta con un valor obtenido por interpolación a partir de datos reales. Esta grilla uniforme tiene valores no solo para los puntos donde se conoce el valor de la variable, sino también para aquellos puntos donde no se tiene observaciones de la variable, es decir los valores son predichos.

Para poder realizar la interpolación es necesario definir valores a usar por determinados parámetros geoestadísticos como nugget, psill y range. Según los valores que se usen, el error que tenga el raster final. La búsqueda manual de los parámetros que generen el raster con el menor error de predicción es una tarea ardua, y prácticamente imposible por que las combinaciones de parámetros son infinitas.

En R la interpolación de lluvia devuelve 2 grillas raster, una con el valor predicho de lluvia para cada celda, y otra con la varianza de la predicción para cada celda.

El error de predicción del raster se evalúa a través del desvío estándar de lluvia definido como la raíz del promedio de la varianza de lluvia de las celdas del raster final. El objetivo final es generar un raster de lluvia con el menor desvío estándar.

En R se evaluaron las técnicas de optimización

Parametro	Psill	Range	Nugget
Rango	1 a 100	1 a 200	1 a 100

Tabla 2. Rango de búsqueda para cada parámetro geoestadístico.

de hiperparámetros conocidas como grid search (GS), random search (RS), y optimización bayesiana (OB) (Feurer & Hutter, 2019). Estas técnicas de optimización buscan encontrar de manera automática aquellos parámetros geoestadísticos que minimicen una función objetivo, en este estudio el desvío estándar.

Las técnicas de optimización requieren que uno defina de antemano para cada variable a ajustar, el rango de búsqueda, es decir entre que valores mínimos y máximos puede estar cada variable, cuando es numérica, y que valores puede tomar si es discreta. En la tabla 2 se muestra el rango de búsqueda para cada parámetro geoestadístico.

Resultados y discusión

A continuación, a modo de ejemplo se presentan los resultados obtenidos para las diferentes técnicas de optimización solo para el día 15/03/2020. En la tabla 3 se ven los parámetros que dieron como resultado el desvío estándar más bajo en la etapa de interpolación para cada técnica de optimización.

Como se ve en la tabla 3, las técnicas evaluadas obtuvieron valores diferentes de desvío estándar en la etapa de interpolación, siendo optimización bayesiana la que tuvo mejor desempeño con el desvío más bajo, seguida por random search.

En cuanto al tiempo de ejecución optimización bayesiana es la que lleva más tiempo demorando 6 veces más en su ejecución en comparación a las otras técnicas.

Conclusiones

Es posible usando diferentes técnicas de optimización de hiperparámetros obtener un raster de interpolación de lluvia con valores de desvío estándar para lluvia diferentes a partir de diferentes valores de parámetros geoestadísticos. Optimización bayesiana fue la técnica que obtuvo el desvío estándar más bajo. Sin embargo, hay que tener en cuenta que el resultado mostrado es sobre una sola fecha, pudiendo en otras obtener resultados distintos. A futuro es necesario seguir explorando y profundizando el conocimiento de estas técnicas para obtener mapas de lluvia con menor incertidumbre para uso en el sector agropecuario.

Bibliografía

- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Interpolation and Geostatistics. En *Applied Spatial Data Analysis with R* (pp. 213-261). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7618-4_8
- Feuerer, M., & Hutter, F. (2019). Hyperparameter Optimization. En F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges* (pp. 3-33). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1
- Tiscornia, G., Cal, A., & Giménez, A. (2016). Análisis y caracterización de la variabilidad climática en algunas regiones de Uruguay. *RIA. Revista de investigaciones agropecuarias*, 42(1), 66-71.

OPTIMIZACIÓN	CORRIDAS	NUGGET	PSILL	RANGE	DESVÍO ESTÁNDAR	TIEMPO DE EJECUCIÓN (segundos)
BAYESIANA	125	1.00	1.02	196.30	1.19	306
GRID SEARCH		23.20	18.70	196.80	5.61	50
RANDOM SEARCH		1.48	5.66	159.89	1.90	52

Tabla 3. Mejores corridas según técnica de optimización para el día 15/03/2020.

Integración de herramientas SIG, bases de datos, R espacial, Rmarkdown en la automatización de informes periódicos de la información en el Programa de Erradicación de plagas ISCAMEN, Mendoza, Argentina

El Instituto de Sanidad y Calidad Agropecuaria Mendoza (ISCAMEN) tiene entre sus objetivos el control y erradicación de distintas plagas cuarentenarias en la zona de influencia. La obtención en tiempo real del análisis de la información registrada es esencial, pero debido a su relativa complejidad es necesario poder integrar de forma eficiente distintas herramientas para obtener un resultado satisfactorio.

El objetivo de la institución es establecer un flujo de trabajo interno basado en servicios las infraestructuras de datos espaciales (IDE) con el consiguiente abaratamiento de costos; respuesta más rápida y justificada frente a problemas concretos, utilización de mapas temáticos como canal privilegiado de comunicación hacia los gerenciadore del Programa; implementación de estándares nacionales e internacionales en los atributos de los geodatos para una correcta compatibilización e interoperabilidad con otras instituciones.

En esta última etapa se definieron proyectos en R con rutinas de trabajo en rmarkdown que integran y automatizan el uso de librerías de lec-

tura, manejo y resumen de bases de datos de distintas fuentes, análisis espaciales descriptivos, de modelado de autocorrelación espacial, gráficos y agregado de datos en mapas. El proceso completo determina el filtrado y control de calidad de los datos obtenidos de las bases de datos crudas, definición del periodo considerado, la distribución espacial de los recuentos en las trampas y valores de la cantidad de moscas capturadas en las trampas expuestas por día (MTD) con proyección de los polígonos y valores sobre los proyectos colaborativos para crear mapas editables OpenStreetMap. Las etapas de procesamiento se resumen en la Figura 1.

En la etapa de lectura de datos se utilizaron las librerías: readxl, ya que los registros semanales se guardan en formato *.xlsx. Para el resumen de la información, manejo de variables temporales y diseño de tablas se utilizaron las librerías: plyr, officer y flextable. El análisis espacial requirió la aplicación de librerías que permitiera agregar shap-es, realizar interpolación y proyectar sobre mapas geográficos, entre las librerías aplicadas se



Figura 1 Etapas de procesamiento de la información integrando distintas herramientas para el análisis de datos con infraestructura espacial

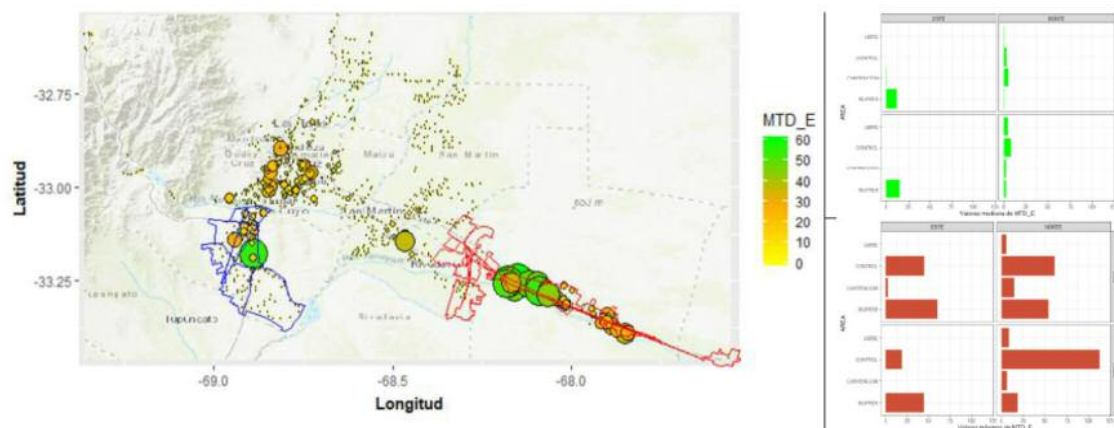


Figura 2. Distribución espacial de los valores de moscas estériles liberadas MTD_E (moscas_estériles x trampas x día) y gráficos de los valores medianos y máximos de MTD_E en los distintos oasis, según las áreas y la ubicación

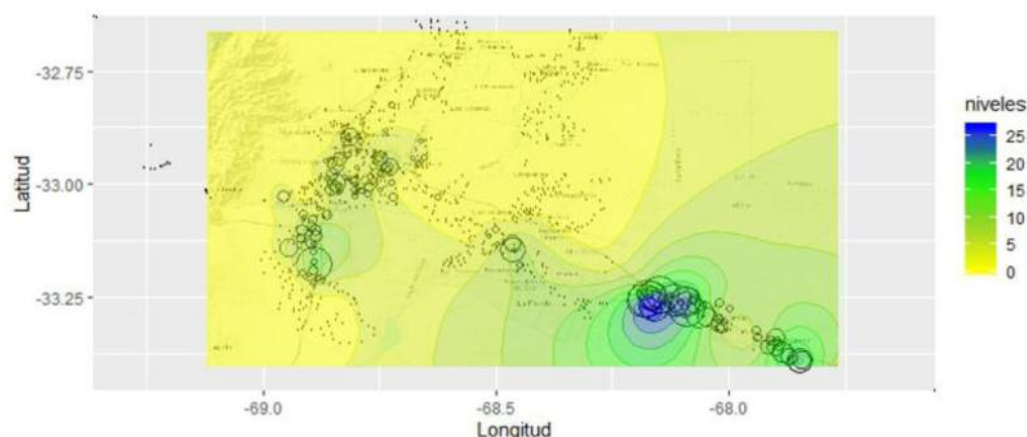


Figura 3 Zonificación espacial de los valores de MTD_F

están GISTools, maptools, geoR, ggplot2, ggiso-band y OpenStreetMap.

A modo de ejemplo se muestra parte de los resultados de mayor interés y rápida visualización que se lograron con datos de liberación de moscas estériles para el control de mosca de la fruta en la semana del 16 al 19 de marzo. A partir del procesamiento de la información semanal se generaron reportes donde además de distintas tablas con estadísticas descriptivas se elaboraron mapas con la distribución espacial ponderada en función de la liberación de moscas estériles con los polígonos de áreas de trabajo y gráficos descripti-

vos por oasis (Este y Norte) y áreas (libre, control, contención y buffer) como los de la Figura 2.

Se determinó el modelo de autocorrelación espacial través de un análisis geoestadístico que permitió la definición de áreas homogéneas (isolíneas) con distinto grado de liberación de moscas estériles se generaron mapas de estratos diferencias como se muestra en la Figura 3.

El uso integrado de herramientas en un entorno R permitió sistematizar los análisis de la información generada por la institución y generar reportes de manera rápida y ágil.

Sesión **Modelos y aplicaciones**

Ventajas del análisis de redes para optimizar la vigilancia y control de Brucelosis bovina en la Argentina

Introducción

En el contexto de la transmisión de enfermedades el análisis de redes tiene como objetivo el estudio de la diseminación de las mismas mediante la descripción de las interacciones entre nodos (ej. establecimientos o individuos). Los nodos con un número elevado de contactos cobran un rol protagónico ya que pueden diseminar las infecciones a muchos otros y por lo tanto son denominados “súper-diseminadores” [1]. Su identificación y las intervenciones dirigidas a estos suelen ser altamente efectivas para el control de las enfermedades. La brucelosis bovina es una enfermedad bajo control oficial en Argentina y se transmite principalmente por el traslado de vacas infectadas. Sin embargo, no se conoce el estado sanitario de todos los establecimientos ganaderos [2]. Por ello nuestro objetivo fue utilizar el análisis de redes en combinación con datos de diagnósticos de laboratorio para identificar aquellos establecimientos con estado desconocido a la Brucelosis y que a su vez poseen un papel clave en la diseminación de la enfermedad para optimizar así la vigilancia de esta zoonosis en la Rep. Argentina.

Métodos y herramientas

La generación de las bases de datos que alimentaron nuestro análisis requirió compaginar fuentes oficiales como los documentos de traslado de animales (DT-e) emitidos electrónicamente durante 2018 y datos acumulados desde el 2014 a 2018 de diagnósticos serológicos de *Brucella* sp. producidos por el laboratorio del Servicio Nacional de Sanidad Animal y Calidad Agroalimentaria (SENASA). El desafío de identificar los establecimientos más riesgosos fue resuelto mediante el análisis de redes a través del

paquete *igraph* de R [3]. Este nos permitió construir, manejar y operar sobre un grafo dirigido y ponderado por el número de animales reproductores trasladados (contacto) entre establecimientos (nodos). A partir del cómputo para cada nodo de una propiedad de centralidad básica, el grado, calculamos una tasa relacionada con la diseminación de enfermedades a nivel de la red, el Número Reproductivo Básico (R_0) [4]. Luego, los establecimientos se ordenaron de modo descendente de acuerdo al grado y se calcularon sucesivos R_0 iterativamente para toda la red, eliminando de a un establecimiento, comenzando por aquellos con los grados más altos. De este modo, para cada iteración, se calculó la contribución relativa al R_0 de la red completa de cada establecimiento. Aquellos establecimientos involucrados en la reducción del R_0 en un 90% se identificaron como “súper-diseminadores”. Dentro de estos últimos, se identificaron los establecimientos con resultados serológicos positivos, negativos y sin estado definido a la *Brucella* sp.

Resultados

Con este método se pudo identificar que casi el 12% (16.731) del total de establecimientos (139.914) son de alto riesgo (Figura 1), lo que indica que relativamente pocos nodos tienen un papel clave en la propagación de la enfermedad. Luego de relacionar estos resultados con los diagnósticos de laboratorio pudimos observar que el 69,3% (11.593) de estos establecimientos no poseen diagnóstico, el 3,5% (579) poseen diagnóstico positivo (Figura 2) y el 27,2% (4.559) son negativos. En conjunto con el paquete *ggplot2* [5], se obtuvo la visualización de todos estos resultados.

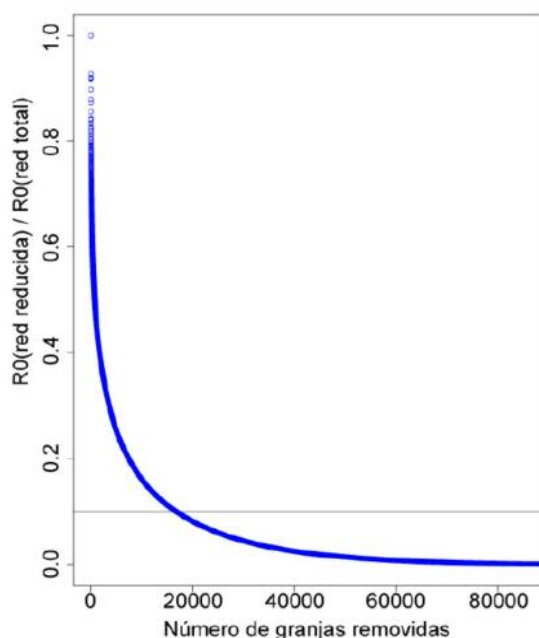


Figura 1: Impacto de la eliminación selectiva de los establecimientos en el valor R_0 (red reducida / red completa) basado en el grado total ponderado.

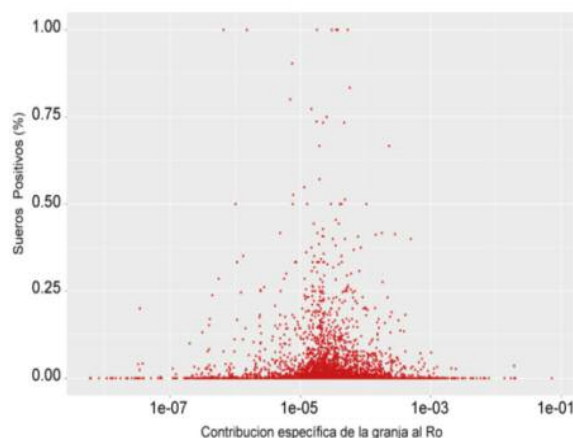


Figura 2: Relación entre la contribución de cada establecimiento al R_0 (eje X) y la proporción de sueros positivos a *Brucella sp.* (eje Y).

Conclusiones y perspectivas a futuro

Con los resultados de esta metodología pudimos comprobar que el análisis de redes es de suma utilidad para resolver problemáticas relacionadas a la diseminación de enfermedades. En combinación con diagnósticos de laboratorios se convierte en una herramienta potente para priorizar y direccionar acciones de control y vigilancia en el marco de una gran población de establecimientos ganaderos, pero con bajos recursos económicos que restringen las posibilidades de muestreo y los análisis de laboratorio como ocurre en Latinoamérica. En un futuro planeamos generalizar este análisis en el plano temporal a fin de visualizar los cambios en la red a lo largo de varios años y mejorar la precisión de los análisis de transmisión de enfermedades.

Bibliografía

- Keeling, M.J., & K.T.D. Eames. (2005). Networks and epidemic models. *J. R. Soc. Interface* 2, 295–307. DOI:10.1098/rsif.2005.0051
- <http://www.senasa.gob.ar/normativas/resolucion-67-2019-senasa-servicio-nacional-de-sanidad-y-calidad-agroalimentaria>
- Csardi, G. & T. Nepusz. (2006). The igraph software package for complex network research. *Int. J. Complex Syst.* 1695, 1–9.
- Woolhouse M.E.J., D. J. Shaw, L. Matthews, W.-C. Liu, D. J. Mellor & M. R. Thomas. (2005). Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule. *Biol. Lett.* 1, 350–352. doi: 10.1098/rsbl.2005.0331
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.

Palabras clave: Respostas de contagem longitudinais, Distribuição de Poisson, Aproximação de integrais, Influência local, Métodos de Monte Carlo e Metropolis-Hastings

Análise de sensibilidade de respostas de contagem longitudinais usando R com aplicação a dados médicos

Resumo

Respostas de contagem longitudinais ocorrem frequentemente em diversas áreas da ciência. Estas respostas são comumente descritas pelo modelo Poisson com efeitos mistos. Uma questão relevante que deve ser considerada em toda modelagem estatística, uma vez que a estimação de parâmetros é realizada, corresponde a análise de sensibilidade. Apesar da popularidade deste modelo, ferramentas de análise de sensibilidade para avaliar seu ajuste ainda são escassas. Assim, este trabalho apresenta uma metodologia de análise de sensibilidade das respostas de contagem longitudinais do modelo Poisson com efeitos mistos usando R. Finalmente, uma aplicação a dados médicos mostra a utilidade da proposta.

Metodologia

Respostas de contagem longitudinais ocorrem frequentemente em diversas áreas, por exemplo, medicina, biologia, psicologia, sociologia, humanidades, economia, agricultura. Estas respostas são comumente analisadas através do modelo Poisson com efeitos mistos (MPEM). Este modelo acomoda a estrutura de correlação existente entre as respostas por meio de efeitos aleatórios (não observáveis). A estimação de parâmetros do modelo é conduzida usualmente pelo método de máxima verossimilhança (ML). Porém, não é uma tarefa computacional fácil, pois a incorporação dos efeitos aleatórios leva à função de verossimilhança dos dados observados a incluir integrais, que não possuem solução analítica e podem ser de grande dimensão. Por consequência, aproximações dessas integrais por aproximação de Laplace (AL) e adaptativa quadratura de Gauss Hermite (AQGH) são necessárias para a obtenção destas estimativas. Neste trabalho, as estimativas

de ML são obtidas usando a função `glmer` do pacote `lme4` de R. Note que esta função está disponível com AQGH quando o modelo possui um intercepto aleatório e com AL para dois ou mais efeitos aleatórios.

Uma questão relevante que deve ser considerada em toda modelagem estatística, uma vez realizada a estimação de parâmetros, e a análise de sensibilidade. Esta permite investigar/avaliar como as diferentes fontes de incerteza do modelo podem afetar as estimativas de ML e, consequentemente, as inferências e tomadas de decisão. No obstante, este tipo de ferramentas para o MPEM ainda são escassas, principalmente para suas respostas de contagem, que são uma fonte de incerteza relevante para o modelo. Neste contexto, este trabalho apresenta uma metodologia de análise de sensibilidade usando a técnica de influência local sob uma estratégia de perturbação indireta para as respostas de contagem de um MPEM. Esta metodologia é resumida em um algoritmo é implementada no R; para mais detalhes vide Tapia et al. (2019).

Como a influência local é uma técnica baseada na função de verossimilhança, seu uso para o MPEM apresenta o mesmo problema das integrais mencionado acima. No obstante, essa dificuldade pode ser evitada, tratando os efeitos aleatórios como dados faltantes, definindo assim a função de verossimilhança dos dados completos (observados e faltantes). Desta maneira, a técnica de influência local é desenvolvida com base na função de verossimilhança dos dados completos, compreendendo o uso do método de Monte Carlo e algoritmo Metropolis-Hastings para a aproximação de matrizes fundamentais no cálculo das curvaturas individuais que permitirão avaliar a influência de uma observação. Na prática,

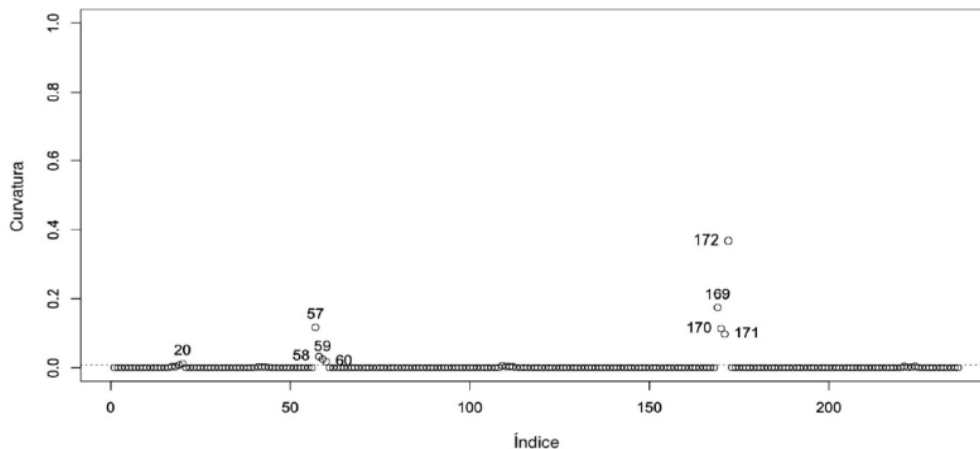


Figura 1: Pacientes identificados como influentes para os dados de epilepsia.

esta metodologia se resume no seguinte algoritmo: (i) realizar uma análise exploratória dos dados usando os pacotes base e ggplot para a obtenção de estatísticas e visualizações, respectivamente; (ii) baseado em (i), formular um MPEM e obter as estimativas de ML usando a função glmer do pacote lme4; (iii) baseado nas estimativas de ML, amostrar $S = 10,000$ observações relacionada aos efeitos aleatórios através do algoritmo Metropolis-Hastings; (iv) com as observações amostradas em (iv), e depois de eliminar as primeiras $M0 = 1000$, aproximar as matrizes fundamentais; (v) com as matrizes obtidas em (iv), calcular as curvaturas individuais usando o pacote matrixcalc e um ponte de corte adequado para identificar as observações influentes; (vi) realizar uma gráfica das curvaturas individuais versus o índice das observações, traçando o ponto de corte, para visualizar as observações identificadas como influentes; (vii) avaliar a magnitude do impacto exercido nas estimativas de ML pelas observações influentes ou um conjunto delas, calculando o erro percentual (PE) para todas as combinações de observações influentes e conjuntos de observações influentes (quando pertencem ao mesmo indivíduo). Por último, se deseja desenvolver um pacote de análise de sensibilidade para modelos mistos com respostas discretas que contenha tanto esta metodologia quanto outras.

Aplicação

Um conjunto de dados de um ensaio clínico com 59 pacientes que sofrem de epilepsia foi apresentado por Thall e Vail (1990). O objetivo deste estudo foi analisar se uma nova droga reduz as crises epiléticas ou não. A resposta de

interés é o número de crises epiléticas vivenciadas por cada paciente durante um período de duas semanas a cada quatro visitas à clínica. Depois de uma análise exploratória dos dados, considera-se um modelo Poisson com efeitos mistos definido por:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{Base}_{ij} + \beta_2 \text{Trt}_{ij} + \beta_3 \text{Base}_{ij} \text{Trt}_{ij} + \beta_4 \text{Age}_{ij} + \beta_5 \text{Visit}_{ij} + \beta_{10} + \beta_{11} \text{Visit}_{ij}, \quad (1)$$

para $j = 1, \dots, 4$, $i = 1, \dots, 59$, e $q = 236$, em que "Base" é o número de crises de um pre-ensaio clínico; "Trt" é um indicador de tratamento o placebo; "Base Trt" é a interação entre "Base" e "Trt"; "Age" é a idade em anos; e "Visit" um indicador dos períodos. A Figura 1 mostra os pacientes identificados como influentes. A eliminação de combinações de pacientes leva a variações (PE) importantes nas estimativas de β_0 , β_3 , β_4 e β_5 , sendo estas as mais sensíveis à estratégia de perturbação indireta adotada. No entanto, com um nível de significância do 5% não foram detectadas alterações inferenciais. Em conclusão, esta metodologia permite obter informações valiosas sobre estimativas de ML e respostas de contagem que precisam de escrutínio adicional neste modelo, permitindo obter uma melhor perspectiva sobre as consequências dos dados quando observações influentes são indentificadas.

Referências

- Tapia, A., Giampaoli, V., Diaz, M.P. e Leiva, V. 2019. "Sensitivity analysis of longitudinal count responses: a local influence approach and application to medical data". Journal of Applied Statistics, 46 (6): 1021–1042. <https://doi.org/10.1080/02664763.2018.1531978>

R en Temas de Industria

Introducción

Se presenta el uso de R en dos problemas, el primero tiene como objetivo la optimización logística del diseño de paletizado y de carga en transporte terrestre, con cálculo de huella de carbono; el segundo construye un conjunto de alternativas óptimas para el geoposicionamiento de una planta de bioetanol producido a partir de los rastrojos de maíz y sorgo en Argentina. Los planteos propuestos se desprenden de situaciones reales; en el caso de la paletización y transporte se emulan y rediseñan softwares existentes hasta crear un código propio en el entorno de desarrollo integrado (IDE) del lenguaje "R", y en el caso de la producción de bioetanol se consideran los datos temporales disponibles en bases gubernamentales.

Metodología y Resultados

Problema 1: Desarrollo de un código de paletizado, carga y transporte con cálculo de huella de carbono.

Se definieron tres etapas de ejecución:

- La creación de un código en "R" (script) que resuelva el problema de diseñar el mosaico espacialmente óptimo para la unidad de carga en pallet (llamado problema PLP) (ver Figura 1).
- La modificación del código creado incorporando la disposición de los pallets en transporte terrestre, atendiendo a las restricciones de espacio y peso, incluyendo en el dataframe solución, la optimización espacial vehicular.
- La incorporación al código anterior del cálculo de la huella de carbono producida por el transporte origen-destino del total de la carga, considerando trayecto, tipo de vehículo y combustible (TonCO₂) (ver Figura 2).

Para determinar las características de importancia que debía heredar el código en "R" se ana-

lizaron estadísticamente (estudio descriptivo, correlacional, de componentes principales y de variabilidad) los programas de paletización TOPS (<https://topseng.com/>) y PLMPack Stack Builder (<https://www.treedim.com/stackbuilder/es/>), fijando un conjunto de tipos y valores de empaque, pallet y vehículos de carga sobre los cuales se trabajó.

El cumplimiento de la segunda y la tercera etapa supuso un trabajo de recodificación en R, incorporando un diagrama de flujo que contemplara:

- Elección del vehículo (nacen las restricciones del transporte: Largo, Ancho, Alto, Kilaje máximo).
- Determinación del diseño de la disposición de las unidades de carga (pallets) en el transporte.
- Cálculo del kilaje máximo por unidad de carga (pallet), y de la altura de la estiba (considerando las restricciones iniciales, y la cantidad de pallets por diseño de disposición en el transporte).
- Diseño de paletizado (mosaico de cajas por pallet), recodificando el script generado en la etapa 1 con la incorporación de la limitación por peso y altura determinadas por el paso anterior.
- Cálculo de huella de carbono por transportación (asociada a envergadura, tipo y combustible del transporte y a la distancia a recorrer, script calculadora siguiendo formulaciones oficiales).
- Decisión de tipo de salidas, variables que se mostrarán y su formato de visualización.

Problema 2: Determinación del geoposicionamiento de una Planta industrial de Bioetanol en Argentina.

Un primer trabajo se basó en filtrar la base glo-

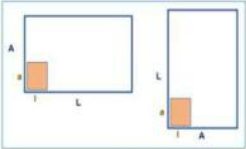
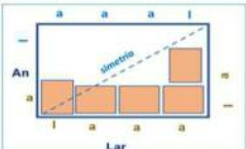
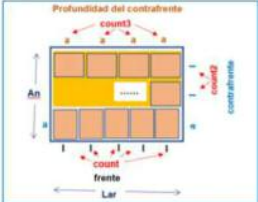
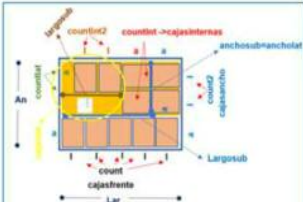
Modelos	Pseudocódigo
 <p>• Distribución "C"</p>	<ol style="list-style-type: none"> 1. Leer archivos de datos 2. Imprimir listado de medidas correspondientes a cada elemento (pallet/ caja) 3. Solicitar la elección del tipo deseado y Escanear respuesta 4. Si la elección se corresponde con un tipo estipulado <i>entonces</i> <ul style="list-style-type: none"> - Asignar las medidas correspondientes en un vector <p><i>Caso contrario</i></p> <ul style="list-style-type: none"> - Solicitar la escritura de las medidas personalizadas - Escanear datos y cargarlos en el vector <ol style="list-style-type: none"> 5. Fin.
 <p>• Distribución "D"</p>	<p>Código</p> <pre> {tipo_caja<-read.table("medidas_de_caja.txt", header=TRUE, sep=",") print (tipo_caja) print ("Ingresa el número de la caja que elijas") a<-scan(n=1) if (a!=9) {caja<-c(tipo_caja[a,1],tipo_caja[a,2],tipo_caja[a,3])} else {print("Ingresa las tres medidas personalizadas: largo, ancho, alto de la caja en mm ") caja<-scan(n=3)} } </pre>
 <p>• Distribución "B"</p>	
 <p>• Distribución "T"</p>	

Figura 1. Diseños de mosaico en pallet y su codificación en R

Especificación mosaico vehículo y Huella de Carbono	Especificación distribución de cajas en pallet
	

Figura 2. Salidas generadas con "qrcode" desde un óptimo del dataframe solución.

bal descargada (dataset agropecuario oficial) para generar los arreglos (estructuras de datos) que contuvieran las variables y datos de utilidad para el objetivo (tipo de cultivo: maíz y sorgo, cantidad producida, provincia, localidad, y campaña). A través de las herramientas gráficas y de visualización del paquete ggplot2 de "R" trabajadas sobre las bases de datos creadas se identificaron las regiones de mayor importancia productiva dentro del país (ver Figura 3, izquierda), posteriormente se utilizó el método de conglomerados (clusters) para el tratamiento de las similitudes-disimilitudes sobre la productividad/localización de cada cultivo, hallándose el número n de clusters óptimo (n=4 para el maíz y n=5 para el sorgo) y cotejando su bondad de ajuste (utilizando la función Kmeans y gráficos estándar del tipo scat-

terplots/pairs, etc.). El filtro final aplicado sobre el universo geográfico redujo a 21 localidades productivamente destacadas para el maíz y 26 localidades para el sorgo.

La distribución de las localidades por conglomerado determina una diferenciación de las regiones de cultivo por peso productivo geconjunto, dando paso a un último trabajo de geoposicionamiento para ayudar a la toma de decisión de la ubicación de la planta, según las coordenadas geográficas (latitud-longitud) de los centroides de cada cluster.

Finalmente se determinan cuatro soluciones posibles para la localización de la planta siguiendo estrategias de ponderación sobre los centroides más productivos de cada cultivo. Todo el proceso se visualiza a través de la plataforma Google

Maps (ver Figura 3, derecha). Modelos matemáticos industriales utilizarán estos datos considerando parámetros (como la estructura vial) para la ubicación final de la planta (<https://www.ingenieriaindustrialonline.com/disenio-y-distribucion-en-planta/metodos-de-localizacion-de-planta/>).

Conclusiones y Trabajos Futuros

En el área logística el script de carga desarrollado es una herramienta de apoyo a la decisión del ingeniero en empaque que puede robustecerse con elementos gráficos (tipo raster) para la visualización del mosaico de paletizado. El código creado también puede complementarse con herramientas externas de criptoseguridad (QR de

trazabilidad logística, blockchain).

El problema de determinación del geoposicionamiento de una planta de bioetanol otorga un conjunto de soluciones geoestadístico-productivas, dejando paso a futuro a nuevos niveles de decisión que involucren variables de proximidad a vías de comunicación/rutas (para facilitar tareas logísticas), cercanía a ciudades (para asegurar la existencia de mano de obra y de centros de distribución postventa), entre otras. El trabajo a futuro podría contener la variante de utilizar herramientas de código abierto (OpenStreetMaps) en R y sus propios paquetes orientados a usos geográficos (gstat, maps, etc.).

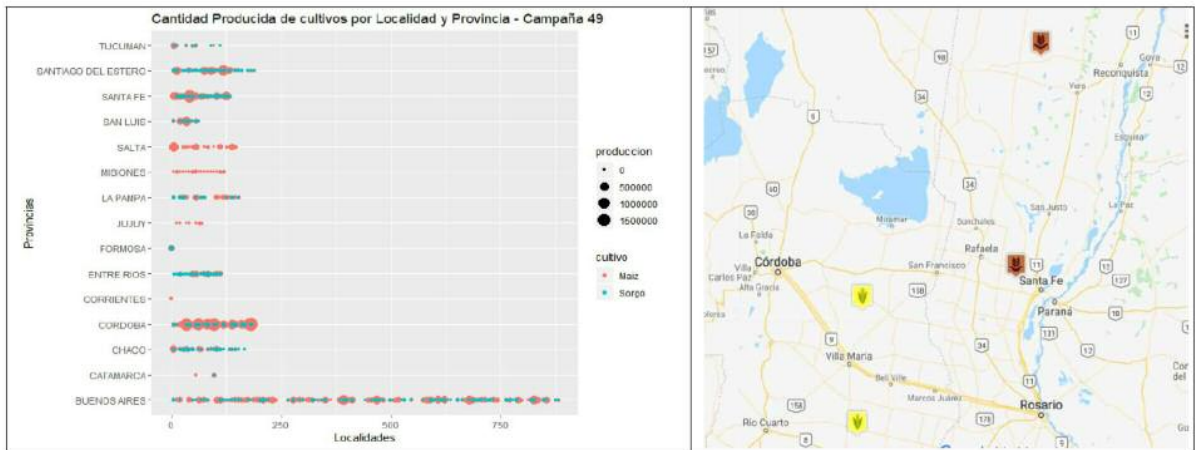


Figura 3. Peso productivo de maíz y sorgo por localidad y provincia (izq.) – Geolocalizaciones estadísticamente sugeridas para la radicación de la planta de bioetanol (der.)