Actas de LATIN**R** 2019

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

Editoras: Yanina Bellini Saibene, Florencia D'Andrea

Diseño Gráfico
Dis. Gráf. Francisco Etchart

Diciembre de 2019

Integrantes

Chairs

Laura Ación
- CONICET-Universidad de Buenos Aires
- R-Ladies Leadership Team
"https://twitter.com/_lacion_"
"https://lacion.rbind.io/"

Natalia da Silva
- Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República
- R-Ladies Montevideo, Uruguay
"https://twitter.com/pacocuak"
" http://natydasilva.com"

Riva Quiroga
- Facultad de Letras, Universidad Católica de Chile
- R-Ladies Santiago / Valparaíso, Chile
https://twitter.com/rivaquiroga

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 Córdoba
- Universidad de Costa Rica, Costa Rica

Ignacio Álvarez-Castro
- Universidad de la República-IESTA, Uruguay

Stephan Arndt
- University of Iowa, Estados Unidos

Mathias Bourel
- Universidad de la República-IESTA-FING, Uruguay

Xavier Buenaño
- Universidad Politécnica de Madrid, ETSI Ingenieros de Minas y Energía, España

Elena Chicaiza
- Instituto Panamericano de Geografía e Historia, Ecuador

Marina Cock
- CONICET-La Pampa, Argentina

Mario Cortina Borja
- University College London, Reino Unido

Andrés Farall
- Universidad de Buenos Aires, EcoClimaSol, Argentina

María Inés Fariello
- Universidad de la República-FING & Instituto Pasteur de Montevideo, Uruguay

Ileana Frasier
- CONICET-INTA - La Pampa, Argentina

Adriana Gil
- Universidad de la Pampa, Argentina

Juan José Goyeneche
- Universidad de la República-IESTA, Uruguay

Erin LeDell
- H2O.ai, Estados Unidos

Estefanía Mancini
- Centre for Genomic Regulation, España

Priscilla Minotti
- 3iA-UNSAM, Argentina

Leonardo Moreno
- Universidad de la República-IESTA, Uruguay

Germán Rosati
- IDAES-UNSAM / CONICET, Argentina

Marcelo Soria
- CONICET-Universidad de Buenos Aires, Argentina

Walter Sosa Escudero
- CONICET-Universidad de San Andrés, Argentina

Lucía Spangenberg
- Instituto Pasteur Montevideo, Uruguay

Melina Vidoni
- RMIT Computer Science, Australia

Comité Organizador

Yanina Bellini Saibene
- Instituto Nacional de Tecnología Agropecuaria
- R-Ladies Santa Rosa, Argentina

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

Paola Corrales
- Universidad de Buenos Aires, Argentina
- R-Ladies Buenos Aires, Argentina

Maria Florencia D'Andrea
- CONICET - Instituto Nacional de Tecnología Agropecuaria
- R-Ladies Buenos Aires, Argentina
- R en Buenos Aires

Patricio Cofré
- Metrics Arts, Chile

Patricia Loto
- Dirección de Sistemas de Lotería Chaqueña, Argentina
- R-Ladies Resistencia-Corrientes

Roxana Villafañe
- Universidad Nacional del Nordeste (UNNE), Argentina
- R-Ladies Resistencia-Corrientes

Gabriela Sandoval
- Pontificia Universidad Católica de Chile
- R-Ladies Santiago, Chile

Luis Verde
- Universidad Austral de Chile, Chile

Daniela Vazquez
- Data Scientist
- R-Ladies Montevideo

Beatriz Milz
- Universidade de São Paulo
- R-Ladies São Paulo

Prólogo

El 25 de octubre de 2017, Heather Turner anunció en el Slack de Organizadores de Grupos de Usuarios de R (RUG) que el Comité de Conferencias de la R Foundation estaba interesado en la aparición de eventos sobre R con un enfoque académico en regiones que no se encuentran actualmente cubiertas por useR!. En menos de una semana ya teníamos organizada nuestra primera videoconferencia para pensar cómo enfrentar el desafío. Esta rápida respuesta no fue solo suerte, sino la consecuencia de un año en el que la comunidad de R, a través de los Grupos de Usuarios de R (RUGs) locales y la presencia en redes sociales, creció significativamente en Sudamérica. Para mediados de noviembre ya estaba todo definido: un nombre, un lugar, una fecha y un comité organizador internacional motivado. La primera edición se realizó en Buenos Aires, Argentina y esta segunda en Santiago de Chile, Chile.

Los tópicos de interés del llamado a presentación de trabajos consistió en:

- Aplicaciones de R en distintas disciplinas de la academia y la industria. Todos los campos del conocimiento y sus combinaciones son de interés (por ejemplo, Ciencia de Datos, Estadística, Matemática, Informática, Ciencias Biológicas y de la Salud, Bioinformática, Geología, Ciencias de la Atmósfera, Ciencias Sociales, Humanidades, Educación, Economía, Periodismo de Datos, entre otras).
- Uso de R en conjunto con otros lenguajes de programación y plataformas
- Desarrollo de nuevos paquetes R
- Uso innovador de paquetes R existentes
- Uso de R en la enseñanza
- Iniciativas innovadoras para el aprendizaje de R
- Investigación reproducible usando R
- Análisis de grandes datos con R
- Aprendizaje automático con R
- Visualización de datos con R
- Análisis de redes con R
- Uso de R para análisis de datos abiertos

Este libro de actas presenta los 53 trabajos contribuidos, aceptados y presentados en la segunda edición. Los mismos están ordenados temáticamente de acuerdo a su presentación durante la conferencia y al formato seleccionado (presentación oral o poster).

Las contribuciones presentan estudios académicos y experiencias en la industria sobre temáticas de: Ciencia de Datos para las políticas públicas, comunidades, datos oficiales, datos abiertos, datos espaciales, datos de gobierno, elecciones y encuestas, género, manejo de datos, desarrollo de paquetes, R en producción, modelos predictivos, bases de datos, aplicaciones Shiny. Estas experiencias corresponden a varios países de Latinoamérica, dentro de los cuales se encuentran: Chile, Argentina, Uruguay, Brasil, Perú, Colombia y Ecuador.

Nuevamente en esta edición la participación y el entusiasmo de la comunidad de R en Latinoamérica queda de manifiesto en la cantidad, variedad y calidad de los trabajos y del networking regional realizado durante la conferencia, inicio de futuros trabajos y desarrollos conjuntos para la región.

Indice

Verónica Canales, Germán Mondragón, Mauricio Vargas,
Pablo León & Patricio Cofré

Resumen mesa temática: "Ciencia de datos para políticas públicas"

El uso de datos para la toma de decisiones en el sector público se ha visto revolucionado con las herramientas de big data y analytics. La posibilidad de procesar grandes volúmenes de información en un menor tiempo plantea nuevos desafíos para las instituciones públicas, las que hoy tienen la oportunidad de enfrentar antiguos problemas con nuevas herramientas.

Los casos que se revisarán permitirán conocer más de cerca cómo se han llevado a cabo soluciones de ciencia de datos para problemas de política pública, considerando la mirada desde la industria y desde agentes públicos, pues cada uno de estos actores es fundamental para la introducción de innovaciones en la administración pública. Asimismo, las presentaciones muestran la manera en la que herramientas de código abierto como R, se plantean como una alternativa importante para el desarrollo de innovaciones en el sector público, tendientes a mejorar la gestión del Estado.

Marco general de la situación de ciencia de datos en relación desarrollo de las políticas públicas

En la actualidad se habla de tópicos como Modernización, Transformación digital, Innovación, Ciencia de datos, Inteligencia artificial, asumiendo que tienen el mismo significado para todas las personas. Sin embargo, se hace necesario aclarar qué se entenderá por cada uno de ellos y cómo se relacionan entre sí. La presentación busca entregar elementos teóricos para entender los conceptos señalados. Esto, con el objeto de abordar las aplicaciones de la ciencia de datos; particularmente, la emergente utilización de R como software open source en la gestión pública. Lo anterior, desde un marco conceptual claro y compartido por los distintos actores involucrados en estas tareas.

En primer lugar, modernización considera los elementos en el plano físico: conectividad, servidores, máquinas, inter-operatividad. En estos casos, por lo general, nos referiremos al hardware involucrado. Cuando se comenzó con el impulso modernizador en el Estado justamente se abordó este tema como el principal.

En segundo lugar, transformación digital se relaciona con el proceso de digitalización de trámites. También se considera la incorporación de tecnología a los procesos que soportan la operación del Estado.

La innovación, por su parte, tiene relación con cambios en la manera en la que se hacen las cosas y con el modo en el que nos relacionamos. Esto puede venir acompañado de elementos tecnológicos o no. La innovación, como es entendida en instituciones como el Laboratorio de Gobierno, debiese estar centrada en las personas. En este marco conceptual, entonces, la ciencia de datos es comprendida como una disciplina que está al servicio de las personas.

Una segunda dimensión que se abordará en esta presentación son las implicancias políticas y éticas vinculadas a la ciencia de datos, por cuanto los datos y particularmente la ciencia de datos, puede ser entendida como una disciplina que logra representar la realidad únicamente de manera parcial. Al desarrollar modelos en base a datos históricos, se construyen modelos predictivos que reproducen su entrenamiento, es decir, se predice el futuro en base al pasado. El desafío propuesto es trabajar con los datos para modificar trayectorias y evitar reproducciones, es decir, prevenir o generar intervenciones tempranas. En este sentido, se propone dar una reflexión respecto a la utilización de los datos para uso empresarial, campañas políticas, etc. y cuestionarse sobre los límites del uso que estos tienen.

Luego, se mencionarán algunos esfuerzos impulsados desde el Estado, como Transparencia y Datos abiertos, con la mirada puesta en las implicancias que tiene para el Estado una mayor apertura en el acceso a los datos.

Finalmente, en los desafíos pendientes, se menciona la relación entre empresas y Estado en lo relativo al manejo de los datos y sus análisis. Se revisarán temas como datos abiertos, capacidades instaladas, software open source (R), entre otros.

Pablo León

Experiencias de trabajo con grandes volúmenes de datos en procesamiento de censo 2017

En los censos de población se recoge un volumen de datos importante. No todos estos datos cumplen con los requisitos de calidad establecidos, los errores pueden tener su origen en el informante, en el registro de los datos por parte del censista, en el reconocimiento óptico de los formularios, etc.

Dada la importancia que tienen esos datos para el desarrollo de políticas públicas, tanto a nivel nacional como a nivel local, y el alto costo de obtener la información, se hace un gran esfuerzo para aprovechar los datos obtenidos. Estos son revisados buscando errores o incoherencias. Una vez identificados conflictos en los datos, se busca ejecutar el mínimo de intervenciones que produzca un conjunto coherente, evitando descartar los registros.

Este proceso de validación y edición se aplica también en encuestas, muchas veces en forma manual, pero en los Censos, por el volumen de datos, solo es practicable por medios automáticos. En Censo 2017 por primera vez se introdujo el uso de R para la validación y edición de los datos. En esta presentación queremos transmitir algunas experiencias de adoptar R para este proceso, relatar lo que creemos fueron los mayores problemas y las soluciones que el proyecto elaboró para atenderlos. Principalmente en lo que se refiere al manejo de un volumen de datos importante, con sus efectos en los tiempos de procesamiento y la gestión del espacio en disco.

Los temas tratados se relacionan con:

- técnicas generales para aprovechar operaciones en que R es eficiente y evitar operaciones en que el performance de R no es el deseado,
- estrategia de particionamiento para permitir procesamiento paralelo,
- organización de la ejecución de procesos que permita maximizar el paralelismo, dentro de los límites de recursos de los equipos computacionales disponibles,
- selección de formato de datos en disco, balanceando la facilidad de manejo, el rendimiento y el uso de espacio.

También se discutirá el efecto de los requerimientos de reproducibilidad en la organización del trabajo y el uso de BBDD relacionales como repositorios de datos.

Creemos que la presentación será de interés no solo para quienes procesen encuestas y censos, sino para todos quienes requieran aplicar procesos de limpieza y depuración sobre volúmenes importantes de datos.

Mauricio Vargas

Uso de APIS y modelos de series de tiempo para estimar dotación de personal

Esta presentación conjuga 2 elementos relevantes para el desarrollo de la ciencia de datos en el sector público. Por un lado se encuentra la relevancia del modelo construido, en términos sustantivos; y por otro, el tipo de desafíos de infraestructura informática que es necesario enfrentar en muchos casos.

Parte del proceso de Reforma Procesal Civil (RPC), implementado por el Ministerio de Justicia y Derechos Humanos, involucra determinar la dotación de personal que será destinada a cada territorio jurisdiccional. Para llevar esto a cabo se contó con estudios de cargas de trabajo para los distintos perfiles (e.g. juez, administrador, administrativo de causas, entre otros) y procedimientos (e.g. Ejecutivo, Voluntario, Sumario y otros) además de los ingresos históricos por materia y juzgado proporcionados por la Corporación Administrativa del Poder Judicial.

Con estos inputs se procedió a ajustar un modelo de redes neuronales de dos capas usando la librería forecast. Pero surge un problema al momento de calcular el output del modelo, y radica en la capacidad del hardware existente en la administración pública.

Pese a que el problema a abordar difícilmente califica como big data, la falta de un servidor o equipos de escritorio adecuados llevó a tener que generar alguna solución alternativa. Esta solución fue crear una base de datos PostgreSQL y usando el paquete plumber se realizaron todos los cálculos del lado del servidor, y del lado del cliente bastó con un computador de escritorio con capacidad no mayor a la de un tablet promedio del año 2018.

El resultado del modelo es una salida JSON que indica con cuántos funcionarios se debe contar como mínimo para abordar las cargas de trabajo trimestrales durante todo el año para cada tupla juzgado-perfil-procedimiento, respetando la legislación laboral vigente. Por ejemplo, la solución que entrega el algoritmo, entre otros elementos, indica que el 3er juzgado civil de Santiago requiere N administrativos de causas el año 2020 para resolver todas las tareas relacionadas a los M ingresos en distintas materias que reciben durante el mismo periodo.

Entonces, el desafío enfrentado durante este proceso fue doble: llevar todo a DigitalOcean, que fue la plataforma que permitió sortear las dificultades de infraestructura informática del Ministerio; y luego proceder a experimentar modelos que respondieran a la necesidad de optimizar la dotación de personal en los territorios jurisdiccionales. La presentación abordará ambos desafíos en conjunto.

Patricio Cofré

Tres experiencias de éxito en el uso de la ciencia de datos para el sector público

Un cambio de paradigma muy notorio de los últimos años, es cómo la innovación dejó de provenir exclusivamente de la investigación desarrollada en universidades. En este sentido, la web y cloud han sido dos oleadas tecnológicas trasformadoras, lideradas desde la industria, que han dejado huellas permanentes en términos de metodologías, técnicas, estándares y tecnología.

Esta presentación aborda la manera en la que los cambios tecnológicos recién mencionados pueden contribuir a una mejor gestión del sector público. Uno de los cambios particularmente notables en términos de elaboración de políticas públicas ha sido el surgimiento de grandes volúmenes de datos, fenómeno que ha estado acompañado de la aparición de tecnologías que permiten procesar estas nuevas fuentes de información y sacar provecho de ellas. A partir del denominado Big Data y de las fuentes de información no estructurada es posible enfrentar nuevos problemas que no necesariamente pueden abordarse a partir de fuentes tradicionales, como encuestas y/o datos administrativos.

Para ejemplificar lo anterior se revisan 3 experiencias concretas: 1) las políticas de "cero papel", 2) la reforma al SERNAC y 3) el sistema anti evasión en el pago del transporte implementado por el Metro de Valparaíso. Respecto a la política de "cero papel" se describe una solución inteligente para detectar de manera automática las instituciones públicas que solicitan dentro de sus trámites alguna documentación en papel. Con esta información es posible generar un dataset que reúne la información producida por todas las instituciones públicas, lo cual disminuye la cantidad de trámites y evita la duplicación de ciertos procedimientos.

La segunda experiencia que se aborda se relaciona con la necesidad de fortalecer el rol fiscalizador del SERNAC. Se ha detectado que ciertos segmentos de la población no utilizan necesariamente los canales formales establecidos por el SERNAC para dirigir sus reclamos. En lugar de ello, utilizan las redes sociales. Considerando esta situación, se encuentra en fase de implementación un sistema automatizado que monitorea constantemente algunas redes sociales, con el objeto de detectar reclamos sistemáticos hacia ciertas empresas. En base a dicha información el SERNAC puede girar hacia una postura más activa en lo que respecta a su rol fiscalizador, sin tener que esperar a que la ciudadanía utilice los canales formales para llevar a cabo reclamos.

La última experiencia abordada corresponde a un sistema de reconocimiento facial implementado en el Metro de Valparaíso, que tiene el objetivo de detectar el mal uso que puedan tener los beneficios relacionados con el costo del pasaje. Así, lo que se intenta es reconocer si el beneficiario corresponde a la persona que efectivamente está utilizando el beneficio. Para ello se recurre a tecnología de visión computacional.

Las tres experiencias abordadas buscan motivar la discusión respecto a las potencialidades que tiene la ciencia de datos en lo relativo al mejoramiento de la gestión pública. Asimismo, se busca poner sobre relieve la capacidad que tiene la industria de generar innovaciones útiles para problemas de política pública.

Richard Detomasi & Gabriela Mathieu

Sistema para gestión de zonas a relevar en campo usando R

Introducción

Este trabajo presenta un desarrollo generado en el marco de la "Encuesta Habitar Urbano en Montevideo y Área Metropolitana"[1], con el fin de agilitar el sistema de asignación de zonas a relevar, sin necesidad del uso de mapas en formato papel. Se analizaron sus implicancias tanto a nivel de la coordinación de la encuesta, como del equipo de encuestadores, considerando el formato de relevamiento por medio de telefonía móvil.

Materiales y Métodos

Para la implementación específica de este desarrollo, se partió de la propuesta del equipo coordinador, que implicaba utilizar LimeSurvey 3.14.8 [LimeSurvey Project Team / Carsten Schmitz, 2012] y su servidor de almacenaje de las encuestas, pero sin contar con un servidor disponible para gestionar las bases de asignación de cargas. Se optó, para mejorar la flexibilidad y el trabajo colaborativo de quienes gestionaron el trabajo de campo, utilizar "Google Sheets" como repositorio de los registros de las asignaciones, y mediante la librería homónima [Bryan and Zhao, 2018] mantener actualizada la base de datos. Este sistema es implementable con gestores de base de datos más potentes (i. e. PostgreSQL-PostGIS).

El segundo componenete del sistema, corresponde a un visualizador "leaflet" [Cheng et al., 2018], incrustado en una aplicación "shiny" [Chang et al., 2019], que permite a quien asigna las cargas evaluar la disposición espacial tanto del equipo de encuestadores, como de las zonas a encuestar, pudiendo aplicar una función de exportación de la carga asignada en formato .kml enviándosela al mail del/la encuestador/a, mediante funciones de las librerías "maptools" [Bivand and Lewin-Koh, 2019] y "mailR" [Premraj, 2015] respectivamente.

El sistema se cierra mediante el relevamiento en campo por parte del equipo de encuestadores, que reflejan el progreso en el servidor del LimeSurvey al que se conecta el gestor de zonas mediante la librería de R "limeR" [Heiss, 2015].

Resultados

Como principales resultados, se destaca la excelente adaptación por parte del equipo de encuestadores al sistema, al mismo tiempo que el circuito demuestra la capacidad y flexibilidad de R en la implementación de esta clase de desarrollos a medida. La interfaz del gestor presenta un diseño sencillo, donde destaca la visualización mediante "leaflet" de la geolocalización tanto de las zonas disponibles para asignar, como de cada encuestador/a que necesite carga para trabajar. Como se muestra a continuación:

Referencias

- Roger Bivand and Nicholas Lewin-Koh. maptools: Tools for Handling Spatial Objects, 2019. URL https://CRAN.R-project.org/package=maptools. R package version 0.9-5.
- Jennifer Bryan and Joanna Zhao. googlesheets: Manage Google Spreadsheets from R, 2018. URL https://CRAN.R-project.org/package=googlesheets. R package version 0.3.0.
- Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. shiny: Web Application Framework for R, 2019. URL https://CRAN.R-pro-

1 Esta encuesta fue llevada a cabo por el Departamento de Sociología de la Facultad de Ciencias Sociales (UdelaR) con apoyo de la División de Políticas Sociales de la Intendencia de Montevideo, conformando un equipo coordinador del que formé parte.

ject.org/package=shiny. R package version 1.3.0.

- Joe Cheng, Bhaskar Karambelkar, and Yihui Xie. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library, 2018. URL https://CRAN.R-project.org/package=leaflet. R package version 2.0.2.
- Andrew Heiss. limer: A LimeSurvey R Client, 2015. R package version 0.1.
- LimeSurvey Project Team / Carsten Schmitz. LimeSurvey: An Open Source survey tool. LimeSurvey Project, Hamburg, Germany, 2012. URL http://www.limesurvey.org.
- Rahul Premraj. mailR: A Utility to Send Emails from R, 2015. URL https://CRAN.R-project.org/ package=mailR. R package version 0.4.1.

Gabriela Mathieu & Richard Detomasi

rECH: una caja de herramientas para procesar la Encuesta Continua de Hogares (ECH) de Uruguay usando R

Introducción

La Encuesta Continua de Hogares (ECH), es una encuesta que releva información de los hogares particulares uruguayos, realizada anualmente por el Instituto Nacional de Estadística (INE)[1]. Sus principales objetivos son obtener indicadores para monitorear la evolución del mercado de trabajo, el ingreso de los hogares y las condiciones de vida de la población (vivienda, salud y educación, entre otros). El paquete rECH -programado en R 3.4.3 [R Core Team, 2017]-, brinda un manejo fácil de los microdatos de la ECH, al permitir la obtención directa de indicadores por parte de personas no expertas, contribuyendo así a la democratización de la información pública y el acceso a los datos abiertos.

Materiales y Métodos

rECH permite la conexión a los microdatos de la encuesta desde 2006 hasta 2018 y a los valores del Índice de Precios al Consumo (IPC) de cada año. Para un manejo más accesible, los microdatos están estandarizados de manera que una misma variable tenga el mismo nombre y formato en todos los años. A su vez, el cálculo de un conjunto amplio de indicadores (mercado laboral, educación, ingresos, pobreza, salud, etc.) se obtiene ejecutando funciones específicas. Las funciones permiten obtener estimaciones puntuales y por intervalo a nivel de hogares o personas -según corresponda-, tablas, gráficos y mapas, así como la apertura por diferentes variables (departamento, región, sexo, etc.) relevantes siempre que la cantidad de casos sea suficiente. Algunos paquetes de R que se utilizan en este paquete son survey [Lumley, 2019, 2004] para las estimaciones e ineq [Zeileis, 2014] para los indicadores de distribución del ingreso, ggplot2 [Wickham, 2016] para gráficos y sf [Pebesma, 2018] para mapas.

1 Los microdatos de las ECH de cada año a partir de 1990 se encuentran disponibles en http://www.ine.gub.uy/web/guest/encuesta-continua-de-hogares1.

Prospectiva

Se espera que este paquete colabore en dar un acceso fácil a la ECH para el público en general, tanto con poco conocimiento en la construcción de indicadores, como del uso del lenguaje R.

También se espera continuar desarrollando esta herramienta, acompasando los cambios de la ECH, así como también incluyendo otras encuestas e indicadores. Una vez esté pronta la versión Beta, se disponibilizará en: https://gitlab.com/calcita/paquete-rech? nav_source=navbar

Referencias

- Thomas Lumley. Analysis of complex survey samples. Journal of Statistical Software, 9(1):1–19, 2004. R package verson 2.2.
- Thomas Lumley. survey: analysis of complex survey samples, 2019. R package version 3.35-1.
- Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 2018. https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.
- Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.
- Achim Zeileis. ineq: Measuring Inequality, Concentration, and Poverty, 2014. URL https://CRAN. R-project.org/package=ineq. R package version 0.2-13.

Análisis de la participación de la mujer en la informática aplicada al sector agropecuario en Argentina

Si bien la cantidad de mujeres en tecnología ha aumentado en los últimos años, todavía es mucho menor respecto a la cantidad de hombres. En particular, algunos estudios sugieren una baja participación de las mujeres en las tecnologías de la información y la comunicación (TICs) en todo el mundo. En Argentina, las TICs aplicadas al sector agropecuario y agroindustrial (AgroTICs) se vislumbran como una disciplina con gran potencial de desarrollo. No ajena a la tendencia mundial, se observa que la participación de mujeres en este campo en Argentina es mucho menor a la de los hombres, pero en nuestro conocimiento este fenómeno no ha sido cuantificado. Registrarlo y cuantificarlo es útil para efectuar un diagnóstico objetivo a fines de accionar en relación con este problema. En este trabajo proponemos explorar la participación de las mujeres en las AgroTICs a lo largo del tiempo. Para ello, analizamos los datos registrados en diez años del Congreso de Agroinformática (CAI), uno de los principales vehículos de divulgación científica de la investigación en Agroinformática en Argentina que se realiza desde 2008. Este congreso reúne trabajos que integran el sector tradicional de la producción agrícola con la agroindustria y la teleinformática, y está incluido dentro de las Jornadas Argentina de Informática. Determinar la forma y la intensidad de la participación de un grupo específico en una disciplina es complejo, pero analizar la participación de un grupo en un congreso informa sobre las interacciones y los comportamientos en la comunidad, reconoce a las personas y grupos clave involucrados e identifica áreas potenciales de mejora.

El CAI tuvo ocho ediciones con una convocatoria de artículos y dos ediciones sin convocatoria de trabajos (2012 y 2015); también tuvo una edición en conjunto con otro simposio (2008). Una persona puede participar en el CAI con los siguientes roles: presidente (chairs), comité científico, comité organizador, orador invitado, moderador y autor. Estos roles no son exclusivos, excepto para los roles de presidente y primer autor. Para este trabajo elaboramos un set de datos con cada participante en cada edición del CAI indicando el rol o roles que cumplió, su filiación, localización, título, tipo de trabajo y posición si fue autor (esta información está disponible en la web y en CD). La información extraída se utilizó para calcular el número total de participantes y clasificarlos por género y tipo de participación. Se utilizó R (3.5.1) y RStudio (1.0.143) y los paquetes readxl, dplyr, tidyr, ggplot2, maps y rworldmap. El código y los datos utilizados para el análisis se encuentran en https://github.com/yabellini/AnalisisCAI

El conjunto total de datos (para todos los años) tiene 1.689 observaciones, con 1.236 hombres y 438 mujeres participantes (15 sin definir). La participación de mujeres fue consistentemente menor que la de los hombres en todos los años del CAI y para todos los tipos de participación (Tabla 1 y Figura 1).

El análisis por año del rol autor muestra siempre una proporción menor de mujeres en el período estudiado, con porcentajes que van del 20% (año 2018) al 36.6% (año 2010) de las mujeres, con un promedio del 26% de mujeres autoras (figura 1). En cuanto al orden de autorías, la proporción de mujeres participando como primeras autoras en los trabajos nunca superó el 35% en todos los años analizados. Con respecto a la productividad de los autores, medidos en cantidad de trabajos presentados al congreso en todas sus ediciones, el promedio de trabajos es de 1,54 (desvío 1,25) para los hombres y 1,68 (desvío 2,24) para

Tabla 1. Número total y porcentaje relativo de participantes por género y rol (2008-2018)

Género	Chair	Comité Científico	Comité Organizador	Disertante	Moderador	Autor	Tota
Hombre	19 (83%)	246 (71%)	9 (64%)	115 (80%)	36 (77%)	811 (74%)	1236 (74%)
Mujer	4 (17%)	99 (29%)	5 (36%)	28 (20%)	11 (23%)	291 (26%)	438(26%)
Total	23	345	14	143	47	1102	1674

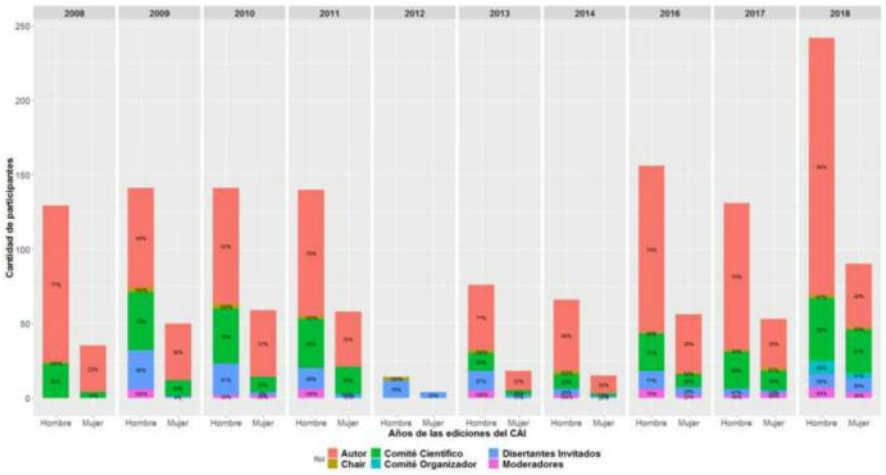


Figura 2. Número de participantes por año, género y rol. Los valores escritos en las barras corresponden al porcentaje de participantes con ese rol respecto al total de participantes de ese año y género.

las mujeres. Los hombres muestran un máximo de 10 trabajos presentados, mientras que para las mujeres el máximo es de 24 trabajos. El 73% y 75% de los autores hombres y mujeres respectivamente, han presentado un solo trabajo en el congreso.

Los datos que mostramos, si bien se limitan al análisis de un único congreso, sugieren que la participación de las mujeres en la disciplina AgroTICs es mucho menor que la de los hombres. Este podría ser un patrón general, que también ocurre en otros países y campos. Consideramos que es importante empezar a registrar estos fenómenos para contribuir a la toma de decisiones orientadas a modificarlo. En ese sentido, la organización del CAI, tomó una serie de medidas para aumentar la participación femenina, tales como: a) convocatoria de un comité científico más equilibrado (Figura 1, años 2016 a 2018), b) asegurar la misma cantidad de disertantes de uno y otro género (Figura 1, Año 2018), c) tener siempre, al menos una mujer en el rol de chair (Figura 1, años 2016 a 2018). El armado de un Código de Conducta, la difusión en redes sociales de los trabajos de las participantes femeninas en fechas claves como el 8 de Marzo, encuestas a los participantes al finalizar el evento y la publicación de este tipo de análisis, pretenden estimular la participación de más autoras femeninas en el CAI. Los resultados de productividad de los autores indican que también se deben realizar acciones tendientes a mantener en el tiempo la participación de los autores y aumentar el de autoras mujeres.

Juan Pablo Sokil, Laura Osorio Gómez

¿Quiénes investigan sobre género?

Introducción

La población mundial está dividida en partes casi iguales entre hombres y mujeres. Sin embargo, apenas existe un 29.3% de investigadoras[1]. Además de la brecha de género entre cantidad de investigadores, también existen sesgos dentro de las disciplinas científicas a la que se dedican, los datos corroboran que los hombres tienen una gran presencia en las ingenierías mientras que las mujeres se concentran en ciencias sociales y en ciertas áreas de la medicina. Los estudios de género han aparecido como una forma de visibilizar las diversas problemáticas de las mujeres, aunque con el paso del tiempo se han extendido a hacia otros grupos. Durante los últimos años han aumentado su participación dentro de la producción científica, la diversidad temática, sumado al creciente interés de la población y las agendas gubernamentales sobre la cuestión de género han sido algunos de los factores que han influido en este crecimiento.

Objetivos

Analizar la producción científica de estudios de género a nivel mundial respondiendo una serie de interrogantes: ¿Cuál es la participación de las mujeres y los hombres dentro de esta área de investigación? ¿Ha habido algún cambio en los últimos años? ¿Que temas investigan las mujeres? ¿Son los mismos que los hombres?.

Materiales

29.708 publicaciones científicas dentro del período 2008-2018, que forman parte de la colección principal de la base bibliométrica Web of Science y pertenecen a una revista cuya temática sean los estudios de la mujer, o que pertenezca al área de ciencias sociales y trate la temática de género[2].

Métodos

Para clasificar a los autores según sexo se utilizaron sus nombres. Para identificar las temáticas de estudio se extrajeron el resumen y título de las

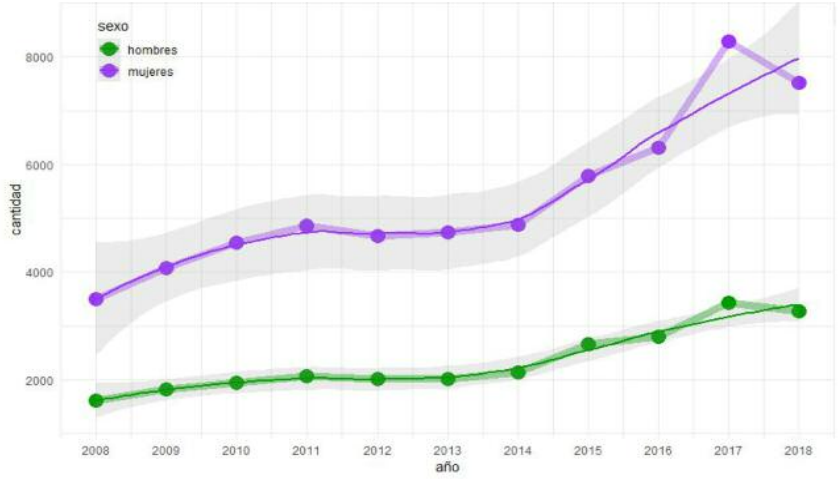


Figura 1: % de participación en publicaciones según sexo.

1 http://data.uis.unesco.org
2 Búsqueda: WC="Women's Studies" OR (TS='gender' AND SU='Social Sciences'), realizada el 22-03-2019

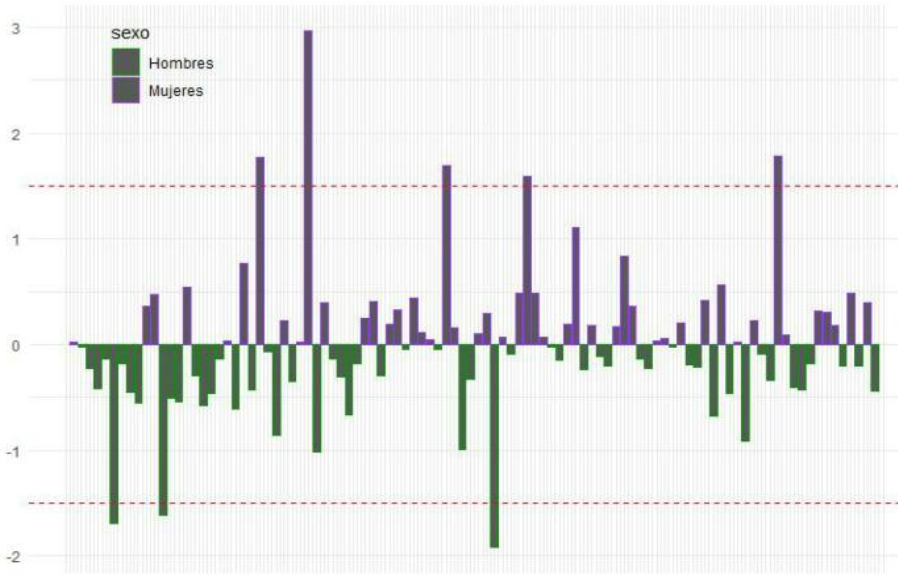


Figura 2: Diferencia porcentual de tópicos según sexo.

publicaciones. se pre-procesaron (lematizado y trigramas) y se aplicó Latent Dirichlet Allocation para identificar los tópicos.

Resultados

La cantidad de autores es constante a lo largo del período, fluctuando entre 69% y 71% de mujeres (Figura 1).

Dentro de los tópicos se identificaron una gran variedad de temas. Las publicaciones con autores hombres tienen mayor presencia en 3 tópicos: T7 - Análisis Estadísticos, T12 - Mercado laboral y T53 Deportes. Mientras que las publicaciones con autoras mujeres se destacan en 5 tópicos: T24 - Identidades, T31 - Feminismo, T47- Violencia de género, T58 - Violencia hacia la mujer y T89 - Salud reproductiva (Figura 2).

Conclusiones

Hemos observado que la investigación sobre estudios de género ha sido históricamente, y aún continúa siendo, un tema mayoritariamente de mujeres. Dentro de los tópicos de estudio se encuentran una gran diversidad de temas, lográndose identificar tópicos más vinculados a mujeres y otros a hombres (aunque con menor notoriedad).

Santiago Melendi, Marilina Santero, Carolina Prado, Rosana Poggio, Natalia Elorriaga, Laura Gutierrez, Pablo Gulayin and Vilma Irazola

Using the 'nasapower' package to merge open source meteorological data to individual-level health data from the CESCAS I cohort study in Argentina, Chile, and Uruguay

Introduction

Scientific evidence suggests that blood pressure (BP) could be related to weather patterns.[1] Increased BP in wintertime has been demonstrated in many studies; however, this hypothesis has not yet been explored in a general population sample from the Southern Cone of America.

Objective

To examine the effect of mean outdoor temperature on levels of systolic and diastolic blood pressure in a general population sample from 4 cities in Argentina, Chile, and Uruguay.

Methods

The CESCAS I study (https://estudiocescas.iecs.org.ar/en/) recruited 7524 men and women, aged 35 to 74 years from randomly selected samples in Marcos Paz and Bariloche (Argentina), Temuco (Chile), and Canelones (Uruguay) between February 2010 and December 2011. BP together with other individual-level demographic and health data was measured using validated instruments.[2]

The NASA Langley Research Center (LaRC) POWER Project funded through the 'NASA' Earth Science Applied Sciences Program (https://power.larc.nasa.gov/) features freely available global meteorology datasets. For the present analysis, daily temperature data for the study cities were obtained and the mean outdoor temperature for the day of the participants' BP measurement was used. The analysis consisted of two steps:

1-Merging datasets: meteorological data was fetched using the "NASA POWER API Client" (https://cran.r-project.org/package=nasapower)[3] of the R computing environment (Version 3.4.1, R Development Core Team, 2017). Coordinates from a single point within each city were used to obtain the mean outdoor temperature for each day of the study period, and this value was merged to the CESCAS I dataset, based on the date of BP measurement for each participant.

2-Fitting linear regression models: multivariate linear regression models were used to assess the association of mean outdoor temperature and season on levels of systolic and diastolic blood pressure. First, temperature was used as the independent variable to report changes in BP per 5 C unit increase in outdoor temperature, adjusting for age (in years) and sex (binary). Second, the mean difference between blood pressure in winter vs. summer was assessed according to study location, sex, age group, cardiovascular risk, history of hypertension and history of diabetes. To weight results accounting for the complex survey design of the CESCAS I study, the package "survey" was used.[4] Graphs were constructed using the "ggplot2" package.[5] Finally, "dplyr" was used for data processing.[6]

Ethical information: the study complies with

1 Modesti PA. Season, temperature and blood pressure: a complex interaction. Eur J Intern Med. 2013;24(7):604-607
2 Rubinstein AL, Irazola VE, Poggio R, et al. Detection and follow-up of cardiovascular disease and risk factors in the Southern Cone of Latin America: the CESCAS I study. BMJ Open. 2011;1(1). doi:10.1136/bmjopen-2011-000126
3 Sparks A. nasapower: NASA-POWER Agroclimatology Data from R_. R package version 1.0.0.9004. https://github.com/adamhsparks/nasapower
4 Lumley T. survey: analysis of complex survey samples. R package version 3.32. 2017
5 Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016
6 Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 0.7.6. 2018. https://cran.r-project.org/package=dplyr

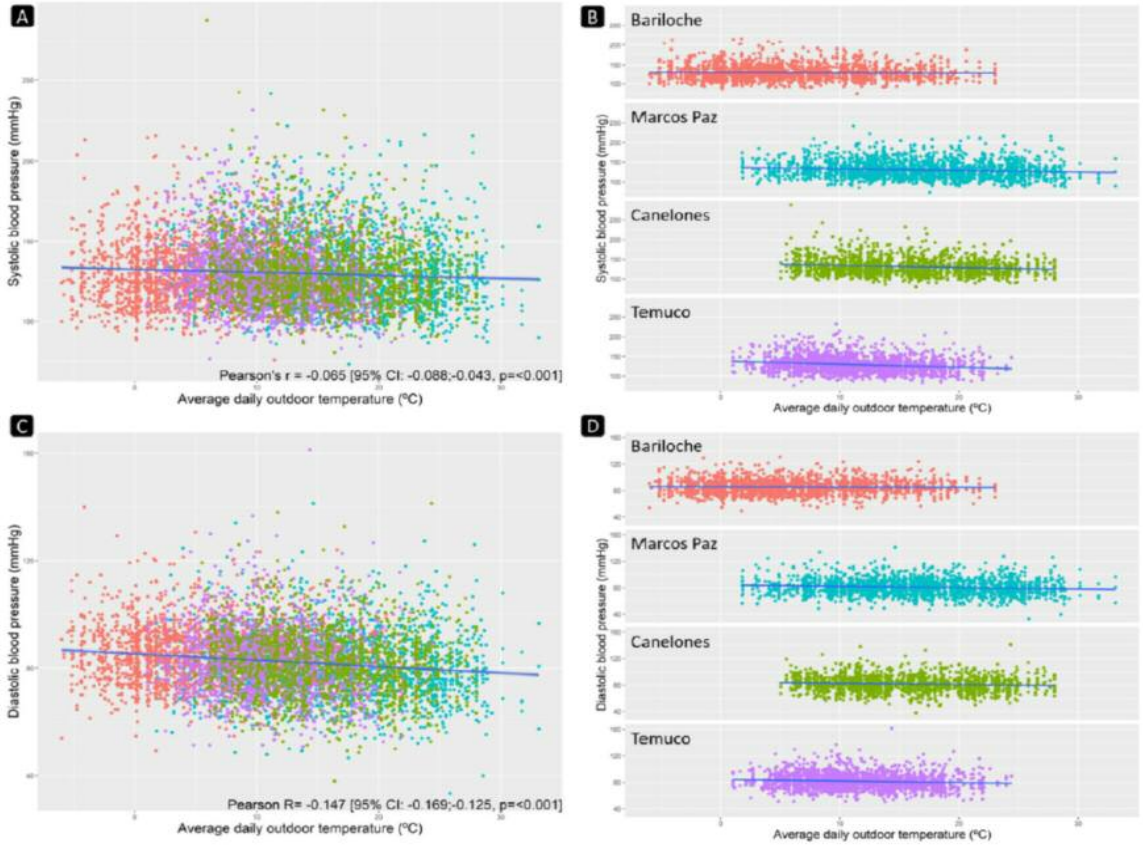


Figure 1. Relationship between average daily outdoor temperature (°C) and systolic (A & B) and diastolic (C & D) blood pressure in four cities from the Southern Cone of Latin America between November 2010 and December 2012. Least square linear method for the line of best fit.

the Declaration of Helsinki. Ethics approval was obtained from the Ethics Review Committees of Argentina, Chile, Uruguay, and the US. Written informed consent was provided by all study participants. There was no racial or gender bias in the selection of participants.

Results

Figure 1 depicts the relationship between outdoor temperature and BP measurements. Per 5 C increases in mean outdoor temperature, an average reduction of 0.64 mmHg (95%CI: 0.94; 0.35) was observed for systolic blood pressure (SBP) and 1.03 mmHg (95%CI: 1.22; 0.84) for diastolic blood pressure (DBP).

In winter, SBP was on average 3.66 mmHg higher than in summer and DBP increased by 1.55 mmHg. However, wide variability was observed among study cities. In Marcos Paz and Bariloche winter rises in SBP were above 6 mmHg while Bariloche showed no difference between seasons for both systolic and diastolic BP. Additionally, the effect of wintertime on BP was more marked for SBD than DBP and accentuated with older age, high cardiovascular risk and history hypertension or diabetes.

Conclusion

Blood pressure levels are associated with outdoor temperature in 3 of the studied cities (Temuco, Marcos Paz, and Canelones). BP levels of the adult population in these cities are higher during wintertime, which should be considered for clinical practice and policymaking. Interestingly, this association was not observed in Bariloche. Probably because despite the existence of seasons, year-long temperatures for Bariloche are consistently lower compared to all other cities.

Richard Vogg

About the presenter: Richard Vogg studied mathematics at the Technical University of Kaiserslautern in Germany. He later obtained a master's degree with a focus in Statistics and Machine Learning and learned to use R as a tool in academic applications. In 2018, he started his first job as a Business Analyst at Evalueserve in Viña del Mar. He currently works in the company's Customer Analytics area for one of the top 5 US banks by managed assets.

Fuzzy merging of company names

Datasets are increasingly large, and this brings along many challenges. Manually input data is especially problematic in that humans are prone to errors and do not always write or type in the same way. Take the answer to the question "Where do you work?" as an example. In our data, bankers wrote down the intended name "McDonald's" in formats as varied as: "McDonalds", "MCDONALDS CORPORATION", "McDonald's Corp", "mcdonlads" etc.

How can we match misspelled fields with their intended counterpart? How can we make sure that the right company name is being considered? This paper will shed some light onto how this problem was approached in a real business case delivered to a top 5 US-Bank by managed assets, one of Evalueserve's clients. Using the stringdist package in R, which implements several functions that measure the similarity of strings, we were able to accurately estimate what the intended input of the bankers had been.

We faced several hundreds of thousands of manually input company names and were interested in knowing whether we could identify clients in management positions at one of the S&P500 Companies in the US. We downloaded the S&P500 companies list as a reference of the correct company names. After cleaning the data, we decided to take the Jaro-Winkler distance between the strings. Unlike other distance measures, the Jaro-Winkler measure yields a higher score if one string is included in the other string (like "McKinsey" is included in "McKinsey & Company). Moreover, it can place more weight on the comparison of the first four letters of the strings.

The Jaro-Winkler distance is defined in two steps. First, the Jaro similarity sim_j between two strings s_1 and s_2 is defined as

$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else, \end{cases}$$

where m is the number of matching characters, $|s_i|$ is the length of string s_i and t is the number of transpositions of letters. The Jaro-Winkler similarity then adds weight to the first four letters of both strings:

$$sim_w(s_1, s_2) = sim_j + lp\left(1 - sim_j\right),$$

where l is the number of matching letters within the first four letters of both strings and $p \in [0,0.25]$ is a weight factor towards the first four letters. We define the Jaro-Winkler distance as $d_j = 1 - sim_j$.

We used the stringdist package to calculate the distance matrix of Jaro-Winkler distances between the hand-typed company names and the list of large US companies. We further determined column-wise minima, and – with these minima – the best fit from the large companies list. By simple thresholding, we selected a decision criterion for selecting the fits we wanted to keep.

The following screenshot is a toy example output with simulated data. "Company" shows the hand-typed company name, "Sim" gives us the minimum Jaro-Winkler distance found within the list of "unifying names", Best_fit is the related company name, and final is the output. If the sim value was lower than the chosen threshold (in this example, 0.15), then we would continue the analysis with the new "unified" company name.

The stringdist package is implemented in a way that it automatically parallelizes if possible. Given that the calculation of the distance matrices is by far the computationally most expensive step, having the option to parallelize makes a big diffe-

company	sim	Best_fit	final
Banco de Chile	0.17904762	Banco Santander	*NA*
McKnsiey and company	0.01750000	McKinsey and Company	McKinsey and Company
toyota	0.10000000	Toyota Motor	Toyota Motor
Volkswagen	0.07500000	Volkswagen Group	Volkswagen Group
Samsung	0.12631579	Samsung Electronics	Samsung Electronics
Apple Inc.	0.10000000	Apple	Apple

rence. We also tried to calculate the distances and best fits with the parallel package row by row which led to a minor improvement in runtime. As a reference: The final method took around two minutes to merge 500,000 company names to their potential counterparts in the S&P500 list – on a machine with the following hardware: Intel i5 (4 Cores) ~2.5 GHz - 8GB RAM.

We faced several challenges throughout the development of this project. Firstly, due to both the quantity of company names and the time limit, we could not implement a supervised algorithm. Secondly, we had to determine the business impact of false positives and false negatives and select the threshold accordingly. Lastly, we had to find ways to run the algorithm repeatedly on smaller chunks of data, as the stringdist package has size limits for the matrix.

In conclusion, the fuzzy merging algorithm helped us to identify 25k potential customers that have a management role in one of the S&P 500 companies. As a result, the bank's relationship with these customers can be optimized. R helped not only to solve this business problem in an effective manner, but also to conduct posterior analyses related to clustering customers.

As a side product, this analysis inspired new applications and ideas across teams within the bank. Fuzzy merging ultimately provides us with the opportunity to glean more valuable insights from applications for which hand-typed or not standardized data is the input.

Luis D. Verde Arregoitia

De la zoología a las aplicaciones generales, unheadR: un paquete para estructurar datos no ordenados

Al momento de formatear, almacenar, y compartir datos con colaboradores o con el público en general, existen ciertas prácticas comunes que pueden generar obstáculos para el posterior uso de estos datos. Generalmente son cambios que hacemos con fines estéticos, o ajustes menores para minimizar el espacio que ocupa una tabla en una página impresa o digital. Aunque sean cambios menores, pueden modificar la estructura de los datos. Por ejemplo: al insertar subencabezados dentro de otras variables para delimitar grupos de observaciones (Figura 1), o partir en dos o más filas los valores de la variable que contiene los nombres de cada unidad observacional (Figura 2). Los datos compactos y fáciles de leer no siempre son fáciles de usar en programas de análisis, por lo que es recomendable siempre compartir tablas y conjuntos de datos en estructura ordenada (en la cual cada variable tiene su propia columna, cada observación tiene su propia fila, y cada valor tiene su propia celda).

Cuando no hay opción y tenemos que trabajar con datos no ordenados que contienen subencabezados o valores rotos, es recomendable reorganizar estos datos de manera programática para ahorrar tiempo y además evitar introducir errores no reproducibles. Este trabajo describe el paquete unheadR (https://github.com/luisDVA/unheadr/), que incluye tres funciones simples que aprovechan las herramientas de 'tidy evaluation' para programar usando funciones de los paquetes dplyr y tidyr mediante evaluación de expresiones no estándar. El paquete fue diseñado para resolver estos obstáculos comunes en la estructura de los datos y generar conjuntos de datos ordenados. Las funciones fueron planeadas para usarse en el contexto de tidyverse, un conjunto de paquetes con una filosofía común y diseñados para utilizarse juntos.

Se describe el desarrollo y uso del paquete unheadR para la re-digitalización de cientos de tablas proporcionadas en artículos científicos sobre biología de mamíferos. Los subencabezados se utilizan ampliamente en ciencias biológicas

a)

Nombre Científico	Edo. Conservación
Leporidae	
Brachylagus idahoensis	Estable
Caprolagus hispidus	Amenazado
Ochotonidae	
Ochotona alpina	Estable
Ochotona iliensis	Amenazado
Ochotona princeps	Estable

b)

Nombre Científico	Edo. Conservación	Familia
Brachylagus idahoensis	Estable	Leporidae
Caprolagus hispidus	Amenazado	Leporidae
Ochotona alpina	Estable	Ochotonidae
Ochotona iliensis	Amenazado	Ochotonidae
Ochotona princeps	Estable	Ochotonidae

Figura 1. a) Uso de subencabezados para agrupar observaciones, b) Valores de la variable de agrupación en su propia columna. Los datos fueron ordenados utilizando unheadR con una expresión regular para identificar las familias taxonomicas dentro la la variable de nombres científicos, identificadas por el sufijo en latín 'dae'.

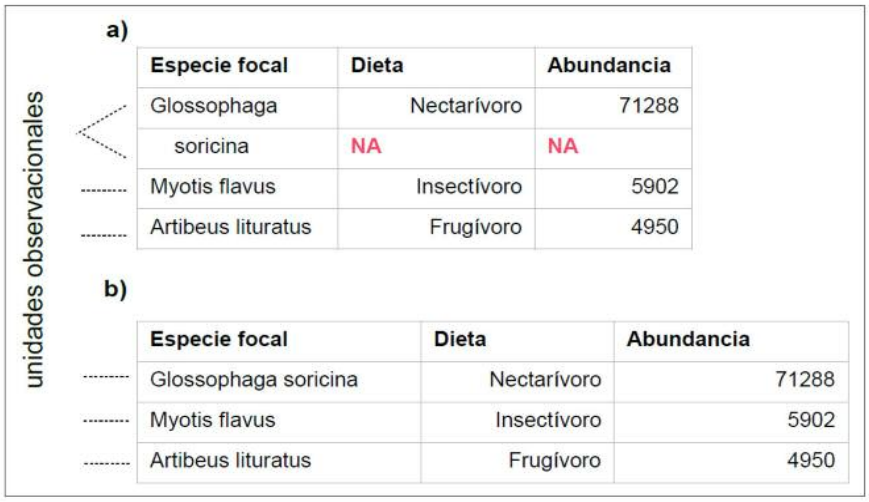


Figura 2. a) Tabla de datos con valores rotos en la variable que contiene las unidades observacionales (especies de murciélago) y valores en blanco, b) Presentación ordenada de los mismos datos mediante las funciones de unheadR.

para reportar grupos de especies que pertenecen a la misma familia o gremio ecológico, y representan un obstáculo significativo para la reutilización de datos taxonómicos. A partir de la implementación inicial del paquete, se definió una nomenclatura general de las diferentes estrategias para ordenar datos, y se delineó una estrategia general para aplicar estas funciones a todo tipo de datos.

Andrés Cruz

Inexact: un addin de RStudio para supervisar la unión fuzzy de bases de datos

"inexact" es un paquete y addin de RStudio que permite supervisar el proceso de fuzzy join, la unión automatizada de bases de datos con discrepancias en los valores de sus columnas comunes. La presente postulación se divide en dos partes: primero se describe el problema a resolver y luego se presenta "inexact".

Veamos los siguientes datos de ejemplo, en los que las observaciones corresponden a países (codificados sus nombres según estándares ligeramente distintos).

Una de las operaciones con datos más comunes es la unión izquierda: en este caso, querríamos añadir la columna "var_b" en "datos_c", para tener un solo data frame con la información de ambas variables. Sin embargo, aquí la columna en común, "pais", es inexacta entre las bases para las últimas dos observaciones, por lo que la unión no es completa.

Es posible solucionar este problema gracias a los algoritmos de pareo aproximado de texto (approximate string matching), que en R están implementados en el paquete "stringdist" (van der Loo et al., 2018). Dichos algoritmos calculan distancias entre textos de acuerdo a criterios como las letras en común, la longitud, etc. Volviendo al ejemplo, podemos cruzar los valores de la columna "pais" para ambas bases y construir una matriz con las distancias correspondientes (de acuerdo al algoritmo "osa" de alineamiento optimo de textos).

A partir de esta matriz se puede realizar un ejercicio corregido de unión de bases, implementado en el paquete "fuzzyjoin" de R (Robinson, 2018). Se comienza buscando, en cada fila de la matriz, el valor mínimo de distancia. Para los países del Cono Sur el pareo es exacto, por lo que sus valores en la diagonal corresponden a 0. Para "Brasil", cuya diferencia con "Brazil" es de sola una letra, la distancia asignada es de 1, lo que denota un pareo razonable. De esta forma, una unión ingenua realizada a través de "fuzzyjoin" — fuzzyjoin::stringdist_left_join(datos_c, datos_d, method = "osa", max_dist = Inf)— funcionará a la hora de añadir en la nueva base el valor de "var_b" para Brasil. El argumento max_dist = Inf

```
(datos_c <- data.frame(
  pais  = c("Argentina", "Chile", "Uruguay", "Bolivia", "Brasil"),
  var_a = 1:5, stringsAsFactors = F
))
       pais var_a
1 Argentina     1
2     Chile     2
3   Uruguay     3
4   Bolivia     4
5    Brasil     5
(datos_d <- data.frame(
  pais  = c("Argentina", "Chile", "Uruguay", "Bolivia (Plurinational State of)", "Brazil"),
  var_b = 11:15, stringsAsFactors = F
))
                             pais var_b
1                       Argentina    11
2                           Chile    12
3                         Uruguay    13
4 Bolivia (Plurinational State of)    14
5                          Brazil    15
```

```
dplyr::left_join(datos_c, datos_d, by = "pais")
       pais var_a var_b
1 Argentina     1    11
2     Chile     2    12
3   Uruguay     3    13
4   Bolivia     4    NA
5    Brasil     5    NA
```

```
matriz <- stringdist::stringdistmatrix(datos_c$pais, datos_d$pais, method = "osa")
rownames(matriz) <- datos_c$pais; colnames(matriz) <- datos_d$pais
matriz
          Argentina Chile Uruguay Bolivia (Plurinational State of) Brazil
Argentina         0     8       8                                27      7
Chile             8     0       7                                29      5
Uruguay           8     7       0                                30      6
Bolivia           7     6       7                                25      5
Brasil            7     5       6                                27      1
```

```
       pais var_a var_b
1 Argentina     1    11
2     Chile     2    12
3   Uruguay     3    13
4   Bolivia     4    14
5    Brasil     5    15
```

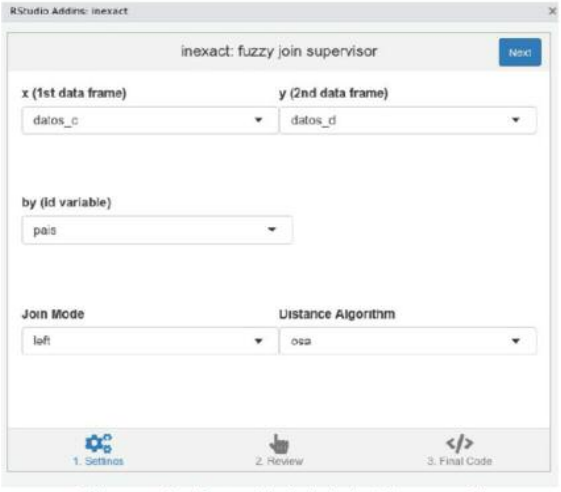


Figura 1. Panel inicial de "inexact"

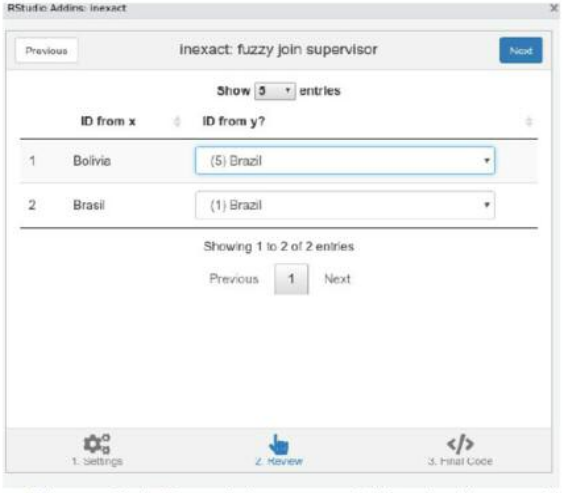


Figura 2.1. Panel de supervisión de "inexact"

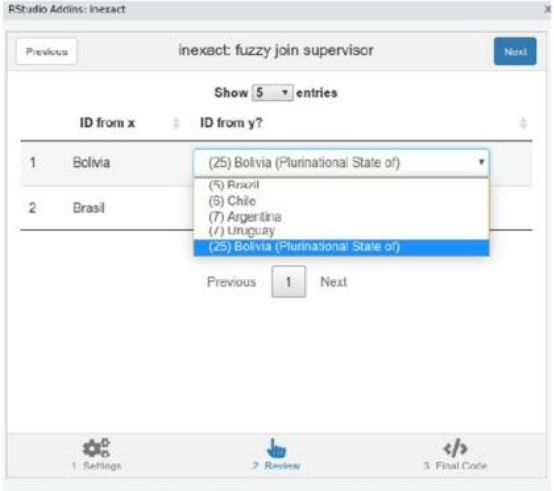


Figura 2.2. Panel de supervisión de "Inexact"

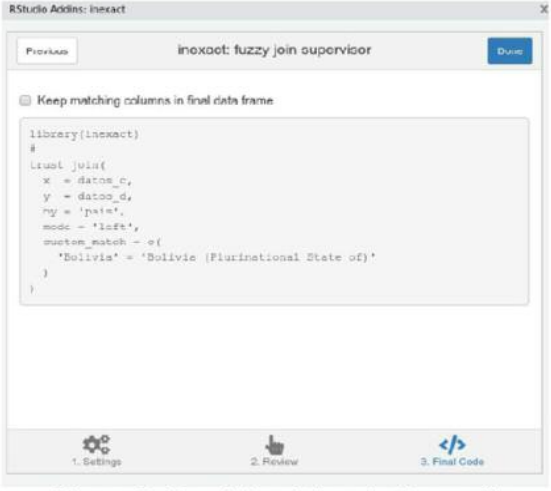


Figura 3. Panel de código de "inexact"

hará a la función completar la unión para cada fila, en cada caso eligiendo el valor mínimo de distancia en la matriz antes descrita.

Sin embargo, esta aproximación no es satisfactoria para el caso de Bolivia. El valor mínimo de distancia en su fila (5) está asignado para "Brazil", cuando sabemos que el pareo correcto es "Bolivia (Plurinational State of)", el valor máximo de distancia (25). En casos como este es que la supervisión humana puede complementar adecuadamente al enfoque automatizado.

Es posible hacer dicha supervisión utilizando los paquetes tratados hasta aquí, generando data frames intermedios que contengan más pareos que los de distancia mínima, para luego hacer filtros según el criterio del usuario(a) a la hora de inspeccionar los pareos visualmente. Sin embargo, este proceso es tedioso e ineficiente en términos de tiempo, en especial para bases de datos que no son pequeñas. "inexact" busca simplificarlo a través de un addin de RStudio, que provee una interfaz gráfica (GUI) especialmente pensada para supervisar la pertinencia de pareos aproximados. La Figura 1 muestra su ventana inicial, con las opciones específicas para la unión del ejemplo. La Figura 2.1 muestra el corazón de "inexact": el usuario(a) puede supervisar todos los pareos imperfectos, que se muestran desde el más conflictivo al menos conflictivo (decrecientemente según la distancia mínima detectada). Luego es posible editar casos específicos (Figura 2.2), siendo la opción por defecto siempre el pareo con menor distancia. La utilización de una GUI no afecta la reproducibilidad del procedimiento, pues su resultado final es código, como muestra la Figura 3: esta es una buena práctica apreciable en otros paquetes que implementan addins de RStudio, como “questionr” (Barnier, Briatte & Larmarange, 2018). El código que permite a “inexact” funcionar tras bambalinas proviene de “fuzzyjoin” (Robinson, 2018), por lo que solo actúa como una máscara en lo que refiere al trabajo computacional de la unión. Tras aplicar el código generado en el paso 3 se consigue la unión requerida, tras la aplicación del proceso automatizado y la supervisión humana.

Luciane Ferreira Alcoforado, Orlando Celso Longo, Lyncoln Sousa de Oliveira, Steven Dutt Ross and Alessandra Dos Santos Simão

AHP implementada em R

A tomada de decisão nas organizações tem sido objeto de constantes pesquisas e estudos comprovando a importância que este tema representa no desempenho dessas organizações. Segundo Gomes et al (2002), um sistema de apoio à decisão (SAD) é uma ferramenta computacional que envolve técnicas de sistemas de informação, inteligência artificial, métodos quantitativos, psicologia cognitiva e comportamental, sociologia das organizações, entre outros, e visam oferecer ao usuário condições favoráveis e acessíveis ao suporte, para de modo prático, melhor escolher uma entre diversas alternativas, minimizando assim a chance de erro na tomada de decisão.

O processo Analytic Hierarchy Process (AHP), baseado em matemática e psicologia, foi desenvolvido na década de 1970 pelo professor Thomas Saaty, Saaty (1970). O AHP pode ser classificado como o mais conhecido e utilizado dos métodos de análise multicritério cuja modelagem se divide em três etapas: construção dos níveis hierárquicos, definição das prioridades através de julgamentos paritários dos critérios estabelecidos e avaliação da consistência lógica dos julgamentos paritários. Neste trabalho apresenta-se a implementação do Método de AHP proposto por Saaty, utilizando-se a linguagem computacional R para automatização do método e apresentação dos resultados de maneira intuitiva.Foi criado um repositório no diretório github, contendo a estrutura do pacote do R, com as seguintes pastas:

- Documentação: arquivos em html para documentar o pacote;
- R: scripts com funções desenvolvidas em R;
- Data: arquivos de dados;
- Man: arquivos para documentos de ajuda para cada função do pacote;
- Vignettes: arquivos de ajuda do pacote

Em adição ao pacote, foi estruturado, a partir da construção hierárquica do problema, uma planilha de dados contendo as matrizes paritárias (ou de julgamento) de cada critério. Para os cálculos envolvidos no método foram implementadas as funções com as seguintes etapas:

- Ler os dados (matrizes de julgamentos fornecida pelo usuário);
- Calcular os pesos e a consistência;
- Retornar tabela com os pesos finais de cada alternativa, informando o índice de consistência dos julgamentos de cada critério considerado no problema.

O programa espera que o usuário tenha um arquivo xlsx contendo várias planilhas, onde cada uma representa uma matriz de julgamento. A estrutura do arquivo pode ser vista na figura 1. Como os especialistas são humanos, Saaty (1991) afirma que o AHP prevê que pode haver inconsistência, então o processo permite que haja uma inconsistência de no máximo 10% para que os resultados possuam credibilidade. Desse modo, a função principal do pacote analisará se cada matriz de julgamento inserida no arquivo é consistente ou não.

Atualmente o código conta com 15 funções totalmente desenvolvidas com o software R. Está hospedado na plataforma github no endereço: https://github.com/Lyncoln/AHP. A escolha da plataforma deveu-se ao fato de tornar o processo de colaboração acessível a todos os integrantes do projeto, além de permitir acesso, comentários e sugestões de não integrantes.

Para problemas com um único nível de critérios, o programa retorna uma tabela completa de proporções para as alternativas, indicando a melhor alternativa a ser escolhida, isto é, aquela que

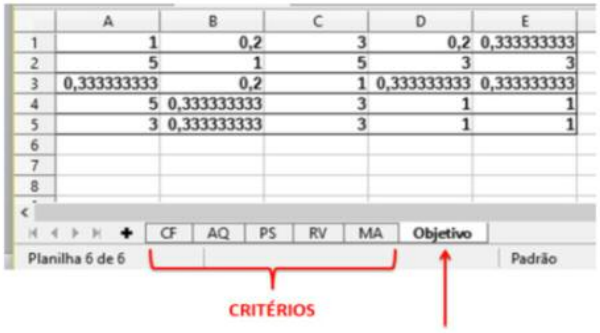
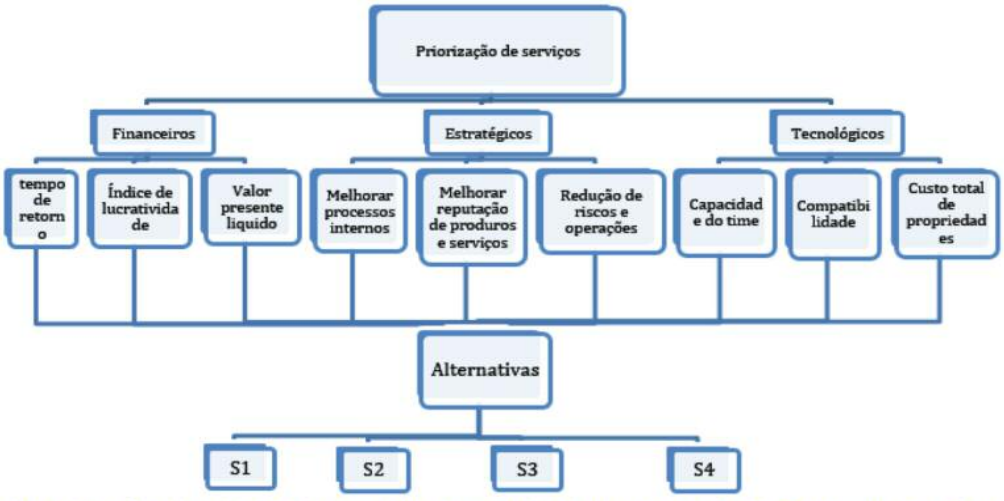


Figura 1 – Estrutura do arquivo de dados xlsx contendo 6 planilhas referentes às matrizes de julgamento. Fonte: Autores, 2019.

```
> tabela_ahp_xlsx("F://GitHub//AHP//Documentação//BD_teste1.xlsx")
[1] "A melhor escolha e a alternativa: A1"
# A tibble: 6 x 7
  criterio Pesos  A1     A2     A3     `Razao de consistencia de saaty` Consistente
  <chr>    <chr>  <chr>  <chr>  <chr>  <chr>                            <chr>
1 FP       100%   35.57% 34.8%  29.63% 7.03%                            Sim
2 AQ       44.75% 8.84%  21.95% 13.96% 4.62%                            Sim
3 CF       9.23%  0.96%  2.53%  5.74%  9.96%                            Sim
4 MA       18.56% 11.75% 1.97%  4.83%  3.32%                            Sim
5 PS       6.12%  2.51%  2%     1.6%   4.62%                            Sim
6 RV       21.35% 11.5%  6.35%  3.5%   0.79%                            Sim
>
```

Figura 2 – Ilustração da tabela de saída do pacote AHP para o exemplo da compra de carro. Fonte: Autores, 2019.




```
> tabela2_ahp_xlsx("F://GitHub//AHP//Documentação//BD_teste3.xlsx",mapeamento = c(1,3,3,3,3,4),
+                  alternativas = c("S1","S2","S3","S4"))
# A tibble: 4 x 2
  alternativas  Pesos
  <chr>         <dbl>
1 S1           0.0691
2 S3           0.152
3 S2           0.388
4 S4           0.391
>
```

Figura 3 – Ilustração de um problema em dois níveis de critérios. Fonte: Autores, 2019.

tiver a maior proporção da linha "Objetivo", conjuntamente com a validação dos julgamentos que são classificados como consistente ou não para cada critério. O exemplo do tutorial do pacote AHP baseou-se nos dados contidos em Costa (2002). O objetivo ou foco principal (FP) é a compra de um carro considerando-se 3 alternativas: A1, A2 e A3. Os critérios considerados nesta compra são AQ (custo de aquisição); CF (conforto); MA (custo de manutenção); PS (prestígio) e RV (preço de revenda). Com base nos dados fornecidos pelo decisor em formato xlsx (figura 1), o pacote efetua os cálculos necessários através da função tabela_ahp_xlsx( ) e retorna a tabela con-

tendo os pesos (proporções) de cada critério e do objetivo final para cada alternativa, bem como a razão de consistência dos julgamentos, informando se o mesmo é consistente ou não (figura 2).

Para problemas com mais de um nível de critérios (tipo composto), a função tabela2_ahp_xlsx( ) retorna como padrão um conjunto de proporçõ-es para as alternativas estudadas.

O exemplo de saída para a base de dados de um problema do tipo composto (GOMEDE, 2012) pode ser visto na figura 3. O objetivo é a priorização de serviços com critérios considerados em dois níveis, no nível 1 tem-se critérios financeiros, estratégicos e tecnológicos e no nível 2 cada critério do nível 1 é sucedido por mais 3 subcritérios em cada. As alternativas são em número de 4: S1, S2, S3 e S4. Observa-se o peso de cada alternativa, sendo a alternativa S4 a melhor escolha com peso de 0.39.

Como ações futuras será implementado generalizações nas funções já programadas para resolver problemas mais complexos com 3 ou mais níveis de critérios.

Referências

- Costa, H. G. Introdução ao método de análise hierárquica: análise multicritério no auxílio à decisão. Niterói, RJ, 2002.
- Gomes, L. F., Gomes, C. F. S., Almeida, A. T. Tomada de Decisão Gerencial: Enfoque Multicritério. Ed Atlas, SP, 2002.
- Gomede, Everton,. Miranda, Rodolfo. Utilizando o Método Analytic Hierarchy Process (AHP) para Priorização de Serviços de TI: Um Estudo de Caso, 2012. URL: http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2012/0041.pdf
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2018. Saaty, T. L. How to Make a Decision: The Analytic Hierarchy Process. European Journal of Operational Research, 48, 9-26, 1970.
- Saaty, T. L. Método de Análise Hierárquica. Rio de Janeiro: Makrom Books, 2Ed, 1991.

Auto-Keras: An R easily accessible deep learning library

Abstract

In the past few years, artificial intelligence (AI) has been a subject of intense media hype. Machine learning, deep learning, and AI come up in countless articles, often outside of technology-minded publications (Chollet and Allaire 2017). Transiting the age of data, it results fundamentally for any researcher with large amounts of information, to consider the application of deep learning models.

With a brief search on the web, dozens of texts are found where it is suggested to apply one or another deep learning model. However, tasks such as featurization, hyperparameters tunning, or network design, by no means, are easy for people without a rich computer science background. In this context, research work began to emerge in the area of what is known as Neural Architecture Search (NAS) (Zoph and Le 2016; Jin, Song, and Hu 2018). The main goal of NAS algorithms is to, given a specific dataset, search for the most optimal neural network to perform a certain task on that dataset. In this sense, NAS algorithms allow the user to not have to worry about any task related to Data Science engineering.

In other words, given a tagged dataset, and a task, e.g., Image Regression, Text Classification, among others, the NAS algorithm will train several highperformance deep learning models and return the one that outperforms the rest.

Several NAS algorithms were developed on different platforms, or as libraries of certain programming languages (Zoph and Le 2016; Jin, Song, and Hu 2018). However, for a language that brings together experts from such diverse disciplines as is the R programming language, there is no NAS tool to this day. In this paper, we present the Auto-Keras R package, an interface from R to the Auto-Keras Python library (Jin, Song, and Hu 2018). Thanks to the use of Auto-Keras, R programmers with few lines of code will be able to train several deep learning models for their data, get the best model and evaluate it.

For example, with the Auto-Keras R library, to train several deep learning models for the public MNIST dataset, and get the best-trained model, it results enough to run:

```
library("keras")
mnist <- dataset_mnist() # load mnist dataset
c(x_train, y_train) %<-% mnist$train # get train
c(x_test, y_test) %<-% mnist$test # and test data
library("autokeras")
# train an Image Classifier for 12 hours
clf <- model_image_classifier(verbose=TRUE) %>%
fit(x_train, y_train, time_limit=12*60*60)
# and get the best-trained model
clf %>% final_fit(x_train, y_train, x_test, y_test, retrain=TRUE)
```

In the present work, Auto-Keras was evaluated using the MNIST and CIFAR- 10 public datasets. For MNIST, after training for 12 hours, Auto-Keras tested 15 models and the best-trained model, obtains, for the test data, an accuracy value of 0.99. For the CIFAR-10 dataset, after training for 24 hours, it trained five models and the returned one got an accuracy value of 0.94. In this work, the Auto-Keras R package was presented. This library allows, with almost no deep learning knowledge, to train models and get the one that returns the best results for the desired task. Very promising results were obtained using Auto-Keras for the two datasets evaluated. Auto-Keras is an opensource R package and is freely available in https://github.com/jcrodriguez1989/autokeras/.

Although the Python Auto-Keras library is currently in a pre-release version and has few types of training tasks developed, it was recently added to the keras-team working group. This leads to an imminent advancement of this library.

References

- Chollet, F, and JJ Allaire. 2017. "Deep Learning with R. Manning Publications, Manning Early Access Program."
- Jin, Haifeng, Qingquan Song, and Xia Hu. 2018. "Auto-Keras: An Efficient Neural Architecture Search System." June 27, 2018. http://arxiv.org/abs/cs.LG/1806.10282.
- Zoph, Barret, and Quoc V Le. 2016. "Neural Architecture Search with Reinforcement Learning." arXiv Preprint arXiv:1611.01578.

alicer: Creando un paquete con soluciones analíticas para Walmart Chile

En el equipo de Customer Analytics de Walmart Chile hemos desarrollado un paquete de R llamado alicer (inspirados por la experiencia del paquete Rbnb de AirBnb[1]) con el fin de acelerar la entrega de análisis y productos digitales basados en datos, mejorar la calidad de estos, y facilitar la inducción de nuevos miembros al equipo.

Algunos de los beneficios más importantes que hemos conseguido en nuestro equipo gracias a la incorporación de un paquete de R interno en nuestro flujo de trabajo son:

Facilitar la conexión y consulta de múltiples fuentes de datos

Nuestro paquete incluye funciones que hacen más rápida y fácil la conexión a distintas fuentes de datos (Data Warehouse, bases de datos transacciones y APIs de servicios como Salesforce). Tenemos listados como dependencias tanto el paquete DBI, como los paquetes que proveen de drivers para conectarse a cada una de estas fuentes (odbc, RPostgres, etc). Además, usamos las funciones ui_ del paquete usethis para orientar al usuario sobre como guardar sus credenciales en .Renviron cuando se conecta por primera vez. El mapeo de las credenciales como variables de entorno facilita también escribir scripts que puede ejecutarse de forma no supervisada en contenedores Docker.

Asimismo, se han creado funciones que facilitan las consultas SQL desde R, actuando como wrappers de DBI::dbGetQuery, pero integrando además funcionalidades como el poder pasar parámetros o variables hacia las queries (mediante DBI::sqlInterpolate), leer queries directamente desde un archivo .sql, retornar siempre un output con clase tibble y, de forma opcional, "cachear" resultados anteriores en un archivo .rds para no tener que ejecutar con tanta frecuencia queries pesadas, pero mantener reproducibilidad de los resultados.

Facilitar la aplicación de técnicas econométricas de evaluación de impacto

En el equipo hemos recibido múltiples preguntas de negocio respecto de cómo impactan ciertas acciones corporativas en la venta de los locales (por ejemplo, inicio de cobro de estacionamientos, remodelaciones, y cese de entrega de bolsas plásticas) y que pueden abordarse mediante la técnica cuasi-experimental de diferencias en diferencias[2]. La aplicación de esta técnica requiere de varios pasos, entre los que se cuenta, en nuestro caso particular, obtener un dataset de venta por local en un formato determinado, calcular una matriz de correlación de pendientes para identificar locales con tendencias paralelas al local de interés, y luego hacer un análisis de regresión creando variables binarias con ciertos criterios.

Para ampliar el uso de esta técnica y permitir que todos los miembros del equipo puedan aplicarla (incluso quienes no tienen experiencia previa con ella) hemos creado una familia de funciones llamada evdid_ que ayudan al usuario a completar todas las etapas del análisis, entregando resultados intermedios que permiten realizar análisis derivados y validaciones (por ejemplo, excluir de la regresión a locales que sufrieron un evento exógeno por el cual ya no se cumple el supuesto de tendencias paralelas). También se provee de funciones que toman el resultado de la regresión de diferencias en diferencias y crean automáticamen-

1 https://peerj.com/preprints/3182.pdf
2 https://es.wikipedia.org/wiki/Diferencias_en_diferencias

te visualizaciones fáciles de leer para mostrarlas a audiencias del negocio y/o incorporarlas en reportes de R Markdown.

Agilizar la manipulación de datos y los análisis exploratorios

Hemos cargado algunas de las tablas de dimensiones más importantes que existen en nuestro Data Warehouse, tales como el árbol de productos vendidos por Walmart y la lista de locales/tiendas (en ambos casos acompañadas de una gran cantidad de columnas o atributos), dentro del directorio data/ del paquete.

De esta forma, un analista que ha importado datos en R usando una query SQL a un cierto nivel de agregación (por ejemplo, venta a nivel de local) puede rápidamente obtener métricas a nivel de comuna, región, u otra agrupación, sin tener que modificar su query, gracias a la gramática de dplyr en conjunto con estos datasets integrados.

Entrega de reportería ad-hoc mediante informes R Markdown parametrizados

Si bien nuestro equipo llevaba varios años usando Tableau para la entrega de múltiples métricas, existía también una gran carga de trabajo debido a solicitudes de informes ad-hoc que requerían escribir queries SQL manualmente, las cuales iban acompañadas de un flujo de trabajo orientado a Excel y PowerPoint, perdiendo replicabilidad y generando mucho espacio para el error humano.

Ahora hemos podido hacer más eficiente la entrega de esos informes gracias al uso de funciones de R que contienen queries SQL parametrizadas, y que se utilizan dentro de reportes de R Markdown, que quedan disponibles en nuestro GitHub corporativo para usarse de nuevo cuando nos llega una solicitud similar. Los reportes de uso más frecuente están disponibles dentro del mismo paquete como plantillas.

Además, dado que hay miembros de nuestro equipo que prefieren utilizar Python, hemos tomado ventaja del paquete reticulate para permitirles a ellos colaborar en estas plantillas, agregando secciones que se ejecutan sin problemas al hacer knitr de los reportes.

Inducción a nuevos miembros del equipo, y promoción de mejores prácticas y principios del tidyverse

Utilizamos vignettes y la documentación del paquete como formas de educar a nuevos miembros del equipo en el uso de R para resolver problemas concretos de analítica en Walmart. También procuramos que los ejemplos incluidos en la documentación ilustren conceptos como el uso del pipe para combinar funciones, y respeten la guía de estilo del tidyverse[3].

Asimismo, intentamos que nuestras propias funciones se guíen por el manifiesto de las tidy tools[4], abrazando el uso de la programación funcional, retornando estructuras de datos simples cuando siempre que sea posible (preferentemente tibbles), aspirando a que todas las funciones realicen muy bien una tarea concreta (en lugar de crear funciones enormes que realicen múltiples tareas a la vez) y que sean concatenables mediante el pipe para realizar tareas más complejas.

3 https://style.tidyverse.org/
4 https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html

Elio Campitelli

Si te gusta la estadística, bancate los metámeros

Introducción

En 1973 Frank Anscombe creó cuatro sets de datos que comparten la media y el desvío de cada variable y su coeficiente de correlación, pero que lucen muy distintos cuando se los grafica (Anscombe 1973). Desde entonces, el cuarteto de Anscombe se usa para ilustrar la importancia de visualizar los datos crudos en vez de confiar en los estadísticos sumarios. Sin embargo, no existe mucha investigación sobre el fenómeno general de "sets de datos disintos con iguales estadísticos" del cual el cuarteto de Anscombe es sólo un ejemplo. Además usar un conjunto de datos creados hace 50 años para enseñar da la impresión de que es un caso único o extraordinario.

En este artículo propongo el nombre de "metámeros estadísticos" en analogía al concepto de colorimetría y presento el paquete metamer, que implementa el algoritmo de Matejka and Fitzmaurice (2017) para la creación automática de metámeros.

Fundamentos de metamerismo estadístico

El Demonio de Laplace no sabe ni necesita saber estadística. Él puede conocer la posición y velocidad de cada partícula del universo y usar ese conocimiento para predecir su evolución. Pero los seres humanos no podemos analizar más de unos pocos números por vez. Si queremos entender el universo tenemos que resumir grandes cantidad de observaciones en unos pocos números. Necesitamos saber estadística.

La mayoría de los métodos estadísticos buscan representar grandes cantidades de datos con unos pocos números interpretables, lo cual implica una reducción de la dimensionalidad. Intuitivamente, parecería que no se puede representar N números conMnúmeros menor que N, aunque esta propiedad debe demostrarse para cada método estadístico. Se pueden usar las identidades de Newton y los polinomios simétricos elementales para demostrar que se necesitan N momentos para describir unívocamente una muestra de tamaño N[1]. Como colorario, existen infinitos sets de datos de N observaciones que comparten los mismos n < N momentos.

Voy a llamar "metámeros" de un determinado set de datos al conjunto datos que comparten el valor de una transformación estadística. El nombre surge por analogía al metamerismo en colorimetría; el fenómeno por el cual nuestros ojos perciben el mismo color a partir de espectros distintos al reducirlos a los mismos 3 números (rojo, verde y azul). Es decir, toda transformación estadística no inyectiva tiene metámeros. El Cuarteto de Anscombe es un ejemplo dramático, pero no debe entenderse como aplicable sólo a los momentos estadísticos. Ninguna transformación

estadística representable como una función contínua es inyectiva si reduce la dimensión del problema (Malek et al. 2010). Como en general se busca que datos similares tengan transformaciones similares, el metamerismo es una consecuencia inevitable de los métodos estadísticos. No es un bug, es una característica. Tampoco debe concluirse que visualizar los datos sea la única solución. Al proyectar los datos en un espacio bidimensional se pierde información y, como la silueta de un sombrero que puede confundirse con la de una boa digiriendo un elefante, los gráficos también sufren de metamerismo.

1Técnicamente unívocamente a menos de una permutación.

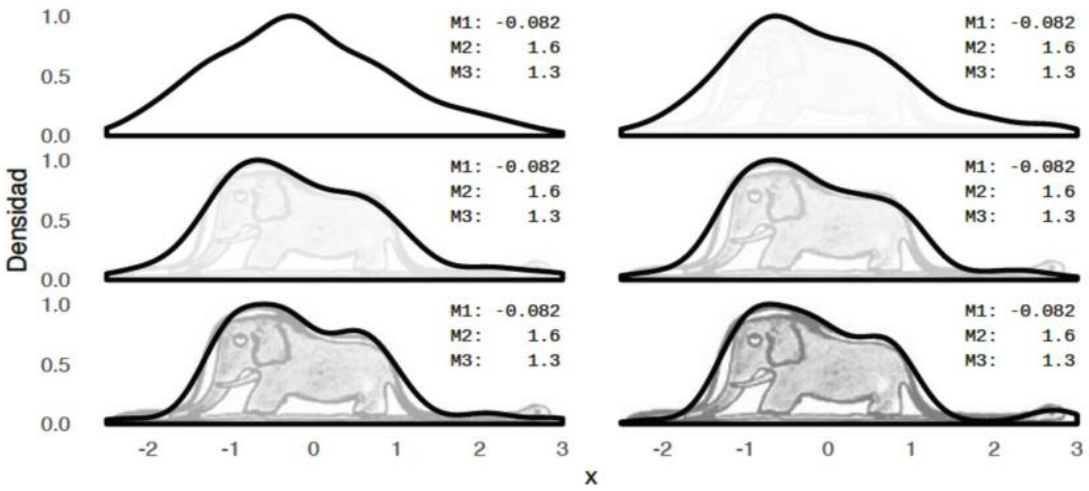


Figura 1: Densidad de probabilidad de 6 metámeros. Todas comparten los primeros tres momentos no centrados hasta 2 cifras significativas (M1, M2 y M3).

```
library(metamer)
metameros <- metamerize(data.frame(x = rt(200, 5)), # dataset inicial
                        preserve = moments_n(1:3),  # función a preservar
                        minimize = elefante,        # función a minimizar
                        annealing = FALSE,          # todos los metámeros minimizan `minimize`
                        N = 250000)                 # número de iteraciones
```

Cómo crear metámeros

El paquete metamer implementa el algoritmo de Matejka and Fitzmaurice (2017) para generar metámeros. Perturba un set de datos iterativamente, verificando que se preserve la transformación estadística de interés y, opcionalmente, que se minimice una función. Al ser completamente genérico, permite ilustrar el metamerismo de cualquier transformación. El siguiente código genera sets de datos que comparten los primeros 3 momentos al menos con 2 cifras significativas pero cuyas distribuciones son substancialmente distintas (Figura 1).

Referencias

- Anscombe, F. J. 1973. "Graphs in Statistical Analysis." The American Statistician 27 (1): 17–21. https://doi. org/10.2307/2682899.
- Malek, Freshteh, Hamed Daneshpajouh, Hamidreza Daneshpajouh, and Johannes Hahn. 2010. "An Interesting Proof of the Nonexistence Continuous Bijection Between Rn˜ and R2˜ for N != 2." arXiv:1003.1467 [Math], March. http://arxiv.org/abs/1003.1467.
- Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17, 1290–4. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Uso de R en ambiente productivos

R como lenguaje es cada día más usado en el ámbito corporativo y no solo académico, sin embargo dentro de las corporaciones muchas veces es importante que los modelos desarrollados queden en ambientes productivos teniendo que cuidar entonces temas como: el performance (cuando se trata de un sistema que debe calificar en tiempo real), Tiempos de procesamiento y posibles overflow de memoria (calificar toda la población de una entidad bancaria). Es por eso que a continuación se explicará como abordar estos problemas sin tener que reescribir los desarrollos R en otro lenguajes como Python, Java, etc.

En este caso se expondrá el uso de herramientas complementarias a R como son h2o y Docker. H2o es un Framework de Inteligencia Artificial, que posee muchos algoritmos tradicionales y no tan tradicionales, con la ventaja de que su procesamiento se hace de manera paralelizada e incluso se puede ejecutar en un cluster Spark, este framework para los conocedores de Caret les puede ser un simil. Por otro lado Docker es un proyecto de código abierto que automatiza el despliegue de aplicaciones dentro de contenedores de software, proporcionando una capa adicional de abstracción y automatización de virtualización de aplicaciones en múltiples sistemas operativos. Para tal motivo se presentarán dos posibilidades, donde claramente se pueden combinar las mismas para hacer aún más eficiente R.

1. R + h2o: Para este escenario se usará R más una versión de h2o (Framework de AI, https://www.h2o.ai/). La idea es procesar toda la data mediante cualquier esquema, realizar cualquier análisis previo a un modelo de ML y posteriormente realizar el modelo bajo h2o desde R. Ya con un modelo generado sobre h2o, el mismo puede ser extraído como objeto y ser cargado y consumido desde otro R, sin embargo, siempre tendremos las limitantes de versiones de R y h2o (caso que veremos más adelante), más los tiempos de respuesta. Por eso se recomienda que el modelo generado sea descargado como un objeto Java (POJO o MOJO). De esta manera se puede compilar sobre cualquier arquitectura y así poner en producción de manera muy rápida y sobre todo eficiente. Bajo este esquema se genera una clase Java que prácticamente no depende de nada salvo un ambiente Java que como bien sabemos hoy día todos los sistemas operativos poseen una versión open u no open.

2. R + h2o + Docker: En este escenario vamos a pensar que no se trata solo del modelo ML, si no que se necesita que el flujo de predicción conlleve algo de tratamiento previo de data, sobre esto y pensando en el escenario anterior se pudiese pensar en trasladar todo el código a Java, sin embargo, ya se pierde el sentido de usar y programar en R. Para esto la posible solución sería crear un contenedor (Docker) con R y el modelo en h2o no como objeto Java si no como objeto de R. De esta manera toda la aplicación, los códigos, y en fin todo el desarrollo cuidando librerías y dependencias externas a R se pueden mantener sin ningún problema. Ese contenedor se puede extraer del computador del creador y ser puesto en otro ambiente sin ningún problema siempre y cuando ambos sistemas cuenten con el respectivo Docker. Ni siquiera hace falta que ambos tengan R, h2o o cualquier otra librerías o dependencia. Claramente con este esquema se debe tener cuidado extra si el tiempo de respuesta debe ser acorde a una transacción en línea (milesimas de

segundos). Allí se debería pensar en usar esquemas de APIS con plumber o temas que no trataremos en esta ocasión. Hoy día en empresas de Buró como es Experian estás metodologias se están adoptando con fuerza, de está manera los desarrollos analíticos pueden llegar a implementarse sin problemas. Ya sea en arquitecturas On Premise u arquitecturas Cloud.

Camilo Gómez-Narváez, Moreno Bevilacqua & Víctor Morales-Oñate

GeoModels: un paquete para el análisis de datos geoestadísticos gaussianos y no-gaussianos

El análisis de datos espaciales se ha convertido en un área de desarrollo muy amplia en distintos campos disciplinarios como la minería, ciencias atmosféricas, geología u otras como la planificación vial y de ciudades basada en datos espaciales o el sistema de información geográfica (GIS). Particularmente, desde la metodología respecto al análisis de este tipo de datos, la Geoestadística ha tomado un rol importante en la modelización y soporte a la toma de decisiones basadas en información que, en muchas ocasiones, es de extrema importancia tomar una decisión correcta.

La investigación en los últimos diez años ha introducido modelos Geoestadísticos de campos aleatorios o procesos espaciales que se utilizan para la modelización particular de datos que tengan asociada una localización espacial o una georreferenciación. Con estos modelos, han surgido también diversos paquetes computacionales que permitan la aplicación de modelos Geoestadísticos, pero la mayoría de estos paquetes y modelos solo consideran el campo aleatorio Gaussiano por lo que es necesario recurrir a la aplicación de transformaciones a los datos para obtener características similares a la gaussianidad o normalidad, teniendo que asumir con que los resultados obtenidos conducirán probablemente a un modelo que no capture toda la información adyacente a la naturaleza de los datos espaciales, además de un modelo con falta de interpretación física el sus parámetros.

Si bien una herramienta importante para la modelización de datos espaciales ha sido el campo aleatorio Gaussiano debido a muchas propiedades, en la práctica, asumir una distribución Gaussiana o Normal en los datos es un supuesto no muy creíble ya que los datos pueden tener características como asimetría, kurtosis, soporte positivo, soporte compacto e incluso pudiendo ser datos discretos. Es por esta razón que el desarrollo de un paquete de acceso gratuito para datos geoestadísticos no-Gaussianos se vuelve una tarea que resolver, en particular el paquete en desarrollo de R, GeoModels, se vuelve una interesante alternativa en el proceso de modelización y estudio de datos espaciales con diversas características en los datos como los detallados en la Tabla 1.

En esta presentación se abordará la estructura metodológica para la simulación y estimación con GeoModels. Primero se mostrarán los aspectos referidos a la simulación de campos aleatorios Gaussianos y no-Gaussianos (Tabla 1) aludiendo a la flexibilidad que entrega GeoModels respecto a otros paquetes de R dado que es posible considerar modelos que sean continuos y diferenciables en media cuadrática debido a la implementación que se ha hecho en el paquete junto con la posibilidad de utilizar una gran cantidad de modelos de covarianza que recientemente han sido propuestos únicamente en la literatura para modelos espaciales o espacio-temporales uni y bivariados tanto en el plano como en la esfera, pero que en GeoModels si se pueden encontrar.

En segundo lugar, se mostrarán las rutinas para la estimación de parámetros de los modelos que, a diferencia de los paquetes existentes, en GeoModels se utiliza una aproximación frecuentista del método de máximo-verosimilitud, pero computacionalmente más eficiente (verosimilitud ponderada compuesta a parejas). El paquete también considera la opción de realizar estimación por máximo-verosimilitud si es que el modelo lo permite.

Así como el paquete contempla las herramientas para el análisis estadístico de datos georreferenciados, también existen otras utilidades como

Soporte de los datos	Modelos y características particulares
1. Continuo	a. Gaussiano b. Gaussiano asimétrico (posible asimetría) c. T de Student (colas pesadas) d. Log-Gaussian (soporte positivo y asimetría) e. Gamma (soporte positivo y asimetría) f. Weibull (soporte positivo y asimetría) g. Wrapped-Gaussian (datos direccionales)
2. Discreto	a. Probit Espacial (datos binarios balanceados) b. Binomial (datos de conteo) c. Binomial Negativo (datos de conteo)

Tabla 1: modelos disponibles en el paquete de R GeoModels.

lo son las funciones NuisParam y CorrParam que permiten una implementación más amigable por parte del usuario ya que por ejemplo estas herramientas entregan los nombres de los parámetros necesarios de los campos aleatorios o funciones de covarianza en una etapa anterior a la simulación o estimación de modelos.

Se finalizará esta presentación con una breve reseña a nuevos modelos para campos aleatorios no-Gaussianos (discretos y continuos) así como también opciones de cómputo eficiente que ofrece GeoModels para implementaciones más rápidas (GPU y en paralelo).

Matthew Upson, Matthew Gregory, Duncan Garmonsway and Matthew Dray

Reproducible Analysis in the UK Government

A key function of government statisticians is to produce official statistics for publication. Often these statistics have a direct impact on government policy, so it is imperative that they are accurate, timely, and importantly: reproducible. At any point in the future, we should be able to reproduce all the steps required to produce a statistic, but manual processes (common in many official publications) can make this challenging. This presentation is about a new approach to producing official statisticsusing the R language that was trialled in the UK Government, and has become a major success story.

Taking inspiration from the fields of reproducible research and DevOps, an approach called 'Reproducible Analytical Pipelines' (RAP) was trialled in the Department for Digital, Culture, Media, and Sport (a central UK Government department). The proof of concept involved taking an existing, highly manual, official publication and creating a reproducible pipeline that allowed the publication to be reproduced instantly. A clear audit trail and high standard of quality assurance was maintained by using version control, test driven development, and continuous integration. Finally the pipeline (comprising a bespoke R package and Rmarkdown document) was published as open source software, allowing complete transparency in how the statistics were produced.

We publicisied this proof of concept with a blog post , and were able to follow up with a second blog post as the project was adopted in other government departments a few months later. We also produced an ebook (written in bookdown), a website , and a massively open online course to help spread the message in the UK Government and elsewhere. Within a year, the project had been endorsed by the UK Statistics Authority , the body responsible for assessing the accuracy of official statistics. As of April 2019, the project has been so successful, that there are now 22 RAP projects under way and 42 champions across 24 UK Government bodies . We even have a hex sticker.

In this presentation we will share some of the technical details behind Reproducible Analytical Pipelines, and how we built a community of champions to take the work forward. We hope that this presentation will serve as an example of how innovation can be fostered within a very large and traditionally conservative organisation, and how modern tools and practices can be used to reduce repetitive and onerous manual labour, improve accuracy, and reduce the time taken to produce official statistics.

Story-telling interativo com dados de despesa pública: uma análise sobre os Estados brasileiros

As story-telling interativas permitem a junção de uma síntese narrativa com a exploração e análises avançadas de dados pelo usuário final. No R isso é feito utilizando ao mesmo tempo Rmarkdown e Shiny. O presente trabalho mostra uma construção que apresenta essa abordagem. Trabalha-se aqui com a evolução de dados de despesa pública, notadamente despesas de pessoal, no escopo da federação brasileira. Os dados utilizados são os produzidos para envio de relatórios ao FMI seguindo uma metodologia dessa instituição. A história começa com a possibilidade do usuário montar dinamicamente indicadores de despesa de pessoal e analisar em um mapa de calor o comportamento desse número para todos os Estados brasileiros, ao longo do período de 2010 a 2018. As opções dadas ao usuário permitem que faça análises sobre até doze indicadores diferentes. No gráfico gerado tonalidades vermelhas revelam indicadores mais elevados e o contrário para tonalidades azuis. Inicialmente o nível crítico que divide as cores é dado pela mediana dos valores apresentados. O usuário pode em seguida modificar esse valor. A narrativa dirigida começa em seguida e analisa três indicadores dos doze possíveis. Encontra-se clusters de Estados formados pelas séries temporais do indicador formado pelo total de despesas de pessoal no critério competência dividido pelo total de receita. Os dois clusters formados são validados em dois gráficos distintos onde fica claro que o algoritmo PAM utilizado gerou uma escolha razoável dos componentes de cada um dos agrupamentos. Em seguida mostra-se uma nova versão do mapa de calor com a evolução temporal do indicador com o foco nos dois clusters formados. A figura 1 mostra esse gráfico.

Pela Figura 1 o leitor percebe claramente que no cluster 1 o indicador apresenta valores elevados desde o início da série temporal, enquanto que o cluster 2 há um comportamento que revela um início azul e que aos poucos vai tomando tonalidades vermelhas, sem chegar ao final da série temporal no mesmo nível do cluster 1.

A análise temporal dos clusters é submetida a dois novos testes gráficos. Dessa vez pelo agregado anual. Um dos gráficos mostra os dois clusters comparados simultaneamente ano a ano através de box-plot. Fica claro nesse caso pela distribuição dos valores que ao longo dos anos o cluster 1 sempre teve os valores posicionais mais elevados do que o cluster 2. O outro gráfico é em formato de violino e mostra também a distribuição ao longo dos anos, mas nesse caso o cuidado foi de comparar se as médias anuais de cada um dos clusters eram significativamente diferentes. Para isso se usou testes ANOVA. Os violinos de cores iguais apresentam também médias estatisticamente iguais. A função do gráfico em violino foi permitir deixar mais claro ano a ano a distribuição para cada um dos pontos possíveis do indicador. O leitor pode ficar mais confortável com essa visualização em aceitar os resultados das comparações de média dada pelo teste ANOVA. Para as análises seguintes o indicador inicial é decomposto em dois outros indicadores: despesas com pessoal ativo em relação ao total de receita e despesa com pessoal inativo em relação ao total de receita. Para esses dois indicadores novamente são feitas comparações das médias anuais usando o teste ANOVA. Os quatro gráficos montados para essa análise contam uma história em que para os dois clusters o que vem se destacando é o aumento das despesas com pessoal inativo em relação ao total das receitas. O estado do Amapá não se enquadra dentro das análises anteriores e rece-

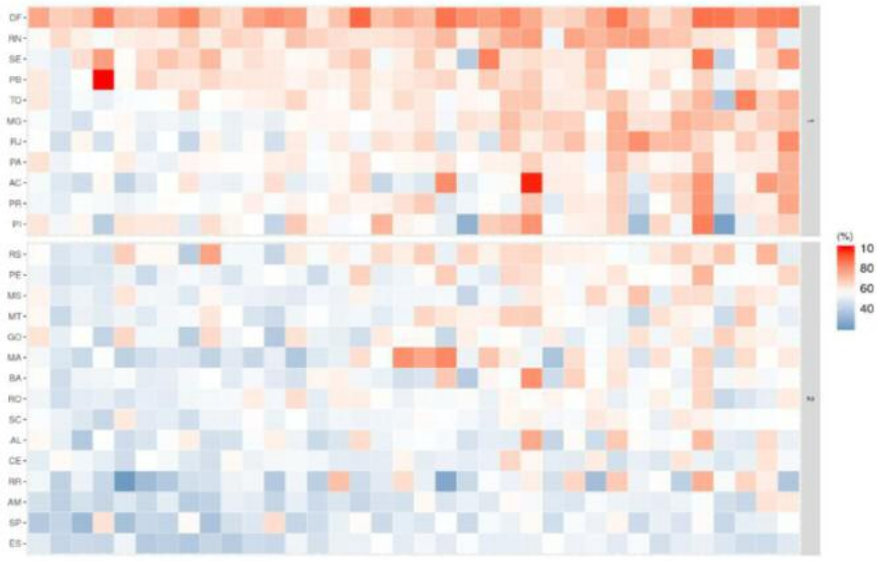


Figura 1: Mapa de calor dos clusters formados pelas séries temporais

beu uma sessão específica onde se analisa as quatro séries temporais que podem dar origem aos indicadores trabalhados nessa história. Dois gráfico, um de linha e outro de barras empilhadas trazem visualizações do comportamento dessas séries. Por fim, a história retorna a possibilidade de interação com usuários. Agora o leitor pode ter acesso a construção de cenários das composições futuras dos clusters através de predições utilizando ARIMA. Para cada um dos estados, o usuário pode escolher o percentual sobre o valor máximo ou valor mínimo das predições ou acreditar no valor mais realista. Essa possibilidade deixa na mão do usuário a elaboração de vários conjuntos de cenários baseados em suas próprias expectativas de evolução do indicador para cada um dos estados. A história se conclui aí. A aposta é que usuários partam de uma exploração inicial de dados, tome contato com uma estrutura narrativa que conduz uma visão específica sobre os dados e se conclui com uma análise mais avançada que permite elaboração sofisticiada de cenários. Esses recursos narrativos e analíticos abrem possibilidades para um maior controle social e incrementos em educação fiscal. Por estar em código aberto, espera-se que a partir do exemplo e em breve pela disponibilização dos dados abertos, ocorra o incentivo a novas ressignificações por parte de amplos setores da sociedade civil. O produto está disponível no seguinte link: https://fabdev.shinyapps.io/analise_clusters_despesas_pessoal_estado/

Inteligencia Electoral: comportamiento y campañas

Introducción

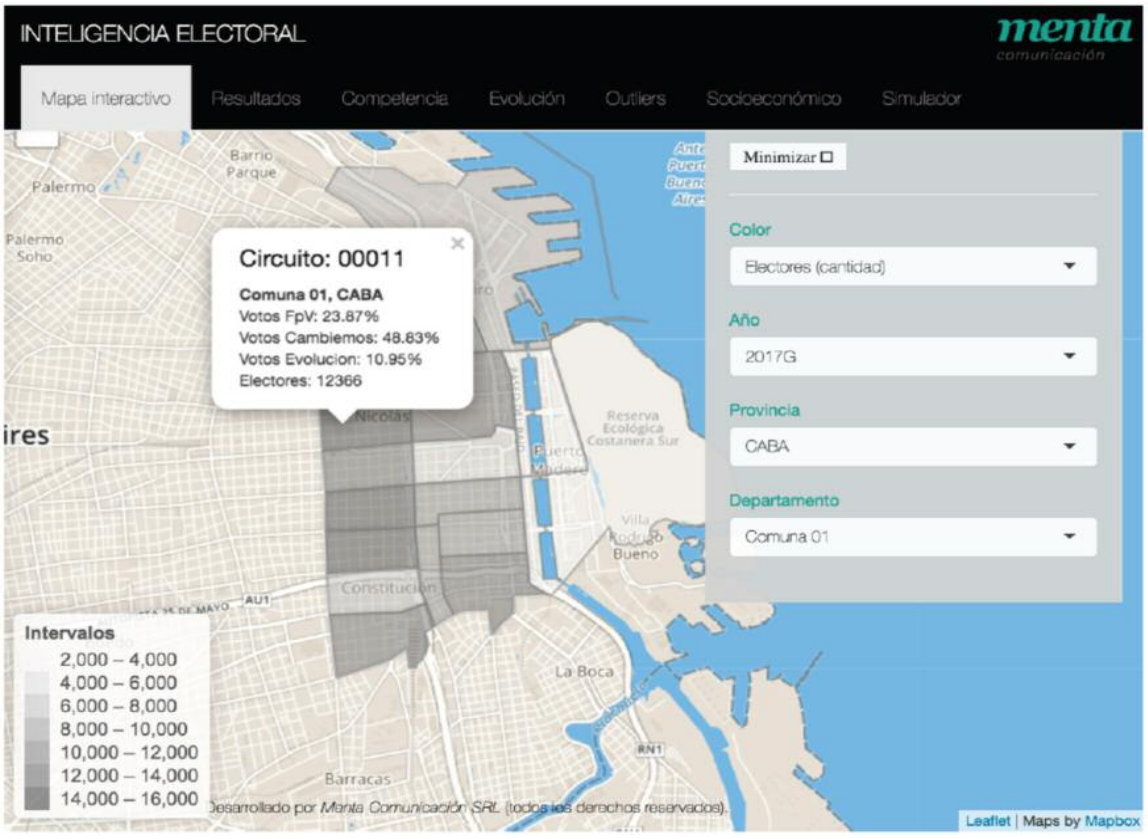
La disponibilidad cada vez mayor de información electoral detallada, junto a una expansión acelerada de recursos y herramientas de fácil acceso para procesarlos abren una oportunidad hasta hace poco tiempo inédita, tanto para estudios académicos como para quienes trabajan en los procesos electorales. La posibilidad de acceder casi en tiempo real a los resultados de las elecciones y cruzarlos con información externa da lugar, entonces, a maneras novedosas de utilizar los datos. Bajo este contexto se desarrolló una herramienta de análisis interactivo de grandes volúmenes de información electoral que permite generar nuevos insights a partir de la detección de patrones y características específicas de los procesos.

Inteligencia Electoral

Este desarrollo permite abordar el comportamiento electoral en (a) distintos niveles de agregación: desde los resultados por mesa -para evaluar valores excepcionales en la distribución de resultados y participación- hasta los datos por departamento, pasando por la georreferenciación de información para cada circuito electoral; y (b) la posibilidad de realizar comparaciones temporales. La aplicación propone una serie de visualizaciones que permiten entender territorios claves a la hora de explicar los resultados electorales agregados incluyendo la distancia entre fuerzas y el peso electoral de los territorios. Partiendo de información publicada por la Justicia Electoral -a nivel nacional y subnacional- y de publicaciones parciales de información geográfica se desarrolló una plataforma enteramente desarrollada en R. Estas herramientas constituyen, a nuestro juicio y según el conocimiento presente respecto del análisis electoral, insumos novedosos para la diagramación y el despliegue de campañas electorales tanto en el territorio como en otros ámbitos (por ejemplo, el digital).

Resumen

Fig. 1: El mapa interactivo permite viajar a lo largo del distrito analizado resultados y comportamientos electorales en la geografía. Todos los datos desagregados a nivel circuito - fuentes de las visualizaciones- son presentados en Resultados, con la posibilidad de aplicar filtros. La pestaña Competencia presenta dos visualizaciones: (a) la primera permite observar rápidamente la distancia porcentual de votos, pudiendo seleccionar del menú distintas listas, procesos electorales y unidades geográficas. (b) La segunda brinda una idea más cabal del peso geográfico de esas diferencias a partir de un gráfico de árbol (treeplot). En Evolución consiste en un gráfico de pendientes para analizar los resultados en el tiempo. Outliers presenta dos niveles de observaciones atípicas al nivel de las mesas al interior de los circuitos electorales de cada elección: observaciones con desvíos moderados y severos para porcentaje de votos de listas, votos negativos y nivel de participación. Incluimos también Un análisis gráfico que relaciona el nivel Socioeconómico con los resultados electorales a nivel circuito. Por último, un Simulador electoral a nivel departamental.


INTELIGENCIA ELECTORAL
menta
comunicación
Mapa interactivo
Resultados
Competencia
Evolución
Outliers
Socioeconómico
Simulador
Minimizar
Color
Electores (cantidad)
Año
2017G
Provincia
CABA
Departamento
Comuna 01
Circuito: 00011
Comuna 01, CABA
Votos FpV: 23.87%
Votos Cambiemos: 48.83%
Votos Evolucion: 10.95%
Electores: 12366
Intervalos
2,000 – 4,000
4,000 – 6,000
6,000 – 8,000
8,000 – 10,000
10,000 – 12,000
12,000 – 14,000
14,000 – 16,000
Palermo
Barrio Parque
Palermo Soho
Reserva Ecológica Costanera Sur
Constitución
La Boca
Barracas
AU1
RN1
Desarrollado por Menta Comunicación SRL (todos los derechos reservados).
Leaflet | Maps by Mapbox

Package "lobbyR"

Actualmente, en Chile y el mundo existe una amplia discusión sobre la difusión y transmisión de datos públicos, entre ellos, se encuentra el "Lobby" o "gestión de intereses". El presente trabajo pretende desarrollar la creación de un package con funciones simplificadas para la extracción de datos de audiencias de Ley del Lobby mediante el requerimiento a la API existente de la plataforma y además, la suma de funciones que permitan realizar web scraping sobre el registro de audiencias de lobby en el Congreso de Chile. Por último, se desarrollará un análisis de redes entre lobistas y congresistas para detectar comportamiento generalizado entre partidos y coaliciones, detección de comunidades y medidas de centralidad.

¿El Problema?

En Chile, la regulación de esta materia se inició desde el 8 de Marzo de 2014 cuándo se publicó la Ley N 20.730, que regula el lobby y las gestiones de intereses de particulares ante autoridades y funcionarios. El objetivo que se contemplaba era suministrar un avance para la transparencia de la actividad pública y su ejercicio.

El acceso a la información detallada de registros por ley lobby se encuentra disponible mediante la conexión a una API. La API de la Ley del Lobby es una interfaz diseñada para programadores que permite integrar los diversos registros de audiencias en servicios e instituciones públicas. La obtención de datos requiere adjuntar algunos Header como un "Api-Key", el cuál se puede obtener en la plataforma principal de Ley del Lobby (https://www.leylobby.gob.cl/admin/perfil). La obtención de los datos es en la extensión ".json " como se muestra en la Fig.1

En este contexto, se detectan dos dificultades: 1) Un acceso ciertamente dificultoso y engorroso para obtención rápida de información y 2) La estructura de los datos puede resultar dificultosa para realizar algunos análisis desde resumen de estadísticas descriptivas y visualización. Cabe mencionar que a la fecha se registran cerca de 347.000 registros de audiencias lo que incurre en una necesidad de automatizar ciertos procesos para el análisis.


```
{
                                    "nombres": "Rodrigo",
                                    "apellidos": "Mora",
                                    "cargo": "Presidente Comisión Defensora Ciudadan
a y Transparencia",
                                    "referencia": "Modificación de resolución que de
niega información por parte del COMPIN 1º región",
                                    "forma": "P",
                                    "lugar": null,
                                    "comuna": null,
                                    "fecha_inicio": "2014-10-27 15:00:00",
                                    "fecha_termino": "2014-10-27 17:00:00",
                                    "sujeto_pasivo_url": "/cargos-pasivos/12",
                                    "institucion_url": "/instituciones/157",
```


Fig. 1 Estructura datos de Audiencias. Fuente: www.leylobby.gob.cl

Registro de Audiencias

Sujeto Pasivo	Fecha	Lobbista representado	Lugar	Materia
José Miguel Castro Bascuñán	30-04-2019 12:00:00	Ormazabal Hijos Ltda. ORMAZABAL ARANCIBIA	Oficina Parlamentaria Antofagasta, Región de Antofagasta	Problemáticas en nuestra planta de revisión técnica.
MARCIA OYARZUN LEIVA (Asesor Dip. Andrés Celis)	30-04-2019 10:00:00	PAOLA DANITZA MORA CORRALES	OFICINA CONGRESO DE VALPARAISO Valparaíso, Región de Valparaíso	TEMA A TRATAR SOBRE LA EDUCACIÓN Y DIFERENCIA DE SUELDOS SOLICITA ASESORÍA LEGAL
Renzo Trisotti Martínez	30-04-2019 9:30:00	Obesidad sin Fronteras	Av. Héroes de la Concepción 2648 Iquique, Región de Tarapacá	Apoyo a la agrupación, fiscalización de dineros del Gobierno a las operaciones detonadas a la obesidad.

Fig. 2. Estructura audiencias Lobby en Congreso. Fuente: https://www.camara.cl

Por otra parte, el registro de audiencias por ley del lobby que contempla a la autoridad parlamentaria (Cámara de Diputados) sólo exhibe las audiencias en su página web sin permitir una transmisión efectiva para los análisis que se pretendan desarrollar (Ver Fig 2). Esto impide una descarga directa de los datos(a la fecha hay cerca de 2000 registros de audiencias entre miembros del congreso), teniendo que recurrir a técnicas como web scraping.

Soluciones y trabajo realizado

Se han implementado funciones determinadas para la extracción de datos de audiencias desde la API con la utilización de "jsonlite", "tidyverse", "httr" y por otra parte, la extracción de datos desde la página web se ha desarrollado con "rvest", "Rselenium", "tidyverse", "plyr", "ggplot2" y "ggpubr". Esto ha permitido el estructuramiento y visualización de datos que se pretenden integrar mediante funciones al package que se pretende crear.

Por último, en la Fig. 5 se muestra un avance en análisis de redes del congreso por coalición y como se encuentran interconectados de acuerdo con los lobistas que han desarrollado audiencias.

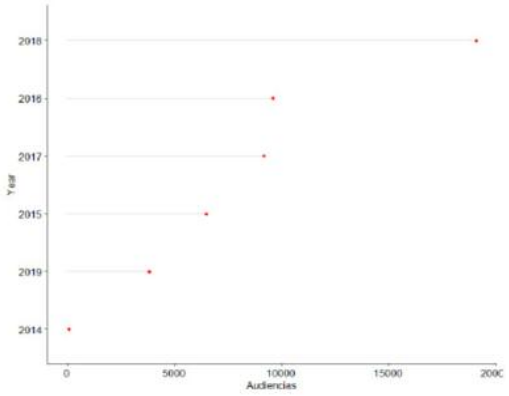


Fig.3 Audiencias Lobby por año.

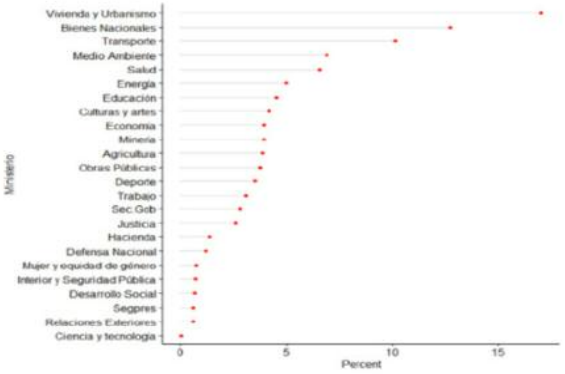


Fig.4 % de audiencias por ministerio.

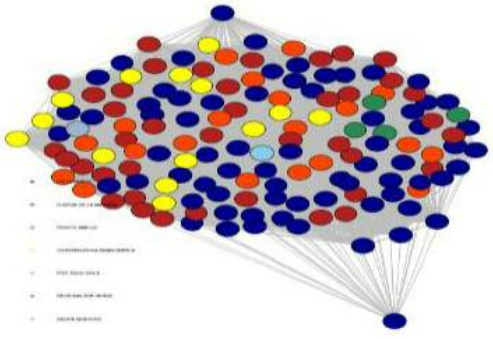


Fig5. Redes de audiencias Lobby en el congreso.

Daniela Vazquez

Abriendo y analizando los Diarios de Sesiones del Parlamento uruguayo con R

Existen diversas iniciativas de datos abiertos en Uruguay, pero aún queda mucho camino por recorrer. Para que un dato sea considerado abierto, debe cumplir con tres características: accesible (publicado en la web), disponible en un formato que sea legible por computadoras y con una licencia clara de uso. En el caso de los Diarios de Sesiones de Diputados y Senadores (DSDS) el segundo punto no se cumple lo que dificulta el acceso y uso de los mismos. En esta presentación se muestra cómo acceder y preparar los DSDS en un formato adecuado para analizarlos, y realizar un análisis usando text mining. Se utilizó el paquete rvest para obtener de forma sistemática los archivos PDF de las sesiones parlamentarias desde enero de 2017 hasta la actualidad desde https://parlamento.gub.uy/documentosyleyes/documentos/diarios-de-sesion y pdftools para extraer el contenido de dichos archivos. Al código fuente del proceso se puede acceder en https://github.com/d4tagirl/uruguayan_parliamentary_session_diary . Con los datos en un formato adecuado se realizó un análisis de text mining con el paquete tidytext. Se calculó la frecuencia de las sesiones y su duración, estimada a través de la cantidad de palabras en el Diario de cada Sesión. Se analizó el texto de las sesiones, centrándose en identificar el sentimiento (usando el Lexicon de Rosas y otros, 2012) para conocer la evolución del sentimiento a lo largo del año (Fig.1) y los temas tratados calculando el tf-idf para medir la importancia de una palabra para una sesión, dado el conjunto de sesiones analizadas. Se nota que el sentimiento empieza a ser más positivo a partir de setiembre de 2017, ante la Renuncia del Vice Presidente Raúl Sendic. Se identifican los temas en las sesiones con sentimientos más extremos (Fig. 2): las dos sesiones más nega-

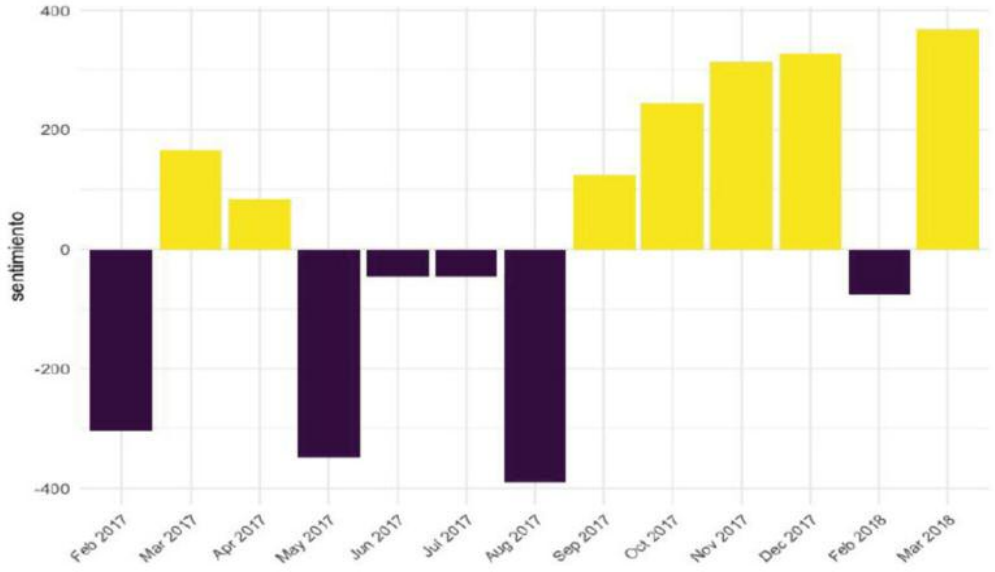


Fig. 1. Evolución del sentimiento de las sesiones de Diputados por mes

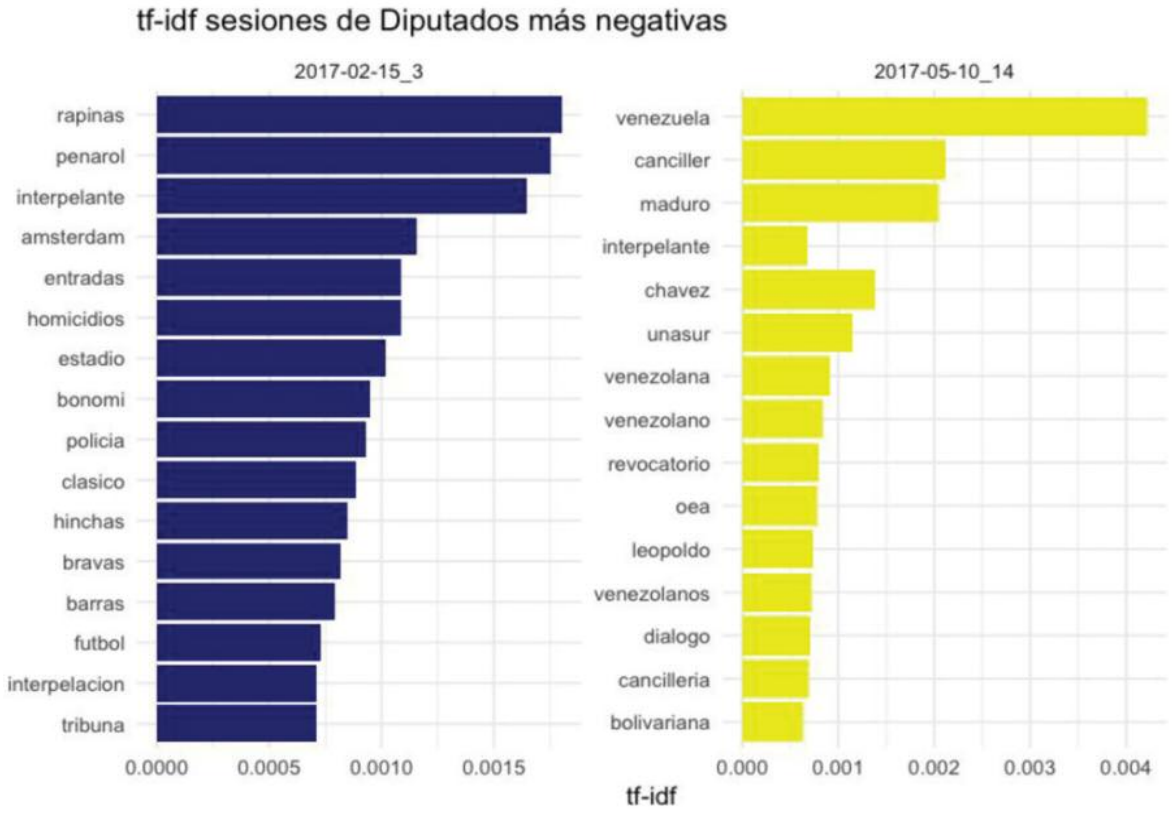


Fig. 2. Tf-idf de los términos con mayor tf-idf, para las dos sesiones de Diputados más negativas

tivas fueron las interpelaciones al Ministro del interior Eduardo Bonomi (extrema violencia en el fútbol) y la interpelación al Canciller Nin Novoa (posición a adoptar ante la crisis venezolana).

En este trabajo se utilizó R de forma exitosa para facilitar el acceso a datos abiertos, que aún necesitan mejorar su acceso, y para analizarlos.

Predicción de heladas usando aprendizaje automático e internet de las cosas

Introducción

En este trabajo se comparten las experiencias de la construcción de un sistema predictivo para heladas agronómicas [2,9] usando R para la experimentación y análisis de resultados. Se utilizaron los enfoques actuales de experimentación para la predicción de temperaturas mínimas diarias que utilizan algoritmos de aprendizaje automático entrenados por lecturas pasadas de los sensores de temperatura y humedad para predecir temperaturas futuras. Sin embargo, al contrario de los enfoques actuales, asumimos que las condiciones termodinámicas circundantes son informativas para la predicción. Por esto, se desarrolló un modelo por cada ubicación, que incluye en su información de entrenamiento las lecturas de sensores de todas las demás ubicaciones que son más relevantes.

Evaluamos nuestro enfoque mediante el entrenamiento de modelos de regresión y clasificación, muchos ya propuestos en la literatura para el problema de predicción de heladas, sobre datos de cinco estaciones meteorológicas distribuidas a lo largo de la Provincia de Mendoza en Argentina. Usamos los paquetes de R caret [4] y doParallel [8] para construir los experimentos. Los algoritmos de aprendizaje automático utilizados fueron: regresión logística, árboles de decisión (C5.0 [3], particionamiento recursivo [7]), random forest [5] y redes bayesianas [6]. Dada la escasez de eventos de heladas, se procedió a balancear el dataset utilizando la técnica de sobremuestreo de minorías sintéticas (SMOTE) [1].

Las métricas utilizadas para comparar los modelos de clasificación fueron: sensitivity o recall (también denominado exhaustividad), precisión, F-measure y para los modelos de regresión RMSE y MAE. Los resultados experimentales muestran que seleccionar a los vecinos más relevantes y entrenar los modelos con SMOTE reduce los errores de predicción de ambos predictores de regresión (random forest y redes bayesianas) para las cinco ubicaciones, en algunos casos hasta un 10%. Además aumenta el rendimiento de los predictores de clasificación, en términos de F-measure , de random forest para cuatro ubicaciones mientras se mantiene sin cambios para la restante, y produce resultados no concluyentes para el predictor de regresión logística. Los valores de F-measure de los mejores modelos resultantes variaron entre 0.7 y 0.9 y los de recall entre 0.7 y 0.8. Otra observación esperable es en el aumento del recall en detrimento de la precisión en comparar los experimentos con y sin aplicación de SMOTE respectivamente. Estos resultados comprobaron la afirmación principal: que la información termodinámica de las ubicaciones vecinas puede ser informativa para mejorar las predicciones de regresión y clasificación, pero también es lo suficientemente buena como para sugerir que el enfoque actual es un recurso válido y útil para los productores agropecuarios o tomadores de decisiones, ya que disminuye los errores en la predicción. En un esquema de redes inalámbricas de sensores (IoT) en el campo no siempre los nodos tienen acceso a internet a un servidor central para recibir la predicción u otra información. Conocer los vecinos más relevantes ayudaría a tomar mejores decisiones de ruteo de la información, que permita alimentar a sus predictores de heladas locales.

Referencias

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research , 16 , 321-357.

- Diedrichs, A. L., Bromberg, F., Dujovne, D., Brun-Laguna, K., & Watteyne, T. (2018). Prediction of frost events using machine learning and IoT sensing devices. IEEE Internet of Things Journal , 5 (6), 4589-4597.
- Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2014). C50: C5. 0 decision trees and rule-based models. R package version 0.1. 0-21, URL http://CRAN. R-project. org/package C , 50 .
- Kuhn, M. (2012). The caret package. R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret .
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news , 2 (3), 18-22.
- Scutari, M., & Ness, R. (2012). bnlearn: Bayesian network structure learning, parameter learning and inference. R package version , 3 .
- Therneau, T. M., Atkinson, B., & Ripley, B. (2010). rpart: Recursive partitioning. R package version , 3 (3.8).
- Weston, Steve, and Rich Calaway. "Getting Started with doParallel and foreach." (2018). URL https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstar tedParallel.pdf
- Ana Laura Diedrichs. (2019, July 29). anadiedrichs/diedrichs2017prediction-frost-experiments: first release (Version v0.1). Zenodo. http://doi.org/10.5281/zenodo.3354574

Sara Acevedo, Cristina Contreras, Carlos Ávila, Javier Rivera & Carlos Bonilla

Uso de R y Shiny en el desarrollo de modelos predictivos aplicados a ciencias del suelo

En ciencias del suelo, la estimación de variables complejas por medio de un modelo predictivo basado en mediciones de propiedades del suelo sencillas se denomina función de pedotransferencia (o abreviado del inglés PTF). Un desafío de la aplicación de las PTFs para su uso por diversos tomadores de decisiones es cómo esta información es entregada de forma amigable y sencilla. La divulgación de este tipo de información es compleja, ya que considera varios pasos en su desarrollo, siendo de interés final para sus potenciales usuarios solo el resultado final o output. Como ejemplo, hidrólogos y científicos de suelo requieren una variedad de datos como input para la modelación de diversos procesos ambientales, los cuales podrían ser accesibles vía software o plataforma. Las aplicaciones web desarrolladas en R utilizando Shiny han mostrado su utilidad en diversos ámbitos, incluyendo la investigación, tanto profesional como académica, en el área de estudio de suelos con enfoque medioambiental y de ingeniería.

Las aplicaciones Shiny han sido usadas como marcos para toma de decisiones medioambientales para la evaluación de eventos climáticos extremos en base a índices agroclimáticos y para la visualización y análisis de series de tiempo de datos hidrológicos. Otro uso interesante de estas aplicaciones es su potencialidad en la entrega de resultados de bases de datos gubernamentales aplicado a análisis de suelo. Por ejemplo, el Departamento de Agricultura de los Estados Unidos (USDA) posee gran parte de su información pedológica del National Soil Information System (NASIS) en esta plataforma. El grupo de investigación del Laboratorio de Biofísica de Suelos ha estado desarrollando análisis de datos de suelo por medio de muestreo en la zona Central de Chile, los cuales ha procesado y analizado por medio de R. Utilizando diversas librerías, se han desarrollado PTFs para diferentes variables objetivo, incluyendo estabilidad de agregados y contenidos de humedad. Las librerías utilizadas incluyen principalmente tidyverse y caret, para la generación de modelos en base a regresión lineal y no lineal. Adicionalmente, parte de los resultados se encuentran publicados como Shiny apps en el sitio web del grupo (biofísica.ing.puc.cl). El objetivo general es generar herramientas de uso fáciles de aplicar y entender por los potenciales usuarios a nivel nacional. Este trabajo presenta un resumen de la experiencia del uso de R en ciencias del suelo tanto en el desarrollo de modelos predictivos, procesamiento de datos y transferencia de conocimiento por medio de talleres interactivos.

Página web, que incluye links a las aplicaciones: biofisica.ing.puc.cl

Daniela Prina, Sabrina López, Stephan Arndt & Laura Acion

Uso de un enfoque de aprendizaje automático para predecir éxito en el tratamiento de adicciones

Introducción

Existen varios métodos para construir modelos de predicción. Los modelos de predicción a menudo se generan mediante algún tipo de regresión lineal o logística. Más recientemente, otros algoritmos de aprendizaje como random forests (RF) o redes neuronales se están usando para la predicción en ciencias de la salud. Estas técnicas más nuevas podrían mejorar la predicción y, en consecuencia, las posibilidades de encontrar el tratamiento más efectivo para cada paciente.

La riqueza de las técnicas de modelado actualmente disponibles suele obligar a juzgar, a priori, cuál será el mejor método de predicción. Super Learner (SL) [1] es una metodología que facilita esta decisión al combinar todos los algoritmos de predicción identificados pertinentes para un problema en particular. Bajo ciertos supuestos, SL genera un modelo final que es al menos tan bueno como cualquiera de los otros modelos considerados para predecir el resultado. Esta propiedad de SL es teórica [1] y está sustentada empíricamente [2].

El objetivo de este trabajo es mostrar el uso de SL según su implementación en H2O a través de su interfaz para R [3] (para más detalles de este trabajo ver [4]).

Métodos

Datos. Para ilustrar diferentes enfoques analíticos, nos centramos en el tratamiento ambulatorio de adicciones para personas hispanas adultas. Se usaron datos administrativos públicos, habitualmente usados por investigadores de adicciones. Estos datos permiten ilustrar el uso de las metodologías dentro de un entorno realista. Outcome. Como es habitual en este tema, el tratamiento completado se consideró un éxito, las demás razones de alta terapéutica (por ej, "en contra del consejo profesional", "persona encarcelada") se consideraron como indicadores de tratamientos exitosos. Predictores. Se incluyeron 28 predictores en el análisis, incluyendo características de las personas, características de los tratamientos, tipo de adicción y problemas coexistentes de salud mental. Modelado. Se ajustaron modelos de regresión logística, regresión penalizada (por ej, LASSO y ridge), random forest, redes neuronales de aprendizaje profundo y SL. Se usó R [5] y la interfaz con R de H2O [3] que optimiza estos métodos para bases de datos grandes. Todos los algoritmos se evaluaron usando el área bajo la curva ROC (AUC).

Resultados

SL mostró la mayor AUC. El rendimiento de SL fue seguido muy de cerca por RF. El algoritmo con el peor desempeño fue la regresión logística. La mejora relativa en el AUC de SL fue menor al 5% en comparación con el peor método de predicción. Además, la AUC para SL tuvo la varianza estimada más pequeña. El resto de los modelos considerados tuvieron varianzas estimadas para la AUC de hasta 20% más altas que SL. Todos los modelos de regresión paramétrica, tanto los enfoques penalizados como los no penalizados, se comportaron de manera casi idéntica con respecto a AUC. Para los modelos que incluyeron todos los predictores e interacciones de 2 vías, LASSO superó a los otros tres modelos de regresión. Las AUC para LASSO fueron mayores que para las regresiones de ridge y las logísticas.

Conclusiones

Este trabajo comparó varios modelos para predecir el éxito del tratamiento de adicciones. Como

era de esperar, SL mostró el mejor rendimiento predictivo. En este caso la superioridad de SL fue magra. Sin embargo, este resultado no era en absoluto evidente antes de realizar el análisis. Creemos que la falta de diferencias importantes en las soluciones de todos los métodos es un resultado relevante. La falta de diferencia sustantiva entre SL y el resto de los modelos usados significa que: a) cualquiera de estos métodos podría usarse para estos datos y b) no hay problemas importantes en los supuestos de los diferentes modelos utilizados. Esto nunca es evidente antes de analizar un conjunto de datos si se usa un único enfoque analítico, lo más habitual en ciencias de la salud. En este sentido, SL sirve como una herramienta para agilizar y mejorar el análisis de sensibilidad para predicción. Además, incluso pequeñas mejoras en la predicción pueden tener un alto impacto dependiendo de cada problema en particular. En este caso, una pequeña mejora en la predicción podría afectar significativamente la salud de los pacientes y los costos de tratamiento. En otros casos, pequeñas mejoras de predicción podrían salvar vidas.

Referencias

- van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical applications in genetics and molecular biology. 2007;6(1).
- Polley EC, Rose S, van der Laan MJ. Super learning. Targeted Learning: Springer; 2011. p. 43–66.
- LeDell E, Gill N, Aiello S, Fu F, Candel A, Click C, Kraljevic T, Nykodym T, Aboyoun P, Kurka M, Malohlava M. h2o: R Interface for 'H2O'. R package version 3.22.1.1. 2019.
- Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One. 2017;12(4), e0175383.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.

Pablo Picardo

Predicción del precio de la vivienda. Comparación de modelos predictivos con CARET

Introducción

En este trabajo se presentan modelos predictivos para el precio de un activo de difícil valuación como la vivienda. Se utilizan dos fuentes de datos novedosas para la ciudad de Montevideo: una proveniente de sitios web obtenidos mediante web scraping (para el período febrero 2018 - enero 2019) y otra de registros administrativos de transacciones. Se implementan tres modelos fácilmente replicables con el paquete CARET Kuhn (2018): modelo lineal, árbol de regresión y bosques aleatorios, luego se compara su poder predictivo. Los resultados arrojan una mejor performance del modelo de bosques aleatorios (random forest) respecto al modelo lineal hedónico, ampliamente difundido en la literatura. Se busca incorporar al análisis de predicción de precios una metodología de creciente aplicación a nivel internacional así como poner a disposición una nueva base de datos procesada y actualizada (ver kaggle dataset).

Antecedentes

Los antecedentes se pueden dividir entre los trabajos que tratan sobre el mercado de vivienda en Uruguay y los que aplican técnicas de aprendizaje estadístico para la predicción de precios de vivienda. Para este punto se encontró un solo trabajo nacional y varias referencias internacionales. Dentro de la segunda categoría, se destaca el trabajo de Mullainathan and Spiess (2017), que realiza una introducción a los modelos de aprendizaje estadístico mediante un ejemplo con datos de precios de vivienda. En éste, ilustran diferentes modelos de predicción y muestran su performance predictiva, destacando la mejor de los modelos de aprendizaje. Este artículo, al igual que Athey (2018) y Varian (2014) fueron la principal motivación para incorporar técnicas de aprendizaje estadístico y resultan referencias básicas para el área económica. La referencia nacional para este trabajo es Goyeneche et al. (2017), en ella se utiliza una base de datos de tasaciones del Banco Hipotecario del Uruguay con el objetivo de predecir el precio contado de un inmueble. Por otra parte, en el reciente artículo de C˘ eh et al. (2018) se realiza un trabajo similar al presentado aquí. Se compara la performance predictiva de un modelo de bosques aleatorios en relación a una regresión lineal hedónica para el precio de los apartamentos en Liubliana, Eslovenia.

Estrategia metodológica

El objetivo es comparar tres modelos en relación a su poder predictivo. La estrategia de comparación elegida consiste en considerar pocas variables explicativas en ambas bases de datos. La elección de las variables se basa en la confiabilidad de las mismas y en la baja presencia de datos faltantes. La modelización se realiza en cada base de datos por separado y por tipo de propiedad. Para comparar los modelos, se realiza validación cruzada estándar con el paquete CARET (Kuhn, 2018) y en una muestra de entrenamiento particionada. Basado en los resultados de la validación cruzada, se toman las medidas de performance en la muestra de testeo considerando cada modelo que emerge de la validación cruzada. Las medidas son la menor Raíz del Error Cuadrático Medio (RECM) y el menor Error Porcentual Absoluto Medio (EPAM). Una vez seleccionado el mejor modelo, se incorporan más variables en ese modelo para mejorarlo en términos predictivos (en el caso del desarrollo de bosques se utiliza el paquete randomForest (Liaw and Wiener, 2002)).

Datos

Los datos de ofertas fueron recopilados a través de la API (interfaz para acceder a la página web) puesta a disposición por mercadolibre.com. Para ello se utilizó un programa elaborado en python. Esto sería posible lograrlo a través de R, utilizando paquetes como rvest (Wickham, 2019), httr (Wickham, 2018), jsonlite (Ooms, 2014), entre otros. La muestra original incluye todas las ofertas de venta de inmuebles para la ciudad de Montevideo para el período febrero 2018 - enero de 2019 inclusive. Se realizaron bajadas sucesivas de datos alrededor del día 25 de cada mes. Luego de la limpieza, la base de datos cuenta con aproximadamente 90.000 observaciones únicas (inmuebles cuyo ID no se repite ni hay indicios de repetición).

La base de datos de transacciones abarca el período enero 2017 - junio de 2018 inclusive e identifica transacciones de compraventa de padrones en Montevideo con destino de vivienda. La muestra seleccionada y procesada cuenta con 12.815 observaciones. A modo de ejemplo, se presenta a continuación la ubicación de las transacciones de apartamentos (las descripciones geográficas utilizan el paquete Leaflet, Cheng et al. (2018) y los gráficos se realizan con ggplot2 Wickham (2016)).

Resultados

Los resultados arrojan que el modelo de bosques aleatorios tiene una mejor performance predictiva respecto al modelo lineal hedónico, considerando los datos de Montevideo, tanto de ofertas como de transacciones. Esto parecería estar explicado por la naturaleza no lineal del problema de predicción. La superioridad del modelo de bosques aleatorios es aún mayor con los datos de ofertas. Con estos datos para apartamentos, el error porcentual absoluto medio (EPAM) evaluado en una muestra de testeo independiente, se reduce en 40% (de 17% a 10 %) al utilizar el modelo de bosques aleatorios respecto al modelo lineal. Para las casas el EPAM se reduce 36% (22% a 14 %) respecto al modelo lineal. Las predicciones de los modelos ampliados, en el caso de las ofertas, presentan un error de USD 29.500 (RECM) para apartamentos, y USD 56.800 para las casas. En el caso de transacciones, la performance del modelo de bosques aleatorios también mejora las predicciones respecto al lineal, aunque en proporciones modestas. En el caso de los apartamentos mejora un 6% (considerando EPAM) y en el caso de las casas la reducción es de 1% (tomando RECM). Los resultados obtenidos pueden establecer una base para desarrollos posteriores que exploren variantes en los modelos utilizados, por ejemplo ajustando los parámetros de variables y poda. Igualmente, se podría completar el análisis contemplando nuevas técnicas de aprendizaje estadístico que no se incluyen en el presente trabajo, principalmente Máquinas de Vectores de Soporte. Adicionalmente es posible incorporar más variables hedónicas (incluyendo variables espaciales y geográficas) y utilizar técnicas de econometría espacial. Adicionalmente, los modelos podrían contemplar por un lado la dimensión temporal y por el otro, fundamentos macroeconómicos.

Referencias

- Athey, S. (2018). The Impact of Machine Learning on Economics. In The Economics of Artificial Intelligence: An Agenda, pages 1–31. University of Chicago Press.
- C˘ eh, M., Kilibarda, M., Lisec, A., and Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5):168.
- Cheng, J., Karambelkar, B., and Xie, Y. (2018). Interactive Maps with JScript Leaflet. R package v. 2.0.2.
- Goyeneche, J. J., Moreno, L., and Scavino, M. (2017). Predicción del valor de un inmueble mediante técnicas agregativas. Serie DT IESTA (17/1).
- Kuhn, M. e. a. (2018). Caret: Classification and Regression Training. R package v. 6.0-84.
- Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3):18–22.
- Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2):87–106.
- Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO].
- Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2):3–28.
- Wickham, H. (2016). ggplot2: Elegant Graphics

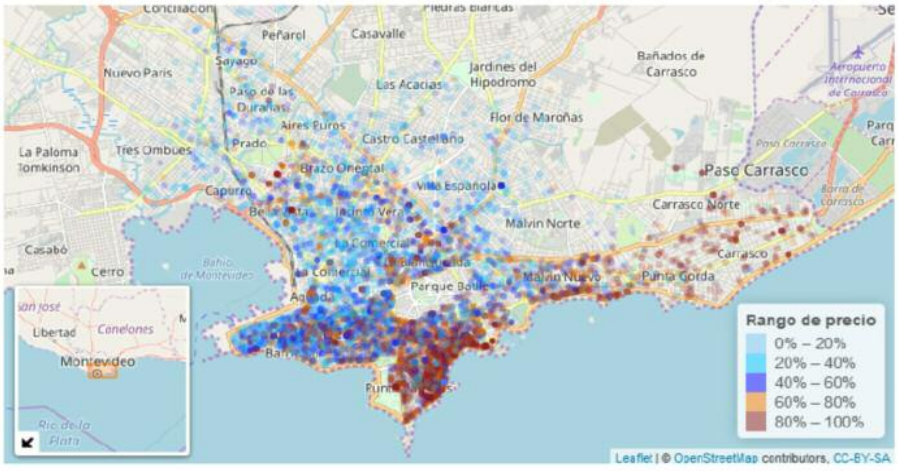


Figura 1: Ubicación de Apartamentos transados. Rango de precios en base a los quintiles por tipo de propiedad. Fuente: Elaboración propia, en base a DGR, DNC, SIG-IM

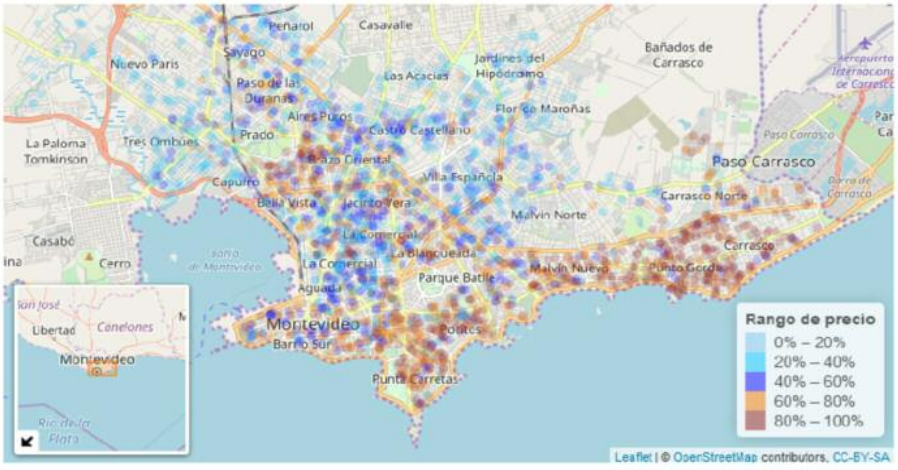


Figura 2: Ubicación de casas transadas

for Data Analysis. Springer-Verlag New York.

- Wickham, H. (2018). httr: Tools for Working with URLs and HTTP. R package v. 1.4.0.
- Wickham, H. (2019). rvest: Easily Harvest (Scrape) Web Pages. R package v. 0.3.3.

Francisco Zambrano

Predicción de la sequía agrícola en Chile: regresión lineal vs deep learning

Resumen

La seguridad alimentaria mundial se ve afectada negativamente por la sequía. Las proyecciones climáticas predicen que la frecuencia y la intensidad de la sequía pueden aumentar en diferentes partes del mundo, lo que es particularmente peligroso para los países en desarrollo. Los pronósticos tempranos de la temporada sobre la ocurrencia y severidad de la sequía podrían ayudar a mitigar mejor sus consecuencias negativas. El objetivo de este estudio fue evaluar si la sequía agrícola se puede predecir con precisión a partir de fuentes de datos casi en tiempo real disponibles de forma gratuita. Como variable de respuesta, utilizamos la puntuación estándar del NDVI acumulado estacional (zcNDVI) en base a los datos de 2000-2017 de MODIS como proxy de las anomalías de la productividad primaria estacional. Las predicciones se realizaron con tiempos de entrega pronosticados entre uno y seis meses antes del final de la temporada, que variaban entre las unidades censales. Las variables predictoras incluyeron el zcNDVI obtenido al acumular NDVI desde el inicio de la temporada hasta el tiempo de predicción; Índices de precipitación estandarizados para escalas de tiempo de uno, tres, seis, 12 y 24 meses derivados de estimaciones de lluvia satelital; dos índices de oscilación climática que incluyen la Oscilación Decenal del Pacífico y el índice ENOS Multivariado; la duración de la temporada de crecimiento; y latitud y longitud. Para cada una de las 758 unidades censales consideradas, las series temporales de la respuesta y las variables predictoras se promediaron para las áreas agrícolas, lo que dio como resultado una serie temporal de 17 estaciones por unidad para cada variable. Los enfoques de predicción utilizados fueron: (i) regresión lineal óptima (OLR), por lo que para cada unidad de censo se seleccionó el único predictor que explicaba mejor la variabilidad interanual de zcNDVI, y (ii) una arquitectura de red neuronal multicapa de alimentación directa a menudo llamada aprendizaje profundo (DL) donde todos los predictores para todas las unidades se combinaron en un solo modelo espacio-temporal. Ambos enfoques se evaluaron con un procedimiento de validación cruzada de un año de ausencia. Ambos métodos mostraron una buena precisión de predicción para tiempos de entrega pequeños y valores similares para todos los tiempos de entrega. Los valores medios de R2 cv para OLR fueron 0.95, 0.83, 0.68, 0.56, 0.46 y 0.37, contra 0.96, 0.84, 0.65, 0.54, 0.46 y 0.38 para DL, para uno, dos, tres, cuatro Cinco y seis meses de tiempo de entrega, respectivamente. Dada la amplia gama de climas y tipos de vegetación cubiertos dentro del área de estudio, esperamos que los modelos presentados puedan contribuir a un sistema mejorado de alerta temprana para la sequía agrícola en todo el mundo.

En este estudio el procesamiento de los datos espaciales se realizó con los paquetes 'raster' [1], 'sp' [2] y 'maptools' [3]. Para el resumen y manejo de datos se utilizaron la serie de paquetes de 'tidyverse' [4]. La implementación de las redes neuronales multi-capa (deep learning) se realizo con el framework de java 'h2o' mediante el paquete 'wraper' de R 'h2o' [5]. Los análisis de regresión lineal con los paquetes 'base' de R. Todos los gráficos se realizaron con el paquete 'ggplot2' [6]. Este estudio se publicó en Diciembre del año 2018 [7] en la revista 'Remote Sensing of Environment', el artículo original se puede descargar del siguiente link https://www.sciencedirect.com/science/article/pii/S0034425718304541.

References

- R. J. Hijmans, Raster: Geographic data analysis and modeling (2019).
- E. J. Pebesma & R. S. Bivand, Classes and methods for spatial data in R. R News, 5 (2005) 9–13.
- R. Bivand & N. Lewin-Koh, Maptools: Tools for handling spatial objects (2019).
- H. Wickham, Tidyverse: Easily install and load the 'tidyverse' (2017).
- E. LeDell, N. Gill, S. Aiello, A. Fu, A. Candel, C. Click, T. Kraljevic, T. Nykodym, P. Aboyoun, M. Kurka, & M. Malohlava, H2o: R interface for 'h2o' (2019).
- H. Wickham, Ggplot2: Elegant graphics for data analysis (Springer-Verlag New York, 2016).
- F. Zambrano, A. Vrieling, A. Nelson, M. Meroni, & T. Tadesse, Prediction of drought-induced reduction of agricultural productivity in Chile from MODIS, rainfall estimates, and climate oscillation indices. Remote Sensing of Environment, 219 (2018) 15–30. https://doi.org/10.1016/J.RSE.2018.10.006.

Mauricio Vargas

Open Trade Statistics: Database, API, Dashboard and Utility Program made with R

Open Trade Statistics (OTS) was created with the intention to lower the barrier to working with international economic trade data. It includes a public API, a dashboard, and an R package for data retrieval. OTS provides data for the period 1962-2017 covering all countries that report to the United Nations, accessing out datasets has no cost and does not need to create an account. The project started when I was affected by the fact that many Latin American Universities have limited or no access to the United Nations Commodity Trade Statistics Database (UN COMTRADE) because the institutional access is paid and can be very expensive in our latin american reality. There are alternatives to COMTRADE, for example the Base Pour L'Analyse du Commerce International (BACI) constitutes an improvement over COMTRADE as it is constructed using the raw data and a method that reconciles the declarations of the exporter and the importer, but you will need UN COMTRADE institutional access to download their datasets. After contacting UN COMTRADE, and suggesting to them my idea of doing something similar to BACI available for anyone but keeping commercial purposes out of the scope of the project, Igot an authorization to share curated versions of their datasets. R is central to this project. Even our API was made with R. I used the Plumber package and nginx enhanced with a secured connection by using Let's Encrypt.

Tradestatistics package, a part of OTS and available on CRAN, provides a really efficient way to interact with the API. OTS was created thinking of people from humanities and social sciences, json files from an API are not optimal, and therefore the package converts the information to tidy data and even handles some joins to provide the best human readable data.

rOpenSci ideas and support were crucial to the development of the project as both the package and the API were highly improved with comments from the community. In addition to the package, I created a shiny dashboard . The idea of the dashboard is to provide a GUI for the package, providing the option to obtain the same data and the possibility of downloading it from the browser in different formats such as csv, xlsx and others.

References

- Vargas (2019). tradestatistics: Open Trade Statistics API Wrapper and Utility Program. R package version 0.2.2. https://docs.ropensci.org/tradestatistics
- Dowle and Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.2.
- Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
- Conway, et al (2017). RPostgreSQL: R Interface to the 'PostgreSQL' Database System. R package version 0.6-2. https://CRAN.R-project.org/package=RPostgreSQL
- Trestle Technology, LLC (2018). plumber: An API Generator for R. R package version 0.4.6. https://CRAN.R-project.org/package=plumber
- Chang and Borges (2019). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.1.8000. http://rstudio.github.io/shinydashboard
- United Nations (2010). United Nations commodity trade statistics database (COMTRADE). http://comtrade.un.org

Paola Corrales & Elio Campitelli

rvad: perfiles verticales de viento a patir de datos de radares meteorológicos

Introducción

El estudio y monitoreo del viento en niveles bajos de la atmósfera es de suma importancia ya que el mismo afecta diversos procesos que tienen un alto impacto en la sociedad. Por ejemplo influye en el desarrollo y severidad de las tormentas de manera directa por la rotación del viento con la altura en los primeros kilómetros. También afectan de manera indirecta a través de procesos de mayor escala, como el transporte de humedad desde el Amazonas hacia el sur de Sudamerica. Sin embargo, las mediciones de viento en superficie tienen una resolución espacial y temporal muy baja y normalmente se realizan únicamente a 10 metros de altura.

Los radares Doppler pueden medir el viento en un volumen de aire cada 5 o 10 minutos. Por esto tienen una gran potencialidad para estimar la variación del viento con la altura. Pero las variables medidas por el radar requieren algoritmos de procesamiento y técnicas de control de calidad de datos. Implementarlos en R permite extender el uso del lenguaje a otras diciplinas como la meteorología, tanto en investigación como en tareas operativas de monitoreo del tiempo.

En este trabajo se presenta el paquete rvad que implementa de la técnica Velocity Azimuth Display (VAD) para estimar un perfil vertical de las componentes horizonaltes del viento a partir del viento medido por el radar.

Algoritmo

Un radar gira sobre su eje enviando pulsos de energía electromagnética en cada ángulo horizontal o azimut. Al terminar cada giro de 360 grados, cambia su ángulo de elevación y repite el proceso. El pulso de radar recorre una determinada distancia (rango) y en el camino puede ser interceptado por gotas de agua, granizo o, en casos sin nubosidad, insectos. Una de las variables medidas por el radar es la velocidad radial o Doppler, que corresponde a la componente radial del viento, es decir, la proyección del viento en la dirección de la propagación del haz de radar.

Valores negativos corresponden a movimientos hacia el radar y valores positivos a movimientos desde el radar, mientras que el valor nulo ocurre en las regiones donde el viento es perpendicular a la trayectoria del haz. La técnica VAD aprovecha el comportamiento sinusoidal del viento radial para cada rango y ángulo de elevación y ajusta estos datos a una función de la forma v cos(q) cos(ϕ) + u cos(theta)sin(ϕ) donde q es el ángulo de elevación y ϕ el azimuth. Estimando la propagación del haz del radar, es posible obtener una altura vertical para los valores del viento zonal (u) y meridional (v). Finalmente, bajo ciertas condiciones es posible realizar un promedio de todas las estimaciones de u y v para obtener un perfil vertical de viento representativo del volumen de atmósfera escaneado por el radar.

Implementación

El paquete rvad implementa el algoritmo VAD presentado por Browning and Wexler (1968) e incluye una serie de controles de calidad que buscan solucionar problemas típicos asociados a los datos de radar. En primer lugar, la función vad_fit() toma vectores con la velocidad radial, el azimuth, el rango y el ángulo de elevación y realiza un ajuste sinusoidal para cada anillo de datos (las observaciones para un rango y ángulo de elevación particular). Además realiza los siguientes controles de calidad:

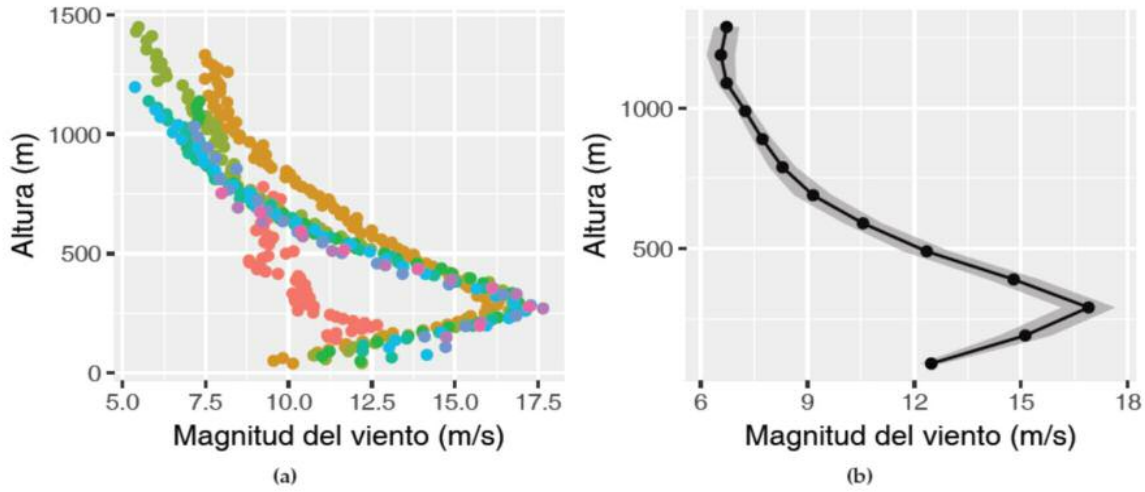


Figura 1: Estimación del viento obtenido a partir de (a) vad_fit() y (b) vad_regrid()

Antes del ajuste:

Cuenta la cantidad de datos faltantes en el anillo y si supera un umbral máximo max_na el anillo es rechazado.

Busca gaps o huecos continuos de datos faltantes y rechaza aquellos anillos con huecos mayores al umbral max_consecutive_na.

Luego del ajuste:

El algoritmo rechaza los anillos cuyo ajuste tiene un R^2 menor al umbral r2_min.

El data frame resultante puede ser visualizando fácilmente con el método plot().

En segundo lugar la función vad_regrid() toma el data frame generado por vad_fit() y genera un único perfil vertical en una grilla regular utilizando un regresión local de orden 1. La función devuelve el valor de u y v para cada nivel de altura y el intervalo de confianza asociado.

Estas dos funciones principales deben ser utilizadas en serie para obtener el pertfil vertical. El paquete las implementa por separado dado que es importante analizar el ajuste individual de cada anillo en función del ángulo de elevación para evaluar la calidad de los datos y ajustar los controles de calidad.

Referencias

- Browning, K. A., and R. Wexler. 1968. "The Determination of Kinematic Properties of a Wind Field Using
- Doppler Radar." Journal of Applied Meteorology 7 (1): 105–13. https://doi.org/10.1175/1520-0450(1968)007<0105:TDOKPO>2.0.


```
library(rvad)
library(ggplot2)
VAD <- with(radial_wind, vad_fit(radial_wind, azimuth, range, elevation))
plot(VAD) +
  scale_color_discrete(guide = "none") +
  labs(x = "Magnitud del viento (m/s)", y = "Altura (m)")
wind_profile <- vad_regrid(VAD, layer_width = 100)
plot(wind_profile) +
  labs(y = "Magnitud del viento (m/s)", x = "Altura (m)")
```

Javier Fajardo, Derek Corcoran, Patrick Roehrdanz, Lee Hannah & Pablo A Marquet

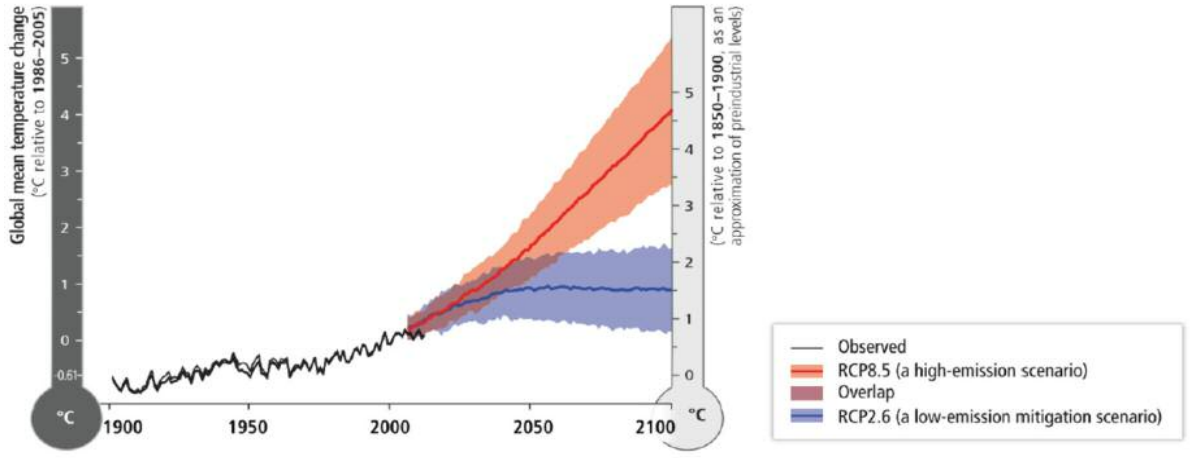
GCM compareR: una aplicación web para evaluar escenarios de cambio climático

Desde un equipo internacional de biólogos de la conservación, liderado desde Chile, hemos desarrollado la aplicación web basada en Shiny GCM compareR, enfocada a la evaluación de escenarios de cambio climático en estudios de ecología y conservación. Gran parte de la investigación del cambio climático y sus efectos sobre la biodiversidad, glaciares y poblaciones humanas recae sobre el estudio de escenarios. Así, paneles internacionales de investigadores han definido diversos futuros (cuya materialización depende de numerosos factores, entre ellos las decisiones tomadas por los seres humanos en cuestiones como la reducción de emisiones o la implementación de estrategidas de mitigación). Los peores escenarios presentan futuros donde la temperatura del planeta subiría unos 4ºC respecto a las de principios del siglo XXI, mientras que los más favorables presentan la posibilidad de contener el aumento por debajo de los 2ºC (IPCC, 2014).

Para estudiar los detalles de estos escenarios y ser capaces de extraer conclusiones y plantear acciones, diversas instituciones a nivel mundial han desarrollado modelos que, traducidos a mapas, son capaces de proporcionar proyecciones de los cambios estimados bajo los escenarios en las distintas partes de nuestro planeta. En la actualidad, más de 35 instituciones distintas han desarrollado estos modelos de circulación global (GCMs, por sus siglas en inglés), presentando cada una de ellas varias alternativas según sean construidos estos modelos. En estas circunstancias, el numero de modelos disposibles para los científicos que tratan de aventurar los cambios que las condiciones futuras pueden suponer para la vida de nuestro planeta ha aumentado enormemente en los últimos años.

La posibilidad de utilizar numerosos y cada vez mejores modelos para interpretar el cambio climático multiplica las posibilidades de los investigadores, al mismo tiempo que genera la necesidad de desarrollar nuevas herramientas que permitan comprender y extraer la máxima información certera de todo este universo de modelos. Sin embargo, esta herramienta aún no existe, y es tarea de los biólogos y ecólogos la evaluación de estos modelos y, eventualmente, el decidirse por basar sus interpretaciones en unos u otros GCMs. A menudo carentes del suficiente conocimiento sobre peculiaridades más propias del campo de la climatología, esta es una tarea que a menudo supera a los científicos, que terminan repitiendo decisiones de otros trabajos ante la imposibilidad de seguir un criterio repetible y adaptable a las necesidades de sus estudios.

Hemos desarrollado GCM compareR para representar ese papel. Escrita en R, e interactiva gracias a Shiny, esta aplicación gratuita, de código abierto y de libre acceso, pone al alcance la todos los investigadores interesados los datos y los flujos de trabajo que les permitirán evaluar cómo los GCMs representan los distintos escenarios, para así entender cuáles son mejores para sus intereses. La aplicación está perfectamente documentada y expande métodos de trabajos existentes en la literatura científica (por ejemplo, ver Snover et al., 2013; Vano et al., 2015), pero cuya aplicación es aún escasa como consecuencia de las dificultades técnicas y complejidad de computación que su análisis requería hasta ahora. La aplicación presenta una interfaz amigable donde el usuario puede definir su escenario de evaluación a partir de una extensiva base de datos de GCMs. Mediante un mapa interactivo con el motor de leaflet, los análisis son adaptables a cualquier parte del planeta, permitiendo hacer comparacio-



nes específicas incluso para aquellas áreas menos estudiadas, como suelen ser las zonas tropicales o los países en vías de desarrollo. Los resultados se ofrecen de forma sencilla en forma de tablas, mapas y gráficos, y son descargables en diversos formatos. La aplicación incluye, además, la posibilidad de generar un reporte que resume todas las condiciones elegidas por el usuario para su análisis y los resultados, embebidos en explicaciones que facilitan la comprensión e interpretación de los mismos. Este reporte está enfocado a incrementar la repetibilidad de los análisis, así como el reporte y comunicación de los resultados.

En definitiva, GCM compareR es una herramienta que nutre a biólogos, conservacionistas y tomadores de decisiones (como ministerios del ambiente y gobiernos) de información objetiva y repetible, y que viene a llenar un escalón difícil de salvar en los flujos de trabajo actuales de estudios de cambio climático. Hemos escrito un artículo científico descriptivo de la aplicación y su utilización que está siendo actualmente evaluado para su publicación por una revista de primer orden de métodos en ecología y evolución. Además de esta publicación, nuestro interés es presentar la aplicación en LatinR 2019. Nuestro equipo de trabajo está formado por biólogos entusiasmado en R. Si bien hemos convertido a R y Rstudio en nuestra herramienta de trabajo, y nos consideramos casi expertos entre los biólogos, carecemos de la formación en progración que tienen los profesionales que vienen de otros campos más relacionados con la informática. Por este motivo, consideramos que LatinR representa un lugar idóneo para ver las impresiones que genera en la comunidad de desarrolladores de R latinoamericana y recibir su feedback.

Citas bibliográficas

- IPCC (2014) Climate Change 2014: Impacts, Adaptation, and Vulnerability.
- Snover, A.K., Mantua, N.J., Littell, J.S., Alexander, M.A., McClure, M.M. & Nye, J. (2013) Choosing and Using Climate-Change Scenarios for Ecological-Impact Assessments and Conservation Decisions. Conservation Biology, 27, 1147-1157.
- Vano, J.A., Kim, J.B., Rupp, D.E. & Mote, P.W. (2015) Selecting climate change scenarios using impact-relevant sensitivities. Geophysical Research Letters, 42, 5516-5525.

Marcos Prunello

InHostShiny: aplicación Shiny para el ajuste y análisis de modelos compartimentales in-host para infecciones virales en poblaciones celulares

Una enfermedad es infecciosa si el agente causante puede ser transmitido de un individuo a otro a través de algún medio. El objetivo de la modelización matemática de una enfermedad infecciosa es describir el proceso de transmisión, generalmente resumido de la siguiente forma: cuando individuos infectados son introducidos en una población susceptible, la enfermedad se pasa a otros individuos a través de sus modos de transmisión, expandiéndose en la población. Un individuo puede ser asintomático durante una etapa temprana y ser diagnosticado como un caso más tarde. Individuos infectados pueden o no recuperarse, ya sea por tratamiento o como respuesta de su sistema inmune, y pueden ganar cierto grado de inmunidad contra una posible re infección. Los modelos matemáticos han sido capaces de proveer claridad sobre los procesos de transmisión y expansión, ayudando a identificar factores claves para el control de la transimisión y prevención y a estimar la severidad y escala potencial de la infección en una población. Un enfoque utilizado para modelar un proceso infeccioso propone particionar la población en compartimentos, por ejemplo, el compartimento S con los individuos susceptibles, I con los infectados y R con los individuos recuperados. La modelización lleva registro de las cantidades de individuos en cada compartimento en un momento dado, planteando un sistema de ecuaciones diferenciales que da lugar a una familia de modelos matemáticos conocidos como SIR.

Este mismo enfoque puede aplicarse para describir procesos infecciosos a nivel microscópico, como la diseminación de una infección viral en una población de células, resultando en modelos conocidos como in-host (en el hospedador). La aplicación que se presenta en este trabajo implementa un modelo in-host para la infección de células T CD4+ por el virus HTLV-I, responsable de varias enfermedades como linfoma/leucemia en adultos y mielopatia asociada a HTLV-1. Para plantear un modelo que permita estudiar la evolución de la infección se toman en cuenta los aspectos biológicos que se detallan a continuación y que dan lugar a una serie de parámetros cuya variación determina el comportamiento del sistema y puede ser estudiada a través de la aplicación Shiny. Fuera de la célula, el virus no genera infección y generalmente se requiere contacto célula a célula para la misma. Además, se transmite verticalmente a células hijas durante la división mitótica, con un crecimiento que sigue la ley de crecimiento logístico. Las células infectadas mantienen la mayor parte de sus funciones celulares y su división es similar a la de las no infectadas, con una tasa de crecimiento constante. Dado que se modela el proceso en sangre periférica donde las células están lo suficientemente mezcladas, se asume una incidencia bilinial. Las células recién infectadas enfrentan una fuerte respuesta inmune por anticuerpos y linfocitos citotóxicos, por lo que sólo una fracción de las mismas resiste las reacciones del sistema inmune y permanece infectada. Por último, se considera que el cuerpo genera células T CD4+ a una tasa constante, las nuevas células son no infectadas y la tasa de remoción de células T CD4+ no infectadas e infectadas es constante. La aplicación permite simular el comportamiento del proceso de transimisión del virus bajo distintos escenarios para los parámetros, proveyendo una interfaz interactiva para analizar la evolución y estabilidad del proceso generado. Para su desarrollo se emplearon las siguientes herramientas:

• El libro de Michael Li, An Introduction To Math Modeling of Infectious Diseases (Springer, 2018), que detalla el modelo estudiado y el análisis de la estabilidad y dinámica global del sistema, presentado en la primera parte de la pestaña Análisis.
• El paquete de R EpiModel, que provee un conjunto de herramientas análiticas y gráficas para el estudio de este tipo de modelos y pueden observarse en las pestañas Gráficos, Resumen y Datos.
• El paquete de R phaseR, que permite estudiar sistemas de ecuaciones diferenciales, facilitando la construcción del plano de fases y el hallazgo y clasificación de puntos de equilibrio, resultados que se presentan en la segunda parte de la pestaña Análisis.

La aplicación se encuentra disponible en https://mpru.shinyapps.io/inhostshiny/ o puede ser ejecutada localmente a través del código:

```
library(shiny)
runGitHub("inhostshiny", "mpru")
```

para lo cual se necesita tener previamente instalados los paquetes EpiModel, phaseR, ggplot2, tidyr, dplyr, stringr and Deriv.

Steven Dutt Ross, Orlando Celso Longo, Luciane Ferreira Alcoforado & Lyncoln Sousa De Oliveira

Shiny Smart City: a cidade inteligente em tempo real

'Cidades inteligentes' é um termo que ganhou força na academia, nos negócios e no governo para descrever cidades que, por um lado, são cada vez mais monitoradas e, por outro, cuja economia está sendo impulsionada pela inovação e criatividade (KITCHIN et.al. 2013). Este trabalho tem o foco no primeiro caso e detalha como as cidades podem ser instrumentalizadas com sistemas digitais integrandos a uma infraestrutura que produz dados continuamente. Tais dados permitem a análise em tempo real da vida na cidade e fornecem a base para cidades mais eficientes, sustentáveis produtivas, abertas e transparentes.

Nesse sentido, vamos apresentar o conceito do Shiny Smart City (SSC). Este sistema tem a função de gerar informações a partir de sensores de cidades intelige ntes em tempo real. Essas informações geram um diagnóstico com o objetivo de antecipar os principais problemas das cidades inteligentes.

A construção de uma cidade inteligente tem como base a existência de uma infraestrutura tecnológica inteligente, isto é, a disseminação pelo espaço urbano de instrumentos eletrônicos para aquisição, tratamento e transmissão de dados (DE FARIAS et al. 2011). E tem como ápice a disponibilização de serviços inteligentes para os usuários dos serviços urbanos (pessoas, empresas, administrações públicas). LEMOS (2013) diz que a própria definição de cidade inteligente é relacionada com o processamento em tempo real. "Inteligente aqui é sinônimo de uma cidade na qual tudo é sensível ao ambiente e produz, consome e distribui um grande número de informações em tempo real." VILACA (et.al. 2014) apontam a importância dos sensores para a cidade inteligente. De acordo com os autores, nas cidades inteligentes os edifícios/construções inteligentes ganham destaque, onde sensores, atuadores, controladores, unidades centrais de programação, interfaces de diversos tipos, redes de comunicação e medidores inteligentes são instalados para garantir uma mel-

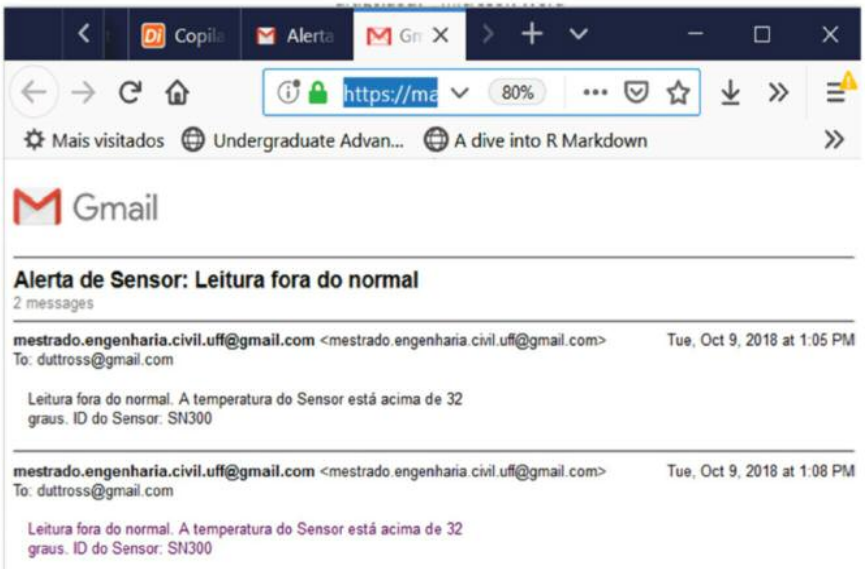


Figura 1- Notificação via email do SSC

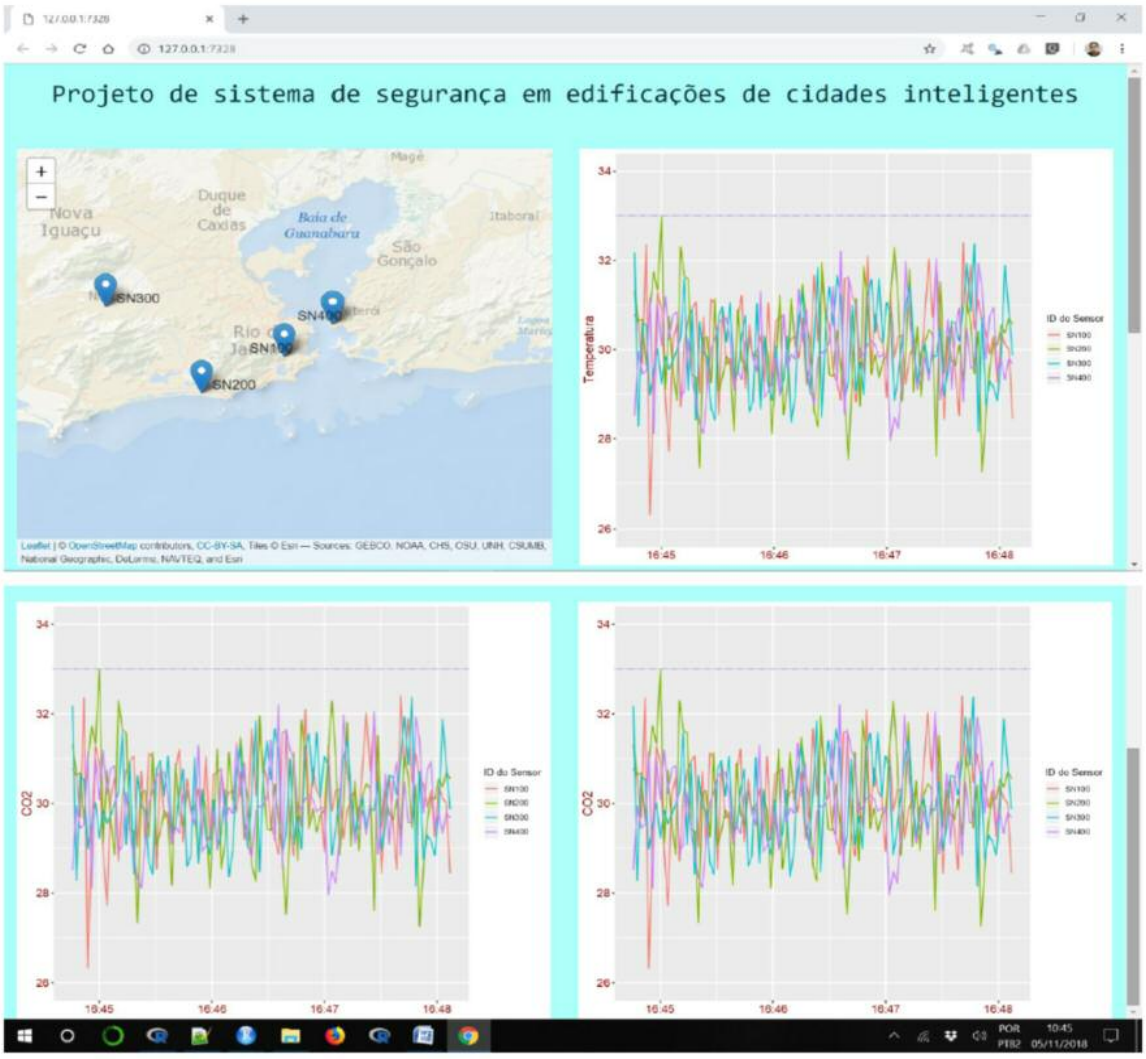


Figura 2- Dashboard SSC

hor performance energética da edificação. Shiny Smart City é um sistema baseado no R e no Shiny com o propósito de ler, carregar, analisar e visualizar dados de sensores de cidades inteligentes. O SSC é projetado para lidar com a informação em tempo real de sensores de última geração. Ele contém uma coleção de ferramentas para análise e tomada de decisão envolvendo as cidades inteligentes. Por exemplo, no caso de uma leitura fora do normal de um sensor ou da ausência prolongada da informação de um sensor, o SSC gera avisos/notificações por email e SMS antecipando os principais problemas das cidades inteligentes, como na figura a seguir:

Além do sistema de notificações/ avisos, o SSC tem um dashboard com indicadores para cada sensor, conforme a figura 2.

Tivemos diversos desafios e dificuldades durante a construção deste projeto. O principal desafio foi a integração entre o Shiny, o Arduino e o PostgreSQL. Testamos diversas alternativas. Entre elas destacamse os pacotes Rduino e ArduinoR. Todavia, a solução que funcionou melhor foi por meio do pacote serial (SEILMAYER, 2019). Este pacote fez a integração do Arduino com o Shiny. Já a integração do PostgreSQL com o R foi utilizado o RPostgreSQL (Conway et al. 2019).

Além deste desafio, a principal dificuldade foi a alimentação "em tempo real" (streaming data). A solução desenvolvida foi no próprio shiny. Este pacote tem uma função chamada reactivePoll. Com esta função, foi possível consultar as informações dos sensores do Arduino de 03 em 03 segundos. Em seguida, a informação do sensor

era guardade no banco de dados no formato PostgreSQL que já esttava conectado com o shiny e a função reactivePoll atualizava o shiny com essa nova informação do Arduino.

O sistema foi instalado, primeiramente como protótipo, no monitoramento do prédio de engenharia de produção da UFF. Este protótipo é composto de uma configuração mínima de tal forma que permite a implementação básica para teste, bem como o seu funcionamento. O programa foi desenvolvido nalinguagem R, que acessa um Banco de Dados no PosgreSQL e calcula em intervalos de 03 (três) segundos a posição do sensor Arduino.Atualmente localizada no Campus da Praia Vermelha, a Escola de Engenharia conta com 2 prédios (Blocos "D" e "E") de, aproximadamente, 15.000 m2 de área física.

Referências

- CONWAY, Joe ; EDDELBUETTEL, Dirk ; NISHIYAMA, Tomoaki ; KUMAR, Sameer Prayaga e TIFFIN, Neil. RPostgreSQL: R-Interface to the 'PostgreSQL' Database System. R package version 0.6-2. https://CRAN.Rproject. org/package=RPostgreSQL, 2017.
- EWERTON P. DE FARIAS, José et al. Cidades Inteligentes e Comunicações. Revista de Tecnologia da Informação e Comunicação, [S.l.], v. 1, n. 1, p. 28-32, out. 2011. ISSN 2237-5104.
- KITCHIN, Rob, The Real-Time City? Big Data and Smart Urbanism (July 3, 2013). A revised version of this paper, including two new sections, has been published in GeoJournal 79(1):1-14, 2014.
- LEMOS, André Cidades inteligentes GV-executivo v. 12, n. 2 (2013) H CHEN, RHL CHIANG, VC STOREY Business intelligence and analytics: from big data to big impact MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012.
- SEILMAYER, Martin. serial: The Serial Interface Package. R package version 2.1.3. https://CRAN.R-project.org/package=serial 2019.
- VILACA,Natalia M.C.A.A. ; FIGUEIREDO, Vinicius N. ; OLIVEIRA, Lorenna Baptista de; FERREIRA, Vitor H. FORTES,Marcio Z. , CORREIA, Weules F. ;PACHECO, Orestes L.C. SMART CITY – CASO DA IMPLANTAÇÃO EM BUZIOS - RJ Revista SODEBRAS –Volume 9 – N 98 – FEVEREIRO/ 2014.
- SUH, Chris ; HOFF, Peter Rduino: A Microcontroller Interface, 2017.
- ZHU, Hao ArduinoR: An easy way to get arduino data into R, 2019.

Antonio Vazquez Brust, Tomás Olego & Germán Rosati

Mapeando la Vunerabilidad Sanitaria en Argentina con R

Introducción

La noción de vulnerabilidad sanitaria está relacionada con los llamados determinantes de salud. Existen ciertos factores y variables que se vinculan fuertemente con el estado de salud -en sentido amplio, biológica, psicológica, social- de una persona o población. Ante a la ausencia o insuficiencia de estos determinantes, se produce un estado de vulnerabilidad. La presentacion resumirá la metodología y principales herramientas basadas en R utilizadas para producir un Indice de Vulnerabilidad Sanitaria de Argentina, que combina varias dimensiones con alta resolución espacial. Proyectado en el mapa, el índice muestra con precisión, incluso a escala intraurbana, regiones donde podría priorizarse la asignación de recursos públicos para mejorar el acceso a salud de población desatendida.

Métodos y herramientas

La recopilación y generación de los datasets que alimentaron al Indice requirió "minar" y combinar fuentes de datos oficiales -muchas veces dispersas- así como repositorios de datos producidos colaborativamente (crowd-sourced). El proyecto se benefició del uso de paquetes de R especializados para la resolución de los muy diversos problemas a enfrentar. Uno de los datasets en particu-

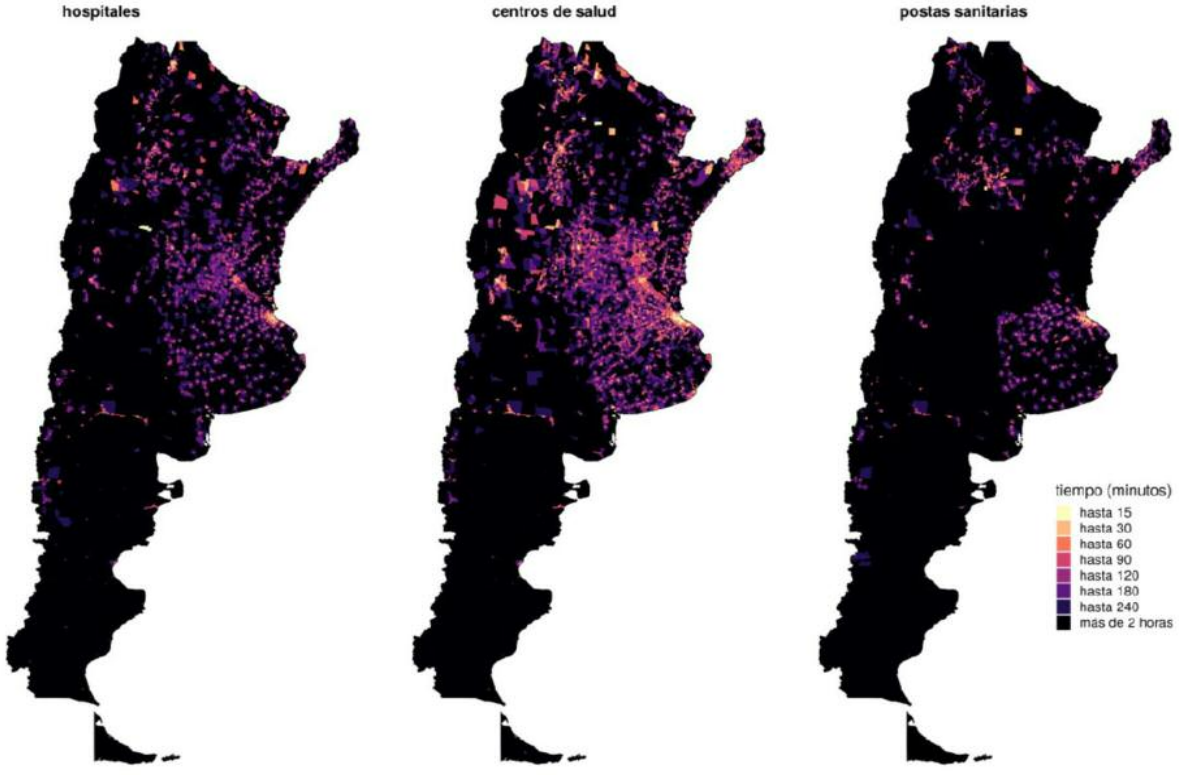


Figura 1. tiempo a pie hasta el efector de salud pública más cercano, por tipo

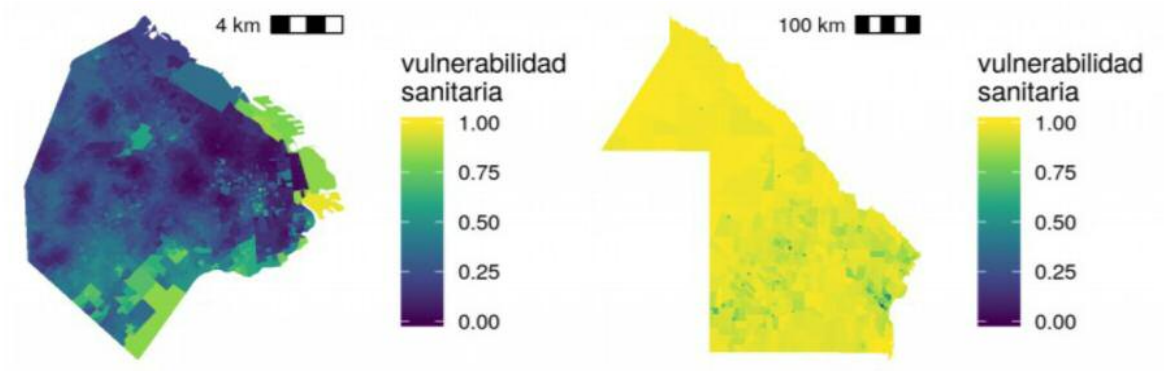


Figura 2. Mapas de Indice de Vulnerabilidad Sanitaria para la Ciudad de Buenos Aires (izq) y la Provincia de Chaco (der)

lar, producido para estimar un nivel de "acceso a la salud pública", requirió compilar tanto la posición de los miles de sitios donde se efectúa la salud pública en Argentina, como el tiempo de viaje a pie desde cada punto del país al sitio más cercano. El desafío fue resuelto diseñando un pipeline de ingesta y procesamiento de datos basado en paquetes especializados: rvest para el scraping de fuentes oficiales no publicadas en formato abierto, ggmap para la georeferenciación de direcciones de hospitales y otros efectores de salud, osrm para la estimación de rutas recorriendo a pie la grilla vial argentina, y purrr para automatizar y organizar el cálculo de rutas y tiempos de viaje para millones de trayectos posibles.

Obtener el índice final requirió combinar el nivel de acceso a salud con otras variables georeferenciadas -densidad poblacional, nivel socioeconómico-, tarea apoyada por el paquete sf que permitió transformar y cruzar capas de información vertidas sobre la geografía censal, y en conjunto con ggplot, una visualización detallada de los resultados para su análisis y comunicación.

Formato de la presentación

Se propone una breve introducción al tema de investigación (2'), una descripción de los pasos de análisis requeridos para el estudio, y del stack de herramientas de R aplicadas (10'), y un vistazo a los mapas obtenidos y sus implicaciones (3') .

Tópicos de interés representados

- Aplicación de R en Salud
- Investigación reproducible usando R
- Uso de R en conjunto con otros lenguajes de programación y plataformas
- Análisis de grandes datos con R
- Visualización de datos con R
- Uso de R para análisis de datos abiertos

Nicolas Sidicaro

Extracción de datos de redes sociales para monitorear el servicio de transporte de la Ciudad de Buenos Aires

En esta ponencia se presenta un estudio sobre el funcionamiento del subterráneo de la Ciudad de Buenos Aires realizado mediante el uso de la herramienta R. El trabajo se basa en el registro y procesamiento de la información pública brindada por la empresa Metrovías S.A, responsable del servicio de transporte subterráneo de la Ciudad desde hace 25 años, a través de su cuenta oficial de Twitter. Los datos de esa red social se procesaron mediante una codificación original elaborada ad hoc para su posterior sistematización y análisis.

El transporte es un derecho esencial de los habitantes de una ciudad dado que afecta el bienestar cotidiano de quienes deben utilizarlo para trabajar, estudiar o realizar diversas actividades. En este sentido, el transporte público cobra una importancia aún mayor ya que permite la movilización de grandes masas de personas a un costo menor que el transporte privado y descongestiona las ciudades.

En la Ciudad de Buenos Aires circulan seis líneas de transporte subterráneo (A, B, C, D, E y H) gestionadas por la empresa Metrovías S.A. desde el 1ro de enero de 1994 cuando se privatizó la provisión de este servicio previamente gestionado por la estatal SBASE. El recorrido total de la red es de 60,8 km, de los cuales 54,7 km los son de tramo comercial, en los que se distribuyen 87 estaciones. En 2016, el conjunto de las líneas transportó a 303.859.273 pasajeros, según informó el Ministerio de Desarrollo Urbano y Transporte de la Ciudad. Pese a la relevancia del conocimiento sobre el funcionamiento del transporte público, tanto del saber especializado como de aquel que afecta el uso cotidiano de los habitantes, la Ciudad de Buenos Aires actualmente no presenta estadísticas de acceso público y rápido sobre las problemáticas presentadas en la prestación del servicio de subterráneos. Esta falta de información obstaculiza los debates públicos sobre la calidad del servicio. Este estudio representa la única fuente de información de acceso público y gratuito actualizada mes a mes con los datos brindados en forma virtual y desagregada por la empresa que presta el servicio. A su vez, abre la posibilidad de replicar esta aplicación en cada ciudad con sistemas de información vía Twitter sobre las problemáticas presentes en los sistemas de transporte. Con respecto a la metodología utilizada, se relevaron los tweets de la empresa que se encuentran formateados de la siguiente manera (Imagen 1): línea que presenta un problema, la característica del problema y la hora de publicación del tweet.

Cada tweet se extrajo con el paquete "twitteR", a través de la función userTimeline. Una vez realizada la extracción mediante las fun-

Texto	Creación	Línea	Fecha	Hora	Mensaje	Justificción
#LineaE - Circula con demora. 08:18	2018-10-31 11:18:42	#LineaE	2018-10-31	11:18	demora	no justificado
#LineaA - Por obras en zona de vias. 05:17	2018-10-31 8:17:40	#LineaA	2018-10-31	8:17	limitado	justificado
#LineaD - Circula con demora. 18:22	2018-10-30 21:22:35	#LineaD	2018-10-30	21:22	demora	no justificado
#LineaB - Servicio limitado. 14:57	2018-10-30 17:57:33	#LineaB	2018-10-30	17:57	limitado	no justificado
#LineaB - Circula con demora. 14:50	2018-10-30 17:50:32	#LineaB	2018-10-30	17:50	demora	no justificado

ciones de “R base” y del paquete “tidyverse” (fundamentalmente se utilizaron los paquetes “dplyr”, “lubridate” y “stringr”), se procedió a desagregar el mensaje de Twitter según diferentes dimensiones de forma de poder analizar cada aspecto en particular como se observa en la siguiente tabla a modo de ejemplo.

El reporte de los datos es accesible de manera libre y gratuita a través de un link de la plataforma Data Studio. Se espera que el estudio constituya un aporte en el acceso a información sobre servicios públicos y amplíe las posibilidades de debate ciudadano.

Stephanie Orellana

Utilizar R como un Sistema de Información Geográfica y hacer análisis reproducible

Los Sistemas de Información Geográfica (SIG) permiten almacenar, organizar, manipular y hacer análisis de datos que posean una referencia espacial (coordenadas). Son utilizados principalmente para realizar cruces de información de diferente tipo (social, ecológica, política, física) para mejorar la toma de decisiones en diferentes áreas, tales como: ciencias de la tierra, ciencias sociales, ingeniería en transporte y aplicaciones más recientes en marketing y ventas, entre otras.

Existen conocidos SIG, como ArcGis (comercial) y Qgis (libre) que son los más utilizados, pero que no siempre permiten almacenar el proceso en un código reproducible, ya que: 1) se centran más en la interfaz y 2) la ejecución de código es vista como herramientas para usuarios más avanzados (ambos permiten código Python en una interfaz de código básica). En el último tiempo, las librerías que permiten realizar análisis espacial en R han tenido una gran mejoría y se encuentran en constante desarrollo, tanto así, que ahora es posible realizar la mayoría de las acciones que permiten los SIG convencionales, con la ventaja de poder generar un código reproducible y realizar labores iterativas en bucle de forma fácil sin tener que aprender nuevos lenguajes ni instalar (o crackear) un software comercial. Además, se están desarrollando librerías que permiten conectar R con los SIG tradicionales para utilizar herramientas más específicas.

Utilizar R tiene otras ventajas, como poder hacer mapas y figuras uniformes (colecciones de mapas), realizar mapas interactivos con leaflet y visualización mediante Shiny.

El objetivo de la presentación será mostrar las diferentes tareas que se pueden hacer con R en este ámbito, mostrando su uso en casos aplicados con entidades vectoriales y raster. También, hacer hincapié en las ventajas de realizar todos los procesos con líneas de código para hacer investigación reproducible, poder realizar cambios "más fáciles" cuando se proponen nuevas metodologías, actualizaciones más rápidas y trazabilidad de errores.

Y mostrar que R no es sólo un software estadístico o para hacer los gráficos, sino que es eso y –mucho más!

Específicamente hablaré sobre los paquetes más conocidos para llevar a cabo un flujo típico de análisis reproducible:

1) Leer datos (paquetes: raster, sf)

2) Manipular datos (paquetes: sp, rgdal, gdalcubes)

3) Analizar datos, generar estadísticas, gráficos o mapas (paquetes: dismo, ggplot, rastervis, leaflet, etc.)

4) Comunicar resultados (Shiny, Rmarkdown)

Se mostrarán dos aplicaciones, la primera es utilizando el paquete dismo, el cual permite realizar modelamiento de nicho ecológico con base en variables físicas/climáticas y a datos de ocurrencia de especies. Mostraré cómo desde del mismo paquete se pueden descargar datos desde gbif (Global Biodiversity Information Facility https://www.gbif.org/) para realizar este tipo de estudios. También mostraré cómo se ha facilitado el proceso de análisis de imágenes satelitales con la nueva librería "gdalcubes", mostraré un pequeño ejemplo de análisis de temperatura superficial con datos del sensor MODIS (Moderate-Resolution Imaging Spectroradiometer) los cuales se pueden descargar de forma libre desde la web de la NASA. Dejaré los códigos de ambos análisis en github para que los participantes puedan acceder a ellos.

Rocío Joo, Matthew E Boone, Thomas A Clay, Samantha C Patrick, Susana Clusella-Trullas and Mathieu Basille

R en movimiento: una revision de los paquetes para analizar movimiento dirigida a usuarios y desarrolladores

Abstract

El creciente avance de la tecnologíia de seguimiento o tracking (e.g. GPS, camaras de video, acelerómetros), esta dando paso a grandes cantidades de datos que permiten monitorear el movimiento tanto de seres humanos como de animales. En paralelo, se van desarrollando variadas y sofisticadas herramientas para procesar, visualizar y analizar datos de seguimiento. Solo en R, hemos encontrado 59 paquetes focalizados en estas tareas (e.g. adehabitatLT, move).

En esta presentación, revisamos y describimos brevemente los 59 paquetes, que llamamos paquetes de seguimiento, basandonos en un flujo de trabajo centrado en este tipo de datos y dividido en tres etapas: preprocesamiento, post-procesamiento y analisis. El analisis lo dividimos a su vez en visualizacion de datos, descripcion de trayectorias, reconstruccion de trayectorias, identificacion de patrones comportamentales, caracterizacion del uso del espacio, simulacion de trayectorias y otros. Utilizando un analisis de redes, evaluamos

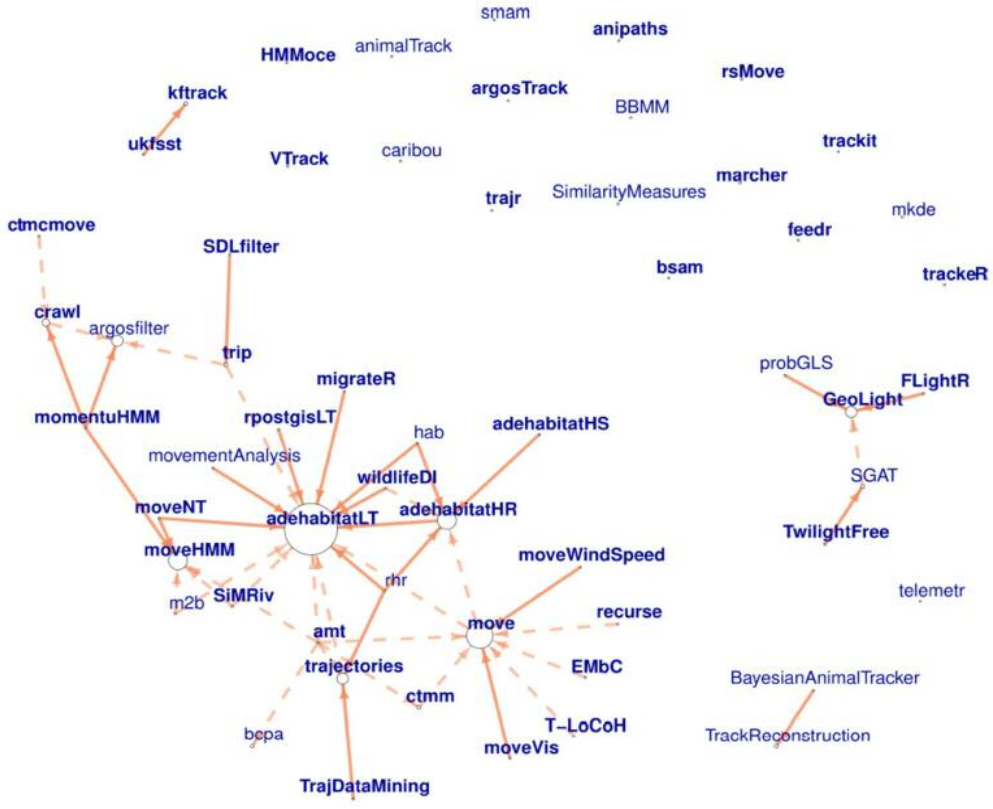


Figura 1: Representacion de la red de dependencias y sugerencias entre paquetes de seguimiento. Las flechas van hacia el paquete que los otros sugieren (flechas discontinuas) o dependen de (flechas solidas). Las fuentes en negrita corresponden a paquetes activos (ultima actualizacion hecha hace menos de 1 año). El tamaño del círculo es proporcional al número de paquetes que sugieren o dependen de éste.

las relaciones entre los paquetes y mostramos que un tercio de ellos trabaja en aislamiento. Este resultado es un reflejo de una fragmentacion en la comunidad de programadores del movimiento en R (Fig. 1).

A la luz de los resultados encontrados, damos algunos criterios para potenciales usuarios para escoger paquetes, y sobre todo, criterios para desarrolladores, con el proposito de maximizar la utilidad de su paquete y fortalecer los vınculos dentro de la comunidad programadora. Si bien estos criterios estan inspirados en los paquetes de seguimiento, pueden ser extrapolados a todo tipo de paquete en R.

El manuscrito relacionado a este trabajo se encuentra en: https://arxiv.org/abs/1901.05935; los datos sobre cada paquete se encuentran en https://doi.org/10.5281/zenodo.3066226.

Consideramos que este trabajo puede ser un punto de partida para discutir, desde America Latina, la relevancia de nuestras propuestas de innovación y desarrollo de la mano de R.

Brenda V. Canizo, Ana Laura Diedrichs, Emiliano F. Fiorentini, Roberto G. Pellerano, Rodolfo G. Wuilloud & Juan M. Juricich

Determinación del origen geográfico de mieles de Mendoza (Argentina) mediante análisis multielemental y árboles de decisión en R

Introducción

La miel es un alimento natural complejo producido por las abejas *Apis Mellifera* a partir del néctar, de las secreciones de plantas o de la excreción de insectos. Actualmente, la autenticidad de la miel se ha convertido en un tema de interés tanto para los productores como los consumidores, quienes reclaman información precisa sobre la procedencia de los alimentos y, por otro lado, los productores quieren evitar la competencia desleal y agregar valor a los productos que ofrecen.

La composición de la miel está relacionada con su origen botánico, que está estrechamente asociado a la zona geográfica en la que se origina, debido a las características del suelo y las condiciones climáticas estacionales, pero también está vinculada a factores como las especies de abejas, la temporada de producción y el almacenamiento, y las prácticas agrícolas en general. La composición elemental parece ser un buen indicador para los sistemas de clasificación geográfica, aunque esta sólo representa el 0.1-0.2% de la composición total de la miel, lo minerales se transportan desde el suelo a las plantas y las flores a través del sistema radicular, pasan al néctar y finalmente a la miel. Según la FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación), Argentina está posicionada como el tercer mayor productor y exportador mundial de miel, exportando alrededor del 95% de su producción total, que es reconocida en todo el mundo por su alta calidad. Dentro de las regiones apícolas de Argentina, la provincia de Mendoza es considerada una de las más importantes del país.

En el software libre R, existe una variedad de algoritmos de clasificación y dentro de ellos, los árboles de decisión son un método de fácil interpretabilidad, pues nos dan un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Además son algoritmos que no son demandantes en poder de cómputo comparado con procedimientos más sofisticados y dan buenos resultados de predicción para muchos tipos de datos.

El objetivo de este trabajo fue evaluar la capacidad de clasificación predictiva de la herramienta estadística multivariable "árboles de decisión" combinada con un análisis multielemental para diferenciar mieles de diferentes regiones apícolas de Mendoza (Argentina) de acuerdo con su región de origen. La originalidad del trabajo se enfoca en la aplicación de distintos algoritmos en R en este tipo de muestras de Mendoza para lograr la primera clasificación regional.

Materiales y métodos

Se recolectaron 154 mieles de tres regiones apícolas de la provincia de Mendoza (Argentina): noreste (Lavalle, Rivadavia y Santa Rosa), Valle de Uco (San Carlos, Tunuyán y Tupungato) y sur (General Alvear, Malargüe y San Rafael). Las muestras se mineralizaron mediante una digestión ácida con HNO3 concentrado en vaso abierto, aplicando un tratamiento térmico sobre una placa caliente durante 45 minutos a 100 C. Las determinaciones analíticas se realizaron utilizando un espectrómetro de masas por plasma acoplado inductivamente (ICP-MS). Los isótopos detectados, en orden creciente del número de masa, fueron los siguientes: 7Li, 23Na, 24Mg, 27Al, 44Ca, 47Ti, 51V, 52Cr, 55Mn, 56Fe, 59Co, 60Ni, 63Cu, 66Zn, 75As, 78Se, 85Rb, 88Sr, 95Mo, 105Pd, 107Ag, 111Cd, 118Sn, 121Sb, 201Hg, 207Pb y 208Pb.

Para el análisis estadístico, debido al desbalance en la cantidad de muestras entre las tres regiones y la importancia de la región del Valle de Uco

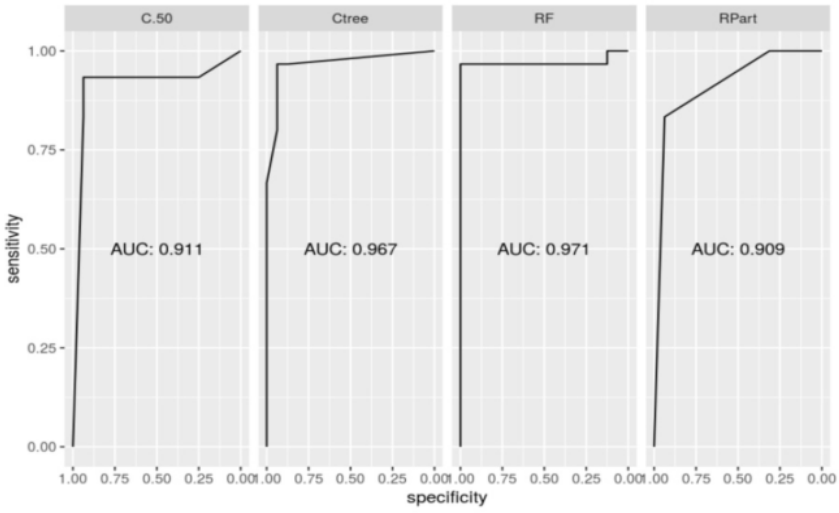


Fig. 1. Curvas ROC de los cuatro modelos probados.

como un poderoso polo apícola gracias a sus condiciones favorables para el desarrollo de la actividad, con ventajas que se refieren al clima y la cubierta vegetal, se procedió a clasificar miel de la región Valle de Uco de las demás regiones de Mendoza. Por lo tanto, todas las muestras que pertenecieron a la misma región fueron asignadas por el mismo código "V" para las mieles del Valle de Uco y "O" para las mieles de las regiones del sur y noreste. Antes de aplicar los algoritmos de clasificación, se realizó un pre-procesamiento del conjunto de datos mediante la estandarización de la matriz. Todos los análisis estadísticos se calcularon utilizando el software R versión 3.3.0 con paquetes de árbol de decisión y caret. Se evaluaron cuatro algoritmos: recursive partitioning and regression trees (rpart), conditional inference tree (ctree), C5.0 y random forest (RF), utilizando la curva ROC (reciver operating characteristic), la cual es una representación gráfica que se utiliza para evaluar un sistema clasificador binario y seleccionar el modelo óptimo teniendo en cuenta AUC (área bajo la curva en su sigla en inglés). Para el desarrollo del proceso de clasificación, fue necesaria una separación del dataset en muestras de entrenamiento (70%) y de prueba (30%) y se aplicaron validaciones cruzadas de k-fold (CV) para la optimización del modelo.

Resultados y Discusión

Luego del entrenamiento del dataset mediante 10-fold validación cruzada para construir los diferentes clasificadores, se seleccionó el modelo con el valor más grande de ROC para cada algoritmo de árbol de decisión. En los modelos de clasificación construidos la variable dependiente fue la región (V y O), y todos los elementos analizados se aplicaron como predictores. Se realizó una comparación del AUC para los diferentes clasificadores de árboles de decisión para la discriminación de origen de la miel en el conjunto de prueba. La Fig. 1 presenta la curva ROC específica para cada algoritmo evaluado para la clasificación de la miel de Mendoza. Los resultados mostraron un rendimiento similar para la predicción de la procedencia de la miel con un AUC superior a 0,9, que es un muy buen resultado que equilibra la sensibilidad y la especificidad. Los modelos de RF (AUC = 0.971) y ctree (AUC = 0.967) tienen el AUC más alto. Entre RF y ctree, este último puede interpretarse fácilmente como un solo árbol en lugar de un bosque (conjunto de árboles que es RF). C5.0 (AUC = 0.911) y rpart (AUC = 0.909) mostraron un poder de predicción similar. Esto se condice con los valores de accuracy obtenidos de 0.96 para RF y ctree, 0.93 para C5.0 y de 0.87 para el algoritmo rpart.

Conclusión

Con la aplicación de cuatro algoritmos de árbol de decisión, RF, ctree, rpart y C5.0, para discriminar el origen geográfico de miel de Mendoza, fue posible alcanzar una excelente clasificación del origen de este producto. Los resultados demostraron que los árboles de decisión son una forma simple y ágil de generar predictores altamente informativos de muestras de miel que se pueden usar para el control de la calidad y la verificación del origen geográfico.

Yuriko Sosa Paredes & Andrés Burga León

Análisis de patrones de respuesta en una evaluación estandarizada bajo el marco del análisis forense de datos

Introducción

Si bien existen maneras legítimas de mejorar los resultados en una prueba estandarizada muchas veces diversos actores del sistema educativo se sienten motivados para ayudar a los estudiantes a resolver las pruebas estandarizadas aplicadas a gran escala o incluso a modificar sus respuestas después de haber rendido la evaluación. Cada vez que existan consecuencias asociadas a los resultados de las pruebas habrá mayores amenazas a la seguridad de los test y por lo tanto posible presencia de trampa. Esta situación ha creado la necesidad de contar con estrategias de análisis de datos que ayuden a identificar este tipo de comportamientos irregulares, las cuales en conjunto se conocen como análisis forense de datos (data forensics). Existen diversos indicadores para analizar la posibilidad de copia, una de ellas es el estudio de la similitud de patrones entre estudiantes. Se ha investigado el funcionamiento de diversos índices, encontrando que tanto el omega (ω) de Wollack, como el Generalized Binomial Test (GBT) muestran resultados adecuados. Hay que considerar que un indicador estadísticamente significativo no necesariamente implica copia. Se presenta una aplicación en las evaluaciones censales aplicadas a los alumnos de 2. º grado de primaria en las áreas de Comunicación (Lectura) y Matemática, durante los años 2014, 2015 y 2016 en Perú.

Materiales y métodos

Descripción de los datos

Se trabajó con tres evaluaciones censales de medio millón de estudiantes de 2. º grado de primaria, aplicadas en los años 2014, 2015 y 2016, donde se evaluaron dos áreas: Comunicación (Lectura) y Matemática. En total, estos alumnos rindieron 42 preguntas en la prueba de Matemática y 46 en la de Comunicación.

Además de las variables de respuesta a cada una de las preguntas, se incluyeron variables de gestión educativa (no estatal, estatal), área geográfica (urbano, rural) y la Dirección Regional Educativa (DRE) a la que pertenece la escuela.

Métodos

El índice ω de Wollack (2003, 2006). Es una aproximación normal a la distribución binomial compuesta para el número de respuestas correctas e incorrectas idénticas entre dos vectores de respuestas; además es un estadístico asimétrico, quiere decir que se puede explorar la similitud al comparar la copia en ambas direcciones (p.ej. X copia de Y, e Y copia de X) y dividir el nivel de alfa (Zopluoglu, 2017). Se utilizó una modificación del paquete CopyDetect2 (Zopluoglu, 2012).

Generalized Binomial Test (GBT) de van der Linden y Sotaridona (2006). Se estima como la probabilidad de observar m o más coincidencias en N ítems; además es un indicador simétrico. Se utilizó una modificación del paquete CopyDetect2 (Zopluoglu, 2012)

Agregados a nivel de sección. Se hicieron todas las comparaciones pareadas entre los estudiantes de una sección. El indicador agregado implicó, a nivel de sección, qué porcentaje de esas comparaciones fueron estadísticamente significativas. Para esto también se generó un script en R, utilizando el tidyverse.

Reportes generales y por Dirección Regional Educativa. Se realizaron reportes automatizados en Rmarkdown para mostrar los resultados a nivel nacional, por estratos y por cada una de las 26 Direcciones Regionales Educativas que tiene el país.

Resultados

A nivel nacional, en los tres años estudiados, la correlación de ambos indicadores es fuerte en Lectura (0,85; 0,88 y 0,95) y moderada para Matemática (0,59; 0,84 y 0,84); por lo que se ha optado por presentar los resultados con solo un indicador, GBT.

En el año 2014, en la prueba de Lectura, a nivel nacional, el 73,5 % de secciones tiene entre 0 y 1% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo. Este porcentaje es mucho más grande en comparación a lo encontrado en el 2015 (45,3 %) y 2016 (50,2 %).

La mayoría de las secciones (aproximadamente el 85 %) tiene menos del 20 % de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo en los años 2014 y 2015. Mientras que en el 2016 es el 73 % de secciones.

En el año 2014, en la prueba de Matemática, a nivel nacional, el 45,2 % de secciones tiene entre 0 y 1% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo. Este porcentaje es casi el doble en comparación a lo encontrado en el 2015 (22,5 %) y 2016 (24,7 %).

La mayoría de las secciones (aproximadamente el 75 %) tiene menos del 20% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo en el 2014. Mientras que en el 2016 es aproximadamente el 64 % y menos en el 2015 (48,3 %).

Referencias

- van der Linden, W. & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. Journal of Educational and Behavioural Statistics, 31(3), 283-304. https://doi.org/10.3102/10769986031003283
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. Journal of Educational Measurement, 40(3), 189-205. https://doi.org/10.1111/j.1745-3984.2003.tb01104.x
- Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. Applied Measurement in Education, 19(4), 265-288. https://doi.org/10.1207/s15324818ame1904_3
- Zopluoglu, C. (2012). CopyDetect: an R package for computing statistical indices to detect answer copying on multiple-choice examinations. Applied Psychological Measurement, 37(1), 93–95. doi:10.1177/0146621612463119
- Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: understanding the status quo. En G.J. Cizek & J.A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 25-46). New York: Routledge.

Log-linear models structure comparison

Abstract

Log-linear models have been largely used for representing the structure of probabilistic models containing context-specific interactions among variables. Automatic structure learning of these models from data is a popular Machine Learning problem, for which the comparison of models' structures is an important aid for the evaluation of algorithms. Nowadays, structure learning, weight learning and inference with this class of probabilistic models have a wide variety of applications in the literature of Machine Learning, such as computer vision and image analysis, language processing, computational biology and biomedicine, among many others. Existent measures such as Kullback-Leibler divergence do not allow the direct comparison of structural differences; instead, they compare the full distribution (structure and parameters of the models), which introduces several limitations. We propose a distance measure for comparing the dependence structures of log-linear models, similar to the Hamming distance commonly used for models represented by undirected graphs. Our method is proven to be a metric, and can be efficiently computed in terms of the number of variables of the domain. It can be used for the comparison of structure learning techniques and for obtaining insights from log-linear models.

Introduction

A log-linear model over a discrete domain $X = \{X_1; ...., X_N\}$ is defined as a set of feature functions $F = \{f_K(X_K)\}$, each one defining a numerical value for each assignment x_K to some subset X_K C X. Given the set F, the parameters of the log-linear model are the weights $\theta = \{\theta_K : f_K \in F\}$ The overall distribution is defined as

$$P(x) = \frac{1}{Z(\theta)} exp(\sum_{f_K \in F} \theta_K f_K(x_K))$$

where Z(o) is the partition function that ensures that the distribution is normalized (i.e., all entries sum to 1) [9, 8, 3, 5].

Unlike undirected graphical models, log-linear models are a more accurate representation when the distribution underlying the data contains context-specific independences (i.e., conditional independences that hold only for some values of the conditioning variables) [1, 6], which makes them better suited for inference tasks in those cases. In addition, there are a number of graph-based representations that allow a qualitative interpretation of these models [2, 7]. Nevertheless, a straight-forward comparison of their independence structure would entail a superexponential number of comparisons. For this reason we have designed a method that takes advantage of the

compactness of the feature representation of log-linear models in order to allow for an efficient computation of discrepancies between two structures, while also providing theoretical guarantees.

2 Approach

Our method consists in computing a confusion matrix, comprising the well-known measures of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), obtained from the dependence structure of the models. To illustrate, for comparing undirected graphical models, the Hamming distance of the edges between two graphs is used as an ad-hoc measure, where the distance is equivalent to the discrepancies between a subset of independence assertions entailed by the structures (given by the pairwise Markov property). In our approach, the comparison is made between a subset of assertions of context-specific independence entailed by the dependence structure, and it is computed directly based on

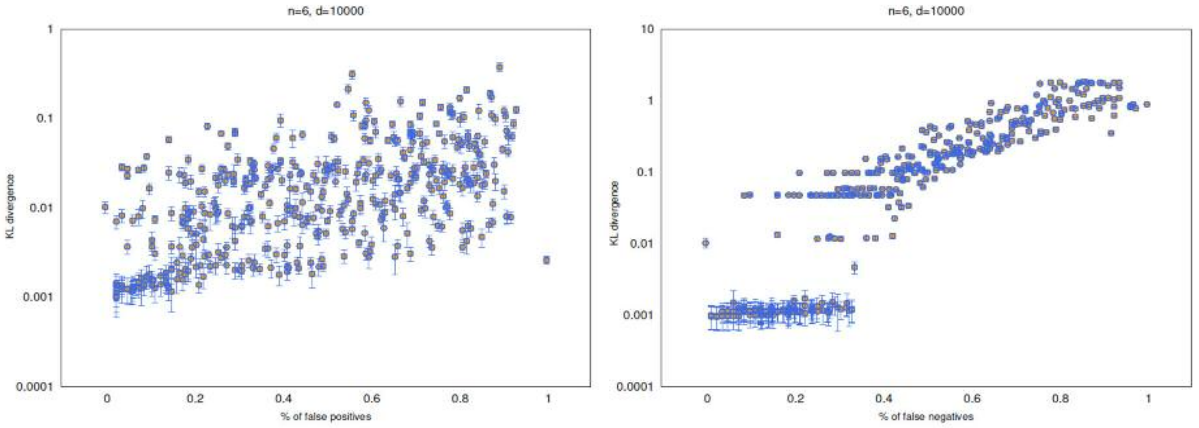


Figure 1: Comparison of errors for the proposed log-linear measure (xaxis) vs KL-divergence (y-axis) for synthetic datasets (6 variables, 10000 observations). Each point in the graph is one structure. X-axis shows % of FPs (left) and % of FNs (right) as computed by our measure.

the features of the log-linear representation. Additionally, we provide proof that the proposed measure is a metric.

2.1 Example: Discriminating False Positives From False Negatives

We illustrate how our measure differs from KL-divergence [4] (the most commonly used measure for comparison of distributions) by generating synthetic graphs with different numbers of false positives and false negatives, and visualising the results. We use R for preprocessing of results and visualisation. We show an example of our results (for a synthetic model of 6 variables) in Figure 1.

References

[1] S. Della Pietra, V. Della Pietra, and L. J. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4):390–393, 1997.

[2] S. Højsgaard. Statistical inference in context specific interaction models for contingency tables. Scandinavian journal of statistics, 31(1):143–158, 2004.

[3] D. Koller and N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. The annals of mathematical statistics, 22(1):79–86, 1951.

[5] S. Lauritzen. Graphical Models. Oxford University Press, 1996.

[6] A. McCallum. Efficiently inducing feeatures of conditional random fields. Proceedings of Uncertainty in Artificial Intelligence (UAI), 2003.

[7] H. Nyman, J. Pensar, T. Koski, J. Corander, et al. Stratified graphical models-context-specific independence in graphical models. Bayesian Analysis, 9(4):883–908, 2014.

[8] J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1988.

[9] P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. Adaptive Computation and Machine Learning Series. MIT Press, 2nd edition, January 2000.

Federico Molina, Natalia da Silva, Ignacio Ignacio Alvarez-Castro y Juan José Goyeneche

Evaluación y monitoreo de plataformas educativas

Resumen

El Plan Ceibal se implementa en Uruguay como una política pública de carácter universal que forma parte de la iniciativa mundial One Laptop per Child (OLPC). El Plan Ceibal genera una enorme cantidad de información a nivel individual en tiempo real del uso de las ceibalitas (computadora o tablet entregada a cada niño que asiste a la educación pública Uruguaya). En este trabajo se presentan herramientas para el monitoreo de plataformas educativas, en especial se analiza la plataforma CREA2, a distintos niveles de análisis (clase, grado, escuela, departamento, etc) y a distintas ventanas temporales con sus correspondientes indicadores. A su vez se presentarán modelos predictivos de aprendizaje educativo de Inglés con información proveniente de pruebas adaptativas de Inglés realizadas en 2015-2016. En este trabajo se utiliza intensivamente las herramientas del tidyverse para la limpieza, transformación y exploración de datos. Se desarrolla una herramienta de sencillo uso para comunicar, evaluar y gestionar los resultados mediante una aplicación shiny que por cuestiones de confidencialidad no podemos compartir aunque la mostraremos en la presentación con los cuidados correspondientes. La aplicación a su vez genera un reporte dinámico con rmarkdown e incorpora visualizaciones interactivas mediante plotly.

Introducción

El Plan Ceibal se implementa en Uruguay como una política pública de carácter universal que forma parte de la iniciativa mundial One Laptop per Child (OLPC). El Plan Ceibal inicia en el año 2007 como un plan de inclusión e igualdad de oportunidades con el objetivo de apoyar con tecnología las políticas educativas uruguayas. Este plan pone la tecnología al servicio de la educación para mejorar su calidad e impulsar procesos de innovación social, inclusión y crecimiento personal. Su visión es inspirar a cada niño y adolescente del Uruguay para que desarrolle al máximo su potencial de aprendizaje, su creatividad y su pensamiento crítico en la era del conocimiento. Resulta crucial explotar al máximo la información disponible de la evolución de los alumnos con distintas métricas que puedan resultar relevante en potenciar las habilidades del estudiante y, ser de utilidad para el diseño de políticas educativas.

Dentro del Plan Ceibal existen distintas plataformas virtuales de aprendizaje, una de ellas es CREA2 que tiene una lógica de red social educativa que dinamiza los aprendizajes mediante la colaboración y comunicación constante entre alumnos y docentes. El presente trabajo busca la implementación de herramientas estadísticas que agreguen valor a los datos de la plataforma educativa CREA2 y los resultados de las pruebas adaptativas de Inglés para los años 2015 y 2016.

Los objetivos generales del proyecto son predecir los aprendizajes educativos de Inglés en base a información de uso de la plataforma educativa CREA2, elaborar herramientas de monitoreo de uso de las plataformas educativas a distintos niveles de análisis (clase, grado, escuela, departamento, etc) y a distintas ventanas temporales con sus correspondientes indicadores y, finalmente desarrollar una herramienta de uso sencillo para comunicar, evaluar y gestionar los resultados. Dicha herramienta consiste en una aplicación web desarrollada en shiny y rmarkdown, buscando que los resultados, tales como visualización estadística, medidas de resumen, indicadores de monitoreo y predicciones se generen en tiempo real y de forma sistematizada. Por motivos de

confidencialidad, no podemos compartir el enlace de nuestra aplicación shiny aunque se mostrarán algunos de los resultados en la presentación con los cuidados correspondientes.

El uso de la plataforma CREA2 se describe mediante un índice de engagement (Marconi, Goyeneche, y Cobo 2017) el mismo es extendido y calculado para distintas unidades de análisis y periodos de tiempo. En base a este indicador y las 6 variables que lo componen, se analiza el nivel de uso de plataformas educativas en los 19 departamentos del Uruguay, de más de 1000 escuelas urbanas, sus grados de educación primaria (4to, 5to, 6to) y el comportamiento a nivel de cada clase. Dado el objetivo del trabajo, es de especial interés utilizar 3 unidades de análisis (además del alumno): escuela, grados y clase. Para cada uno de ellos debe analizarse la distribución del índice lo cual se realiza mediante la mediana de engagement para cada unidad de análisis. Con la posibilidad interactiva de replicar los cálculos para cualquier cuantil de interés. Es decir, calculado el engagement para cada alumno, se calcula la mediana de engagement para la escuela de interés, luego se replica el calculado para cada grado de la misma y finalmente para cada clase. De esta forma es posible hacer una comparación macro a nivel de departamentos y, también cotejar a nivel micro la evolución de las escuelas de un mismo departamento, región geográfica, barrio de interés y aún más interesante escuelas similares en base a características sociodemográficas pero de diferentes regiones geográficas. Finalmente podemos observar si las variables que componen el índice muestran un comportamiento homogéneo o heterogéneo. Estas indagaciones se realizan actualmente tanto espacialmente como temporalmente. O sea, tenemos la posibilidad de comparar una escuela versus el promedio de escuelas de dicho departamento para el año 2015 y realizar un análisis de corte transversal, así como un parangón para todos los meses del año de la misma unidad de análisis, realizando un examen de datos de panel. De forma tal, que el encargado de analizar el funcionamiento de una escuela de interés en cuanto al uso de la plataforma CREA2, tenga una visión lo más detallada y desagregada posible.

Nuestros resultados preliminares muestran un comportamiento desigual entre las escuelas de Montevideo, con relativamente bajo engagement, y del interior (departamentos diferentes de Montevideo) con alto engagement. Existe una clara concentración en los límites del rango del índice, una explicación para la conglomeración en torno a cero es la existencia de aproximadamente 42 mil estudiante de todo el marco de alumnos de 2015 que no utilizan en ningún momento CREA2.

Adicionalmente, existe variabilidad intra establecimientos. Esto es, diferencias entre los grados de una escuela. Hay centros con grados con engagement elevados y otros grados con valores cercanos a cero, generamos un ranking a nivel de región (Montevideo o Interior) y de departamento, en cual rankeamos a la escuela y la comparamos para cada grado, observando comportamiento dísimiles. Dichos casos son de especial interés puesto que incitan a pensar en la fuerza del efecto maestro. Es decir, que debería existir una fuerte asociación entre el valor del índice y el uso de la plataforma dependiendo si el maestro la utiliza o incita a usarla sea en clases o no.

Referencias

- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, y Richard Iannone. 2018. rmarkdown: Dynamic Documents for R. https: //rmarkdown.rstudio.com.
- Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, y Jonathan McPherson. 2018. shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny.
- Marconi, C., J. Goyeneche, y C. Cobo. 2017. «When Teachers and Machines Achieve the Best Combination: A National Comparative Study of Face-to-face and Blended Teaching and Learning». DATA ANALYTICS: The Sixth International Conference on Data Analytics.
- Sievert, Carson. 2018. plotly for R. https://plotly-r.com.
- Wickham, Hadley. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. https://CRAN.R-project.org/package=tidyverse.

Patricio Aroca, Javier Fernández and Esteban López

Eficiencia de los gobiernos locales en educación: Un enfoque desde el espacio para el caso chileno

Las municipalidades en Chile están encargadas de proveer educación pública. Sin embargo, la realidad de cada municipalidad en torno a la disponibilidad de recursos y contextos socio-económicos es altamente heterogénea. En este contexto, es necesario tener medidas que tomen en cuenta las distintas realidades internas y externas de las municipalidades en la provisión de servicios educacionales. Tener información cuantitativa sobre el desempeño relativo de las municipalidades que consideren sus diferencias internas, y además de cuáles son los factores externos que influyen en estos niveles de desempeño es importante para el contexto actual en donde se está debatiendo cual es la mejor forma de entregar educación y cuál será el rol de la educación pública.

Este trabajo sigue una larga literatura que establece el concepto de la eficiencia (entendida como la capacidad relativa de generar un resultado en base a insumos disponibles) como la forma de medir el desempeño. En el caso de la medición de la eficiencia de gobiernos locales, la literatura ha estudiado las variables internas y de contexto que afectan a la eficiencia en la provisión general de servicios públicos por parte de las municipalidades. Es así que variables sociodemográficas, económicas, políticas y geográficas han mostrado ser importantes a la hora de definir el nivel de eficiencia de los gobiernos locales. Sin embargo, existe poca literatura que aborde el problema de la provisión de un servicio en particular por parte de la municipalidad, y aún menos, que aborde el problema de la provisión de educación pública.

Los objetivos principales de este trabajo son dos. En primer lugar, proveer un indicador de eficiencia relativo de la provisión de educación pública por parte de las municipalidades Chilenas (Primera Etapa). Segundo, mostrar que variables de contexto sociodemográficas, educativas y espaciales tienen efectos en la eficiencia, con especial énfasis en los problemas de conectividad que sufren las comunas (Segunda Etapa). Utilizando datos de gastos municipales en educación (SINIM), como información de resultados e información de los puntajes SIMCE de estudiantes de establecimientos municipales de las comunas, se calculó un indicador de eficiencia mediante la metodología "análisis envolvente de datos". La principal ventaja de este método es que minimiza la cantidad de supuestos sobre la función de producción educativa de las municipalidades. Para la segunda etapa, mediante un modelo de regresión lineal y utilizando datos de variadas fuentes que entregan información comunal (SINIM, RECH, Evaluación Docente, Resultados SIMCE y Google Maps), se investigó la importancia estadística de que variables de contexto sociodemográficas, educativas y espaciales pueden tener sobre la eficiencia de las municipalidades en la provisión de educación. Como robustez a los resultados, se utiliza la metodología doble bootstrap propuesta por Simar y Wilson (2007), que es usualmente utilizada en la literatura de eficiencia.

Los resultados de este trabajo sugieren que existe una gran heterogeneidad en los niveles de eficiencia entre las municipalidades de Chile: A medida que nos alejamos de la región metropolitana los resultados en eficiencia son en promedio más bajos. Con respecto a las variables contextuales, existe un fuerte componente espacial-territorial que explica la heterogeneidad en los niveles de eficiencia. En particular, la conectividad muestra ser una variable relevante y que puede explicar más del 30 % de la eficiencia de las municipalidades en algunos casos. Además, ciudades de más de 100 mil habitantes son 26.7% más eficien-

tes en promedio con respecto a ciudades de menos de 5 mil habitantes. Estos resultados son importantes en el desarrollo de la nueva reforma educacional (Ley N 21040, 2017), que separa la administración de los establecimientos educacionales públicos de las municipalidades. Este trabajo permite hacer una evaluación más comprensiva con el contexto municipal en el sistema anterior y permite realizar una planificación que tome en consideración estos factores en la conformación de los entes administradores en el nuevo sistema de educación pública.

Gabriela Sandoval

Reglas de Asociación, una aplicación en retail del mercado outdoor

Una empresa comercial del área outdoor cuenta con información de sus ventas diarias, cada transacción indica la cantidad de productos comprados, su código, categoría, subcategoría, marca y precio. Además, la empresa oferta la posibilidad de ser cliente "preferencial", lo cual funciona como una subscripción que le da acceso a descuentos y/o beneficios.

Reglas de Asociación es un método de análisis no supervisado que indentifica reglas conjuntivas con alta probabilidad, en otras palabras, se buscan patrones de compra frecuentes (compras cruzadas), lo cual lo convierte en una técnica adecuada para el análisis de bases de datos comerciales. Por lo tanto, en el contexto de la planificación de marketing y el uso de técnicas de análisis no supervisado, el objetivo de este trabajo es extraer información de interés para la empresa, a través del uso de Reglas de Asociación.

El paquete arules de R permite construir un conjunto de datos con formato de transacciones e implementa el algoritmo apriori propuesto por (Agrawal et al.,1995), que realiza un computo factible incluso para bases de datos grandes. Una vez procesados los datos se identificaron 114.830 transacciones, que involucran alrededor de 14000 productos divididos en 27 categorías y 161 subcategorias.

Las reglas de asociación identificadas se discutieron con las personas responsables de la planificación de marketing, para enfocar los resultados en sus productos/subcategorías/marcas de interés; algunas de las reglas confirmaron el conocimiento empírico y otras permitieron identificar oportunidades y debilidades de los procesos de venta.

Comentarios

La presentación de este trabajo exige mantener el sigilo de la información.

Carlos A. Catania, Jorge Guerra

Visualizing Discriminative Power of Symbol-based Network Traffic Models

Motivation and Goal

In the context of network security, a behavioral model aims at capturing the long term characteristics of the network traffic following an anonymous and efficient process. Such behavioral models conform the basis of several machine learning algorithms for detecting malicious traffic. In particular, the Stratosphere IPS (http://www.stratosphereips.org) behavioral model represents the history of network connections by using a set of 50 symbols. Each time a new network flow is observed, the corresponding symbol is concatenated according the values of three features (i.e. size, duration and periodicity). An example of a behavioral model is shown in Fig. 1.

Before developing a detection method for a particular malicious behavior, it is useful to have at least an intuition of the potential discriminative power of the symbol-based behavioral model. With that goal in mind, we developed STF-PATTERN-VIZ, an open source application on top of the Shiny R package. The application consists of a set of interactive visualization components for analyzing the discriminative power of symbol-based patterns present in Stratosphere IPS behavioral models (see Fig. 2).

Overview

The idea behind STF-PATTERN-VIZ consists of applying a basic n-gram analysis on the symbol-based behavioral model. Given a set of labeled behavioral models, the application provides a visualization considering the most frequent n-grams of a given size. The visual components offer information about different aspects of the selected n-grams. The aspects considered are basically:

1) The total number of times a given n-gram is observed in all the behavioral models.

2) The number of times a given n-gram is observed in different behavioral models and

3) The labels of the models where the n-gram was observed.

Such information is mainly provided by a simple visualization tool in the following way:

- Each n-gram is represented by a rectangle. Right now the app shows only the first 300 n-grams
- A subdivision inside the rectangle indicates the same n-gram was seen in different behavioral models. The more subdivisions, the more observations of the n-gram pattern in different behavioral models.

By default, Normal behaviors are represented in Blue while Malicious in Orange. By using this simple strategy the user can have an idea about the discriminative power of n-grams in a given dataset. Clearly, a rectangle painted with only one color indicates that a particular pattern (of size n) was only observed in behavioral models with the same label. On the other side, a rectangle painted with different colors indicates class overlapping for that particular pattern. Notice that color differentiation is not restricted to only Normal or Malicious. If the user wants to use the color scheme for discriminating between, for instance, DNS traffic DGA and Normal, she can do it easily using regular expressions.

Fig 1 . A behavioral model representing all the connection based on UDP to port 53.

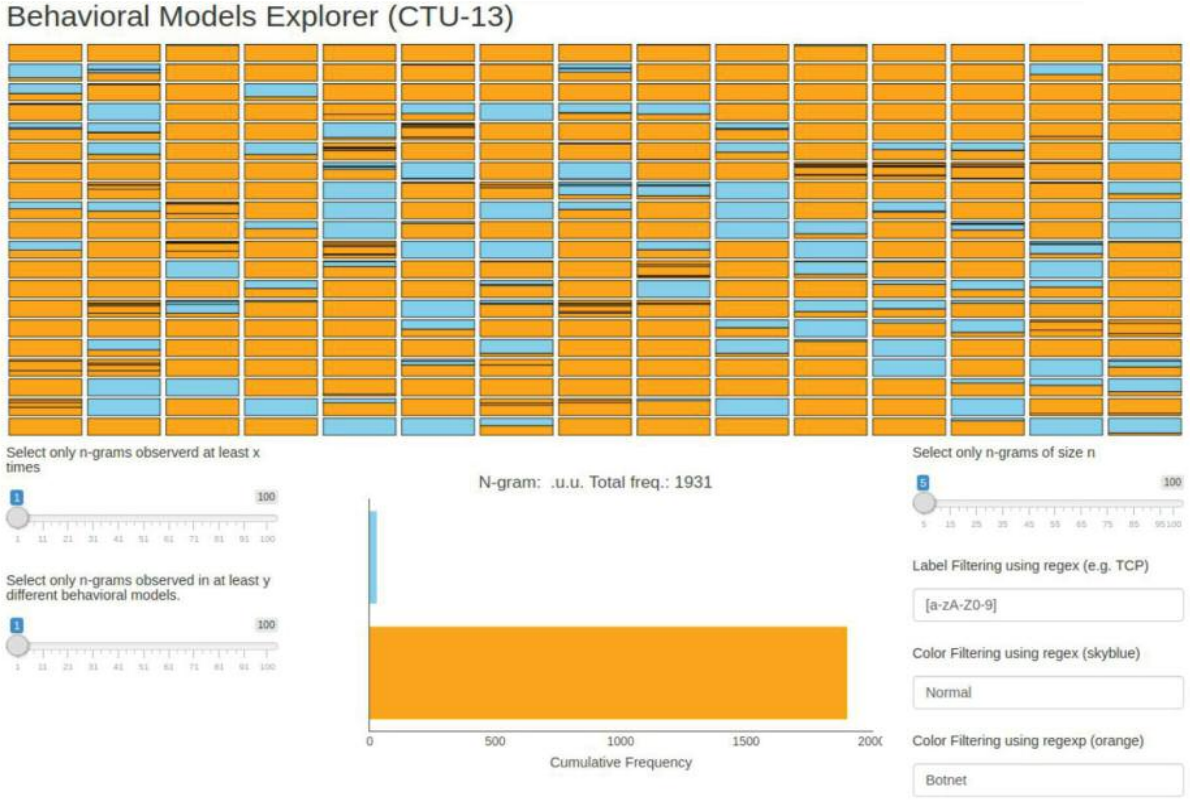


Fig 2. A detailed view of the STF-PATTERN-VIZ interface. In this case, the first 300 n-grams with n = 5 are shown.

Detailed View

It is possible to access to detailed information about a particular n-gram by clicking the corresponding rectangle. In the bottom of the screen, a Frequency histogram for the selected n-gram is shown. The histogram you will find information about the different behavioral models where that n-gram pattern was observed.

Filtering n-grams

The application support some basic pattern filtering.

- Filter n-grams of size n
- Filter n-grams observed at least x times
- Filter n-grams observed at least in y different behavioral models.

By default the application will consider all the behavioral models present in the dataset. However, it is possible to filter by using the Label information. Additionally, it is possible to filter by Label. Since the standard STF Label contains information regarding the different protocol layers, it then possible to filter by protocol layer 4 or 5. Such filtering is done by using standard regular expressions.

Live Demo

A live demo is available at (https://harpomaxx.shinyapps.io/stf-pattern-viz/). The demo contains a subset of the CTU-13 Dataset. Some minor modifications were made in the label description for facilitating the aggregation. A detail about the labeled modification and how the CTU-13 subset was generated is found at http://rpubs.com/harpomaxx/ctu13bis

Conclusions and Future Work

Even thought Stf-pattern-viz was developed as a tool for the Stratosphere model behaviors, it seems clear it can easily be adapted to a pattern analysis of any other n-gram oriented problem, such as text-analysis or DNA sequences, among others. Given the aforementioned extensible capacity to sequence analysis and the visual power of this solution, in the near future we plan to create an R package that will facilitate the application of the tool to new n-gram oriented problems.

Claudia Cifuentes Donald

Herramientas de Análisis en Educación Superior

Este trabajo está basado en la experiencia de la instrucción de educación superior más grande en Chile, que cuenta con más de 122.000 estudiantes, tiene 26 sedes a lo largo del país donde ofrece programas de estudio en su Centro de Formación Técnica, Instituto Profesional y Universidad tanto en jornada diurna como vespertina. Esta Institución representa alrededor del 10% de la matrícula total del país. Adicionalmente, sus tres instituciones son no selectivas lo que conlleva a que, según estadísticas internas de la entidad, la mayoría de los estudiantes de primer año son la primera generación de su familia en ingresar a la educación superior.

La gran cobertura que posee la institución presenta diversos desafíos, entre ellos el trabajo con la gran cantidad de información disponible de los estudiantes, egresados, académicos y del contexto de cada una de las sedes. Además, dados los crecientes requerimientos de acreditación y las necesidades propias de las distintas sedes, áreas académicas o áreas de casa central es preciso contar con herramientas que permitan el procesamiento y el análisis de ésta para la toma de decisiones.

Esta tarea presenta diversos retos, por ejemplo, el trabajo con variadas fuentes de información como lo son los procesos institucionales de matrículas, egresos, prácticas y registros de progresión académicas o el trabajo con encuestas institucionales. Además, dada la complejidad del sistema es necesario contar con formas de entregar la información con variadas aperturas que permitan a los interesados entender con facilidad y los ayude a resolver sus interrogantes de manera oportuna, lo que significa el procesamiento rápido de grandes cantidades de información.

En la dirección de análisis institucional en específico se utiliza R con distintas finalidades. Entre ellas se utiliza para la construcción de modelos, generación de reportes, cálculo de muestras para encuestas institucionales, análisis de texto, entre otros. Para ello se han utilizado una gran cantidad de paquetes, entre ellos ggplot2, dplyr, RODBC, rmarkdown, forcasts, tm, worldcloud, caret, knitr, randomForest, etc.

Dado que la información institucional no se puede compartir y que los datos de los estudiantes son confidenciales, a continuación, se observan ejemplos de algunos gráficos hechos con grupos con información que no es representativa de los resultados institucionales.

Jesús Eduardo Gamboa Unsihuay

Análisis de comentarios de la página de Facebook del Centro de Admisión y Promoción de la Universidad Nacional Agraria La Molina

En la actualidad, las instituciones deben establecer planes de marketing que vayan más allá de la perspectiva tradicional, involucrando el enfoque digital, pues debido al avance de la tecnología, cada vez más usuarios pueden acceder a internet en diversos medios, especialmente smartphones. Un adecuado plan de marketing digital es la carta de presentación permanente (las 24 horas, los 7 días de la semana) de una organización hacia el segmento de mercado al que va dirigida. Este plan le permite establecer canales efectivos de comunicación y responder a las necesidades de sus usuarios o clientes, así como marcar presencia digital permanente en las redes.

Por otro lado, a diferencia de diversos países de Latinoamérica en los que el proceso de admisión a las universidades es unificado, en Perú es establecido por cada institución superior de manera independiente y de acuerdo con los lineamientos de la Ley Universitaria 30220. En ese sentido, el Centro de Admisión y Promoción (CAP) de carreras universitarias de la Universidad Nacional Agraria La Molina (UNALM) cuenta con distintas funciones, una de ellas es la de mantener contacto con un segmento de mercado juvenil (para convencerlos de estudiar una carrera en la universidad) y adulto (son los padres o apoderados quienes generalmente tienen la decisión de pagar la inscripción en el proceso de admisión universitario). No obstante, el plan de marketing digital del CAP – UNALM precisa reajustarse a las necesidades y deseos de este público objetivo.

En este contexto y conscientes de la responsabilidad ambiental, el proceso de inscripción se caracteriza por denominarse "cero papeles" ya que todos los procedimientos se llevan a cabo vía web, sin necesidad de dirigirse hasta el campus universitario a realizar trámites presenciales, sino que éstos se realizan solo luego de haber alcanzado una vacante en el examen de admisión. Esta característica resalta la importancia de establecer un plan de marketing digital en vista de que se da mayor preferencia al contacto digital que al físico, mediante el uso de redes sociales (Facebook, Whatsapp, Instagram) y la página web.

La página oficial en Facebook del CAP-UNALM sirve como medio de contacto con los postulantes y como publicidad para los potenciales aspirantes a la universidad. Se ha recolectado datos acerca de las publicaciones realizadas durante los seis procesos de admisión que se llevaron a cabo desde el 2017-I hasta el 2019-II, es decir en el periodo que va desde agosto del 2016 hasta agosto del 2019, así como los comentarios que las personas han manifestado en cada una de dichas publicaciones.

El conjunto de textos recabado permite realizar el análisis de comentarios en cuanto a su contenido y sentimiento, con la finalidad de determinar los temas de discusión y la posición de los usuarios respecto a ello (positivo, neutral, negativo). Los temas de discusión son determinados mediante la técnica conocida como Topic Modeling – Asignación Latente de Dirichlet, mientras que las posturas respecto al contenido, mediante el análisis de sentimiento. Para el estudio de datos textuales se utilizan distintos paquetes del software R, entre los que se tiene tm, topicmodels, syuzhet y la familia tidyverse.

Mediante Topic Modeling se han determinado principalmente seis temas de discusión: "invitaciones", "etiquetas", "consultas", "elogios a la universidad", "talleres de carreras" y "preparación para postular". Sin embargo, la importancia de cada tema no es la misma a través de los distintos procesos de admisión, pues en los dos últimos se

incrementó la invitación entre usuarios para asistir a talleres y/o postular, así como las etiquetas (invitación tácita). Por el contrario, las consultas por comentarios de Facebook disminuyeron, esto debido al reforzamiento de canales de comunicación alternativos.

Respecto al análisis de sentimientos, las publicaciones y comentarios con sentimiento positivo superaron el 50% del total a partir del 2017-II, sin embargo, dicho porcentaje presenta tendencia decreciente debido al incremento de texto de valoración neutral. Los tópicos mejor valorados son: talleres de carreras, preparación para postular y elogios a la universidad. En contraste, los textos relacionados a consultas presentan una valuación negativa considerablemente alta (20%).

Esta evaluación de la interacción textual vía los comentarios en la página de Facebook servirá como base para el establecimiento de un plan de marketing digital con miras a futuros procesos de admisión, en el que se fortalezcan los contenidos que presentan valoración positiva y se evalúen los inconvenientes relacionados a las consultas de los usuarios, dada su apreciación negativa.

Palabras clave: proceso de admisión, educación superior, minería de texto, análisis de contenido, análisis de sentimientos

Explorando o ENADE com o R na Visão de Gestor de Curso

O desenvolvimento de modelos de mensuração da qualidade de cursos superiores é assunto bastante discutido na literatura internacional dada a diversidade de dimensões e de interesses na avaliação. No Brasil uma das métricas de mensuração da qualidade do ensino superior e das instituições é o ENADE, com provas realizadas em ciclos trienais de avaliação. Este trabalho tem como principal objetivo investigar os dados de avaliação de 2015 disponibilizados pelo INEP e verificar dentre a composição do Conceito Preliminar de Cursos (CPC), as variáveis em que o gestor deverá concentrar seus esforços. Com este intuito foram utilizadas três técnicas estatísticas: Análise Exploratória de Dados, Análise do Componente Principal e Análise Fatorial Exploratória. Os resultados indicam que a Organização Didático-pedagógica, Infraestrutura e Instalações e Oportunidades de Ampliação da Formação são as variáveis de maior influência na composição do CPC.

Introdução

A prática de ranqueamento de cursos e instituições, e sua apropriação principalmente nos mecanismos de marketing de instituições de ensino particulares é comum em todo o mundo. No Brasil, o ranqueamento é realizado com base na pontuação no ENADE, Exame Nacional de Desempenho dos Estudantes, cujo objetivo é avaliar o rendimento dos alunos dos cursos de graduação, ingressantes e concluintes, em relação aos conteúdos programáticos dos cursos em que estão matriculados. Seu principal instrumento é o CPC, Conceito Preliminar de Curso, um indicador de qualidade que avalia os cursos de graduação. Seu cálculo e divulgação ocorrem no ano seguinte ao da realização do ENADE, com base na avaliação de desempenho de estudantes, no valor agregado pelo processo formativo e em insumos referentes às condições de oferta – corpo docente, infraestrutura e recursos didático-pedagógicos (INEP, 2017). O CPC é composto por 37 variáveis. Na edição de 2015 foram avaliados 8121 cursos de 1758 instituições de ensino espalhados em 753 municípios de 27 estados brasileiros, com um percentual de participação de 81.4% dos estudantes nas provas. Há que se investigar a relação entre a diversidade destas variáveis com o vetor qualidade. Para isso é necessário que se racionalize o universo de variáveis em um conjunto delimitado de dimensões. Diante do exposto, este trabalho é resultado da analise das variáveis que constituem o ENADE por meio da redução das dimensionalidades. Acredita-se que seja possível extrair perfis de variáveis representativos dos cursos superiores capazes de descrevêlos em arranjos significativos e de fácil compreensão.

Método

Esta pesquisa utilizou um conjunto de técnicas estatísticas distintas para a análise de dados: Análise do Componente Principal (ACP) e Análise Fatorial(AF). Utilizaram-se os dados do senso de 2015 disponíveis site do INEP. Para leitura utilizou-se o pacote readxl (WICKHAM e BRYAN,2019), manipulação dos dados foi utilizado R, e o pacote dplyr (WICKHAM et al., 2017) e os gráficos realizados com o pacote ggplot2 (WICKHAM,2016). Na análise exploratória dos dados foram analisados: o número de cursos por regiões, o percentual de cursos por organização acadêmica e categoria administrativa. Uma inspeção na variável resposta, CPC, permitiu identificar que 12,97% dos cursos não obtiveram notas. Embora para variáveis preditoras as informações estivessem completas.

Foi realizada a imputação dos dados utilizando o método K-NN com o pacote DMwR v0.4. (TORGO, 2013). A padronização dos dados deve anteceder a ACP (LAROSE e LAROSE, 2015) e foi realizada com a utilização da função scales contida na instalação básica do R. Tanto a ACP como a AF fizeram amplo uso do pacote psych (REVELLE, 2017). Dada a sua importância, este pacote constitui elemento chave para a aplicação das metodologias estatísticas desta pesquisa.

Após a análise do correlograma, prejudicada pelo grande número de variáveis preditoras, partiu-se para a ACP, o critério do Scree Plot (HAIR et al., 1998) indicaria o uso de 5 fatores correspondendo a 78,7% da variabilidade. Foi ainda adicionada uma camada de robustez a análise com a AF, que requer que as variáveis estejam minimamente correlacionadas. Neste sentido, foi realizado teste de Esfericidade de Bartlett que indicou a adequação da AF.

Discussão e conclusão

Após a aplicação da ACP e AF foi possível identificar os agrupamentos das variáveis. Verificou-se que o primeiro componente tem afinidade com a estrutura pedagógica do curso (Organização Didático-pedagógica), sua estrutura (Infraestrutura e Instalações) e perspectiva de futuro do ponto de vista do aluno (Oportunidades de Ampliação da Formação). Concluímos que esse componente apresenta um perfil do que o curso é capaz de entregar ao aluno, que é refletida nas notas que o mesmo atribui no momento da avaliação. O segundo componente observamos uma concentração dos pesos nas variáveis associadas a duas dimensões: alunos e professores. No que se refere aos alunos, pesam mais as notas avaliadas de acordo com o desempenho dos mesmos nas provas nas áreas de Formação Geral e Conhecimentos Específicos. Já em relação aos professores tem destaque as variáveis Quantidade de Mestres e Quantidade de Doutores nos cursos. O terceiro componente dá destaque as variáveis Concluintes Inscritos e Percentual de Concluintes Participantes (CP) com nota no ENEM, Exame Nacional do Ensino Médio. Portanto indica que este componente busca determinar a situação dos alunos no censo bem como a contabilização da participação dos mesmos. O quarto componente não apresenta diferença em relação ao terceiro em termos de composição. A utilização da ACP e AF apresentaram quais variáveis são passíveis de melhor explicarem o que influencia no desempenho do curso no CPC e portanto devem ser objeto de maior preocupação por parte dos gestores os cursos e instituiçõ-es. Os gráficos, tabelas e códigos deste trabalho podem ser encontrados em https://github.com/gui-souza/ACP_AF_Aplicada.

Em áreas como Ciências Sociais por exemplo, 60% da variabilidade é considerada satisfatória dada a imprevisibilidade da natureza do comportamento humano (LAROSE; LAROSE, 2015, p. 103). Concluímos que a identificação e análise dos três primeiros componentes são relevantes e constituem contribuição importante desta pesquisa, vez que os mesmos contabilizam 60,1% de toda a variabilidade.

Bibliografia

- HAIR, J. F. et al. Multivariate data analysis. [s.l.] Prentice hall Upper Saddle River, NJ, 1998. v. 5
- INEP. Avaliação in Loco - Glossário dos Instrumentos de Avaliação Externa. jul. 2018.
- LAROSE, D. T.; LAROSE, C. D. Data mining and predictive analytics. [s.l.] John Wiley & Sons, 2015.
- REVELLE, W. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois: Northwestern University, 2017.
- TORGO, L. Data Mining with R, learning with case studies. [s.l.] Chapman and Hall/CRC, 2010.
- WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. [s.l: s.n.].
- WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York,2016. ISBN 978-3-319-24277-4. Disponível em: http://ggplot2.orgi.
- WICKHAM, H.; BRYAN, J. readxl: Read Excel Files. [S.l.], 2019. R package version 1.3.1. Disponível em: https://CRAN.R-project.org/package=readxli.
- WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. [S.l.], 2018. R package version 0.7.6. Disponível em: https://CRAN.R-project.org/package=dplyri..

David Morales & Stephanie Orellana

Metodología para la estimación espacio-temporal de la demanda de riego (Evapotranspiración real) mediante algoritmos de reproyección y paralelización web

Evapotranspiración

La evapotranspiración corresponde al flujo de vapor de agua desde la superficie del suelo (evaporación) y la transpiración de las plantas hacia la atmósfera (transpiración). Los flujos de vapor desde la superficie son un elemento esencial en el balance hídrico, ya que representan uno de los mecanismos de conexión más importantes entre los procesos de la superficie terrestre con la dinámica atmosférica (Liu et al. 2016). A pesar de su importancia, desde el punto de vista metodológico, es una variable muy difícil de medir a escala regional (Rodell et al. 2004; Howell 1990).

El proceso mediante el cual tiene lugar la evapotranspiración, es el resultado de interacciones complejas entre fenómenos físicos, determinados principalmente por factores climáticos, composición y estado del suelo y condiciones de la cobertura vegetal presente. De manera de poder estudiar el fenómeno de forma más precisa, en la literatura se han propuesto definiciones y modelos conceptuales que buscan separar el efecto de las distintas componentes forzantes de la evapotranspiración, de modo de analizar los elementos que actúan como limitante del proceso. De esta forma, se define el concepto de evapotranspiración potencial (ETP) como la máxima tasa de evaporación en condiciones óptimas de la vegetación y sin restricción de agua. Una de las ecuaciones fundamentales para el cálculo de ETP es la de Penman-Monteith (Shuttleworth, 1993), caracterizada por la ecuación:

$$\lambda_v ETP = \frac{\Delta(R_n - G) + \rho_a c_p (e_s - e_a)/r_a}{\Delta + \gamma}$$

Otro concepto de gran utilidad es el de evapotranspiración de referencia (ETo). Este se introdujo para estudiar la demanda evapotranspirativa de la atmósfera, independientemente del tipo y desarrollo de los cultivos y de las prácticas de manejo (Allen et al., 2005). Este concepto permite comparar valores medidos o estimados en diferentes localidades o en diferentes épocas del año, debido a que no depende del tipo de superficie analizada, siendo afectada únicamente por parámetros climáticos. La ecuación utilizada para calcular este parámetro es la ecuación FAO Penman-Monteith (Allen et al., 2005), según la ecuación: –Error! No se encuentra el origen de la referencia.

$$ETo = \frac{0{,}408\,\Delta\,(R_n - G) + \gamma \frac{900}{T + 273} u_2 (e_s - e_a)}{\Delta + (1 + 0{,}34\,u_2)}$$

Tanto ET_P como ET_o son definiciones conceptuales para casos idealizados, fundamentalmente dependientes de las condiciones atmosféricas y de balance de energía en el sitio. Normalmente la tasa de evapotranspiración real por parte de la cobertura está por debajo de estos valores, debido a que las condiciones de la cobertura vegetal, tipo y estado del suelo y disponibilidad hídrica muchas veces difieren de manera importante a las consideradas por las ecuaciones precedentes.

Sin embargo, estos modelos conceptuales ofrecen la posibilidad de generar set de datos independientes útiles para evaluar el desempeño de modelos hidrológicos como los considerados en este estudio de balance hídrico.

Metodología de cálculo para el producto de validación

Para validar los resultados del modelo hidrológico VIC respecto a las salidas del sistema por evapotranspiración, se desarrolló y aplicó una metodología basada en el balance de energía de la

Capas de información	Fuente	Resolución temporal	Disponibilidad temporal	Formato y resolución espacial
Información de superficie				
Temperatura máxima del aire	CR2MET V1.3	Diaria	1979 - 2015	Raster, 5km
Temperatura mínima del aire	CR2MET V1.3	Diaria	1979 - 2015	Raster, 5km
Radiación Global Horizontal	Explorador Solar (Depto. Geofísica U. Chile, Min. Energía)	Diaria	2004 - 2016	Raster, 1km
Información Satelital				
Albedo de superficie	MODIS (MCD19A3)	Cada 8 días	2001 - presente	HDF, 1km
Temperatura de superficie	MODIS (MYD11A1)	Diaria (día y noche)	2001 - presente	HDF, 1km
Clasificación de Cobertura de Suelo	Zhao et al., 2016	Año 2014, a temporal	2014	Raster, 30 m

Tabla 1. Datos de entrada utilizados en la metodología de cálculo de ETr

superficie, calculada a partir de información climática de superficie, complementada con información proveniente de plataformas satelitales.

Los de insumo para la metodología, se presentan en la Tabla 1 a continuación.

Como primer paso de la metodología de cálculo, a partir de la ecuación general de radiación neta, se considera el cálculo de la radiación neta disponible en superficie, según la siguiente ecuación:

$$Rn_{i,j,k} = (1 - \alpha_{MOD\ i,j,k}) * Rg_{i,j,k} + 0.0864 * \varepsilon_{A\ i,j,k} * \sigma * (\overline{T}_{i,j,k} + 273)^4 - 0.0864 * \varepsilon_{C\ i,j,k} * \sigma * (\overline{T}_{si,j,k} + 273)^4$$

Donde i, j, k corresponden al pixel i, día del año (dia juiano j) y año k. De esta forma $Rn_{i,j,k}$ corresponde a la radiación neta disponible en el pixel i, día j y año k, $\alpha_{MOD\ i,j,k}$ al albedo de la superficie desde MODIS a nivel diario, $Rg_{i,j,k}$ la radiación global horizontal diaria obtenida desde el explorador solar, $\varepsilon_{A\ i,j,k}$ y $\varepsilon_{C\ i,j,k}$ la emisividad del aire y superficie respectivamente a escala diaria, $T_{i,j,k}$ la temperatura promedio del aire diaria, $T_{si,j,k}$ la temperatura promedio de la superficie a nivel diario y σ la constante de Stefan-Boltzmann ($5{,}67 * 10^{-8}$).

La temperatura promedio del aire diaria está definida como:

$$\overline{T}_{i,j,k} = \frac{T_{x\ i,j,k} + T_{n\ i,j,k}}{2}$$

En donde $T_{x\ i,j,k}$ y $T_{n\ i,j,k}$ corresponden a las temperaturas del aire máximas y mínimas diarias a dos metros obtenidas del producto CR2MET.

Por su parte, la temperatura promedio de superficie está definida como:

$$\overline{T}_{s\ i,j,k} = \frac{T_{MOD-d\ i,j,k} + T_{MOD-n\ i,j,k}}{2}$$

En donde, $T_{MOD\text{-}d}$ y $T_{MOD\text{-}n}$ corresponden a las temperatura diurna (~13:50 hrs) y nocturna (~01:40 hrs) de la superficie captada por el sensor MODIS a nivel diario. Por su parte, las emisividades de la cobertura y el aire, se definen, respectivamente como:

$$\varepsilon_{C\ i,j,k} = 0{,}95$$

$$\varepsilon_{A\ i,j,k} = 1{,}72 * \left(\frac{e_{A\ i,j,k}}{10 * (273 + \overline{T}_{i,j,k})}\right)^{1/7}$$

En donde $e_{A\ i,j,k}$ es la presión de vapor a saturación del aire a nivel diario en el pixel i, día j y año k, el cual se define como:

$$e_{A\ i,j,k} = 6{,}11 * e^{\left(\frac{T_{n\ i,j,k} * 17{,}27}{237{,}3 + T_{n\ i,j,k}}\right)}$$

A partir de la obtención de la radiación neta disponible en cada pixel, es necesario determinar la fracción de la energía utilizada en la generación de los flujos de calor sensible y latente. Este último será finalmente la evapotranspiración de la cobertura a considerar como elemento de validación del componente de validación obtenido desde el modelo VIC.

Esta metodología considera el análisis de las diferencias de temperatura entre la superficie y el aire circundante siguiendo lo propuesto por Jackson et al. (1981) y el trabajo seminal de Idso et al. (1981) . En términos simples, a mayor diferencia de temperatura, la energía disponible en la superficie se traduce en mayores flujos de calor sensible (calor), mientras que menores diferencias entre la temperatura de la superficie y el aire circundante indican mayor mayores flujos de calor latente en forma de evapotranspiración entre la superficie y la atmósfera.

En términos operativos, un primer paso consiste en identificar, para cada uso de suelo y por cada una de las cuencas, el píxel cuyo valor promedio de temperatura de superficie sea el máximo a nivel diario, a partir de la serie de temperatura superficial obtenida desde el sensor MODIS, tal como se presenta a continuación:

$$\overline{T}_{sx\ j,k} = \max(\overline{T}_{s\ i,j,k})$$

Debido a las diferencias entre las características y propiedades de las distintas coberturas existentes en las cuencas analizadas, una primera consideración, fue segmentar el análisis diferenciando las distintas coberturas de suelo presentes. Por lo anterior, el valor de $T_{sx\ j,k,l}$ considera el término l, indicando que esa selección se realiza para los "l" tipos de coberturas presentes a partir del producto de cobertura de superficie.

Luego, a partir de las variables previamente definidas, y la obtención de la radiación neta para cada pixel, la resistencia aerodinámica de la cobertura es posible de obtener según la siguiente expresión:

$$ra_{\ i,j,k,l} = \frac{1200 * (\overline{T}_{sx\ j,k,l} - \overline{T}_{i,j,k,l})}{Rn_{i,j,k}}$$

Finalmente, la evapotranspiración real desde la superficie, será obtenida utilizado la siguiente expresión:

$$ET_{i,j,k,l} = \frac{Rn_{i,j,k} - \frac{1200 * (\overline{T}_{s\ i,j,k,l} - \overline{T}_{i,j,k,l})}{ra_{\ i,j,k,l}}}{2{,}5}$$

En donde ETi,j,k,l es la evapotranspiración real a nivel diario para todo el dominio considerado. Para efectos del este proceso de cálculo, se considera j como variable entre valores 1 y 365 (día juliano) y k definido entre los años 2004 y 2015, definidos a partir de la disponibilidad de datos de radiación y producto grillado de temperatura, respectivamente.

El procedimiento fue realizado mediante un script escrito en lenguaje R a través de la interfaz RStudio con la ayuda de las librerias "ncdf4", "phylin", "data.table", "rgdal", "gdalUtils", "raster", "akima" y "palallel" para realizar el procedimiento de paralelización.

Resultados

El algoritmo desarrollado, permite capturar la variabilidad espacial de la señal de evapotranspiración a nivel de cuenca, tal como se presenta en la Figura 1.

En lo que respecta a la variabilidad temporal, el proceso de cálculo responde a la temporalidad esperada para las variaciones de la evapotranspiración en función de la estacionalidad de las temperaturas, así como a la disponibilidad de agua, de manera particular, en zonas con cobertura vegetal importante.

Una primera validación del producto generado se implementó contrastando los valores obtenidos de Evapotranspiración real calculada (ETr) con la Evapotranspiración Potencial (ETo) reportada por la estación agrometeorológica de la red Agroclima para la localidad de Pirque[1] (Figura 2).

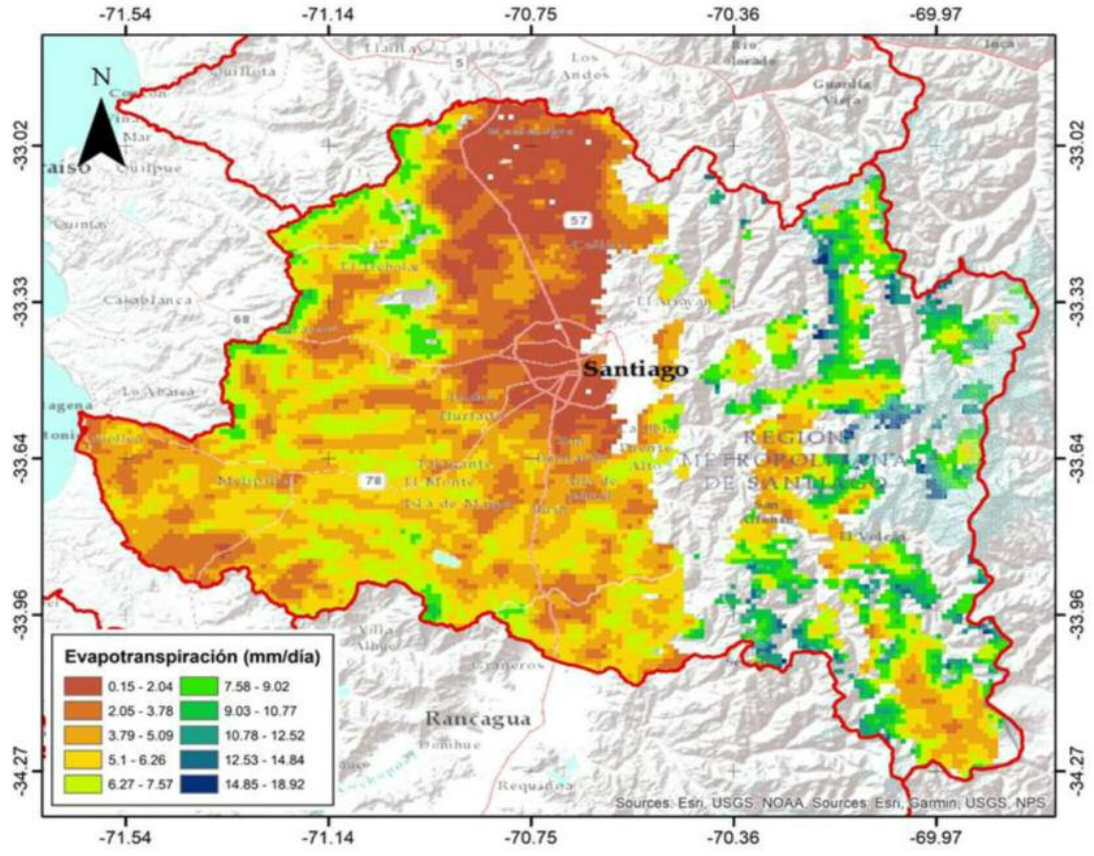


Figura 1. Evapotranspiración real estimada para en la cuenca del río Maipo para el día 25 de enero de 2008.

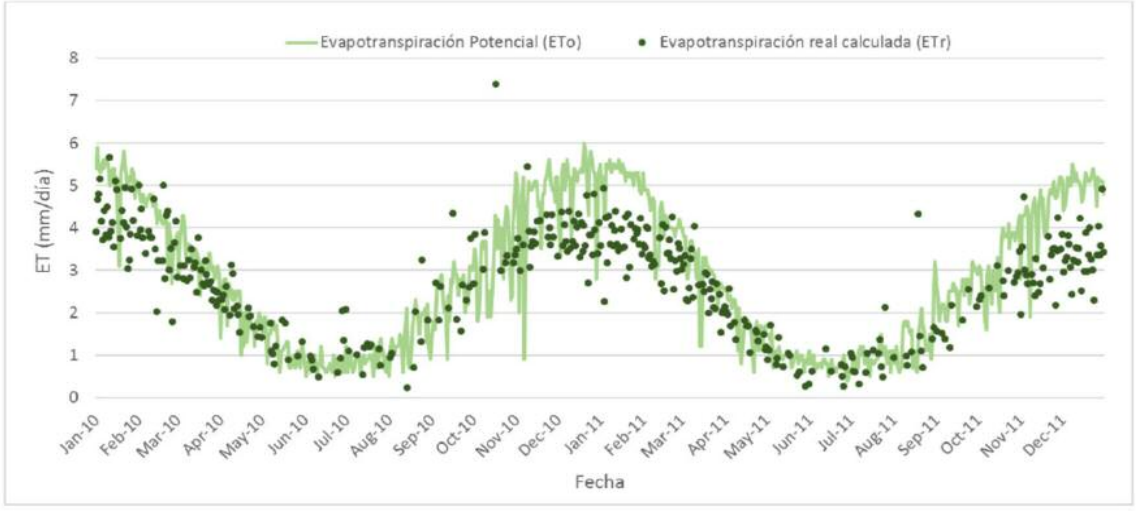


Figura 2. Series de evapotranspiración de referencia calculada por la estación agrometeorológica en la estación Pirque de la red Agroclima (ETo) y Evapotranspiracíón real calculada (ETr) en el pixel correspondiente a la estación para los años 2010 y 2011.

Referencias

- Allen, R. G., Pereira, L. S., Smith, M., Raes, D., & Wright, J. L. (2005). FAO-56 dual crop coefficient method for estimating evaporation from soil and application extensions. Journal of irrigation and drainage engineering, 131(1), 2-13.
- Chao, Y.; Feng, D., Yua, L., Wang, X., Chen, Y., Hernández, H.J., Galleguillos, M., Estades, C., Biging, G., Radke, J. & Gong, P. 2016. Detailed dynamic land cover mapping of Chile: accuracy improvement by integrating multi-seasonal land cover data. Remote Sensing of Environment 183, 170–185.
- R.D. Jackson, S.B. Idso, R.J. Reginato, P.J. Pinter Canopy temperature as a crop water stress indicator Water Resource Res., 7 (1981), pp. 1133-1138
- Howell, T. A. (1990). Relationships between crop production and transpiration, evapotranspiration, and irrigation. Agronomy (USA).ISO 690.
- S.B. Idso, R.D. Jackson, P.J. Pinter, J.L. Hatfield Normalizing the stress – degree – day parameter for environmental variability Agric. Meteorol., 24 (1981), pp. 45-55
- Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X. and Sang, Y.F., 2016. A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. Journal of Hydrology, 538, pp.82-95.
- Rodell, M., Famiglietti, J.S., Chen, J., Seneviratne, S.I., Viterbo, P., Holl, S. and Wilson, C.R., 2004. Basin scale estimates of evapotranspiration using GRACE and other observations. Geophysical Research Letters, 31(20).
- Shuttleworth, W. J. (1993). Handbook of hydrology. Civil Engineering.

Giorgio Boccardo & Felipe Ruiz

Desafíos de la enseñanza de R en contextos hispanoparlantes: una herramienta interactiva para el campo de las ciencias sociales

El objetivo de esta ponencia es presentar una herramienta de aprendizaje interactivo sobre estadística descriptiva y lenguaje de programación R, orientada a estudiantes e investigadores/as en ciencias sociales de contextos hispano parlantes. La enseñanza y aprendizaje de R en contextos de habla hispana presenta diferentes desafíos. En términos endógenos al lenguaje de programación, éste registra una curva de aprendizaje lenta para quienes se encuentran aclimatados/as al uso de softwares con interfaces centradas en botones, en los que no se requiere conocimientos avanzados de sintaxis. En términos exógenos al lenguaje de programación, tanto para quien enseña como para quien aprende se observa una limitada disponibilidad de recursos de aprendizaje en español. Ambas cuestiones limitan la masificación del uso del lenguaje de programación R en diferentes campos disciplinares hispanoparlantes.

De forma específica, estas barreras han desincentivado la incorporación del lenguaje R (pese a ser una funcionalidad de licencia abierta) como herramienta para la recolección, procesamiento, análisis y visualización de datos en el campo de las Ciencias Sociales, acrecentando la brecha entre la producción científica de países hispano parlantes y países desarrollados.

Con base en la experiencia académica desarrollada desde 2015 en diferentes asignaturas de estadística en la Facultad de Ciencias Sociales de la Universidad de Chile, los autores de esta ponencia desarrollamos una herramienta de aprendizaje interactivo para el uso de R, aplicada a la enseñanza de Estadística Descriptiva en Ciencias Sociales. La principal característica de este recurso de apoyo docente es que permite complementar el trabajo de enseñanza en aula, logrando también un ajuste a los ritmos de aprendizaje diferenciados de cada estudiante a lo largo de un semestre académico.

Durante 2018 desarrollamos con el paquete rmarkdown una primera edición del manual RStudio para Estadística Descriptiva en Ciencias Sociales en formato de libro digital estático (disponible para descarga libre en Researchgate y con materiales de apoyo en Github), orientado a estudiantes de los primeros años de la carrera de sociología. Sin embargo, al poco tiempo esta herramienta se difundió entre académicos/as e investigadores/as a nivel nacional y latinoamericano, evidenciando la carencia de este tipo de recursos en español (desde enero de 2018 ha sido descargado 1627 veces desde Researchgate). Esta primera sistematización permitió sintetizar una aproximación propia a la enseñanza de R basada en la experiencia docente y las necesidades de aprendizaje detectadas en diversos grupos de estudiantes. Para esta Conferencia Latinoamericana sobre el Uso de R en Investigación y Desarrollo presentamos una segunda edición corregida y aumentada, en formato de libro digital interactivo, desarrollada mediante los paquetes rmarkdown y bookdown. De este modo, se espera contribuir decididamente a la difusión del aprendizaje y enseñanza del lenguaje de programación R en el mundo hispano parlante, específicamente, en el campo de las ciencias sociales.

En específico, presentamos una herramienta que:

Primero, articula la enseñanza de R con una exposición y tratamiento rigurosos de cuestiones metodológicas y de estadística, cuyo uso mediante tal lenguaje de programación se busca enseñar. De tal forma, y recogiendo la tradición de manuales de referencia en el campo producidos para el

habla inglesa – como los de Andy Field (2012) o las más recientes publicaciones de Hadley Wickham y Garret Grolemund (2016) – quien se enfrente al aprendizaje de R encontrará en este material un recurso que busca apuntalar de forma integral la enseñanza: se apuesta por una perspectiva que no disocie el aprendizaje y uso del lenguaje de programación, de cuestiones de metodología, estadística, lógica y rigurosidad científicas en general.

Segundo, se propone desplegar una enseñanza orientada al uso cotidiano del lenguaje R que cualquier investigador o investigadora en Ciencias Sociales despliega, ya sea en el campo académico o profesional: por ello, se pone énfasis en enseñar herramientas de análisis y visualización de datos sociales a nivel descriptivo. En ese sentido, la perspectiva de enseñanza utilizada es fundamentalmente práctica y aplicada, ya que busca facilitar el aprendizaje de R a investigadores e investigadoras sociales que precisen incorporar este lenguaje de programación en su trabajo cotidiano.

Tercero, tanto las sintaxis y bases de datos utilizadas en los ejemplos se encuentran disponibles en la plataforma Github. Con esto se busca potenciar al máximo la interactividad con este recurso de aprendizaje: cualquier persona que desee replicar los ejemplos y ejercicios incluidos puede acceder a los datos y sintaxis de forma gratuita; a la vez, toda persona que desee sugerir ediciones de estilo, formato o contenidos a la versión en curso del libro, puede hacerlo mediante el sistema de control de versiones proporcionado por Git.

Finalmente, con su publicación pública y gratuita vía las plataformas ResearchGate y la página web asociada al libro digital en formato interactivo (anclada en los hosting proporcionados por bookdown y RSstudio) se espera contribuir a los principios de la Open Science que los autores comparten plenamente.

En definitiva, se trata de una herramienta interactiva de aprendizaje con amplio potencial de crecimiento en cuanto a su alcance y contenidos. El desarrollo alcanzado hasta su segunda edición es prometedor: las constantes innovaciones producidas en torno al lenguaje de programación R a escala mundial y la propia demanda generada por la comunidad científica en Chile y Latinoamerica alientan a profundizar en el tipo de contenidos, dinámicas de enseñanza y formas de presentación utilizadas, para seguir apuntalando la difusión del uso de R para el procesamiento y análisis de información en el mundo de habla hispana.

Gabriel Zanlorenssi

Use of R to work with public data in data journalism

Submission - LatinR

This submission consists in an oral presentation about the use of public data in data journalism, with R as the main tool of work.

Why R?

Based on the experience with Brazilian public datasets, it emphasizes how R, especially Tidyverse packages, makes easy to work with large amounts of data to build stories and visualizations. Without the use of R, it would be impossible to do high standard infographics in one day or two, considering the intense and busy work routine in a newsroom.

R in the workflow of data journalism team

The presentation will cover how R fits in the workflow of an internationally awarded data journalism team, also considering the integration with other tools like Adobe Illustrator, used by graphic designers.

Why public data?

As the fake news proliferate in the digital medias, the public demands reliable information, often relied on data evidences. In our experience, the data visualizations with the highest number of visualizations were based on public data. That is data that is available for everyone to access, download, work with and distribute. However, this whole process is too hard for general public, and it is why the journalism - and its use of R - is important to provide qualified and reproducible information.

Sistema de soporte de decisiones en Shiny para el balanceo de bicicletas en una red de bicicletas compartidas

Resumen

Dentro de la infraestructura de una Smart City, existen servicios y sistemas que en su conjunto permiten la generación de entornos inteligentes, como las plataformas digitales y la gestión de diversos servicios de transporte. Uno de los sistemas de transporte que ha tenido un amplio despliegue en las ciudades durante los últimos años corresponden a los sistemas de bicicletas compartidas(BSS). Uno de los BSS desplegados en Santiago corresponde al sistema BikeSantiago, abarcando 14 comunas con un total de 350 estaciones. Sin embargo, uno de los principales problemas que se generan en la operación diaria de este tipo de sistemas corresponde al problema de rebalanceo de bicicletas(BRP) el cual además comprende distintas configuraciones de acuerdo al tipo de red que se considere, el cual evidencia un creciente interés en la literatura durante los últimos años. En este contexto se propone un Sistema de Soporte Decisiones(DSS) compuesto por un modelo de optimización y una aplicación

A continuación, se muestra el modelo de optimización entera el cual permite re balancear la red de bicicletas compartidas.

Función Objetivo

$$Min\ CTE = \sum_{i=1}^{I}\sum_{j=1}^{J} DS_{i,j,t}SX_{i,j,t} + \sum_{j=1}^{J}\sum_{i=1}^{I} DE_{j,i,t}EX_{j,i,t} \quad (1)$$

Restricciones

$$B_{i,t} = B_{i,t-1} + \sum_{i=1}^{I}\sum_{j=1}^{J} FE_{j,i,t} - \sum_{j=1}^{J}\sum_{i=1}^{I} FS_{i,j,t} \qquad \forall\ i \in I \mid t = 1 \wedge i \neq j \quad (2)$$

$$B_{i,t} < M_{i,t} \qquad \forall\ i \in I \mid t = 1 \quad (3)$$

$$FE_{j,i,t} < B_{j,t} \qquad \forall\ j \in J \wedge i \in I \mid t = 1 \wedge i \neq j \quad (4)$$

$$FS_{i,j,t} < B_{i,t} \qquad \forall\ i \in I \wedge j \in J \mid t = 1 \wedge j \neq i \quad (5)$$

$$FE_{j,i,t} = FS_{i,j,t} \qquad \forall\ i \in I \wedge j \in J \mid t = 1 \wedge j \neq i \quad (6)$$

$$EX_{j,i,t} = \frac{FE_{j,i,t}}{H} \qquad \forall\ i \in I \wedge j \in J \mid t = 1 \wedge j \neq i \quad (7)$$

$$SX_{i,j,t} = \frac{FS_{i,j,t}}{H} \qquad \forall\ j \in J \wedge i \in I \mid t = 1 \wedge i \neq j \quad (8)$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} SX_{i,j,t} + \sum_{j=1}^{J}\sum_{i=1}^{I} EX_{j,i,t} < 2n(n-1) \qquad \mid t = 1 \quad (9)$$

$$B_{i,t} < (1+\delta)\left(\frac{M_{i,t}}{2}(1+\alpha_i)\right) \qquad \forall\ i \in I \mid t = 1 \quad (10)$$

$$B_{i,t} > (1-\delta)\left(\frac{M_{i,t}}{2}(1+\alpha_i)\right) \qquad \forall\ i \in I \mid t = 1 \quad (11)$$

$$\sum_{i=1}^{I} \alpha_i = 1 \quad (12)$$

$$SX_{j,i,t}, EX_{i,j,t}, B_{i,t}, FE_{j,i,t}, FS_{i,j,t} \in INT \quad (13)$$

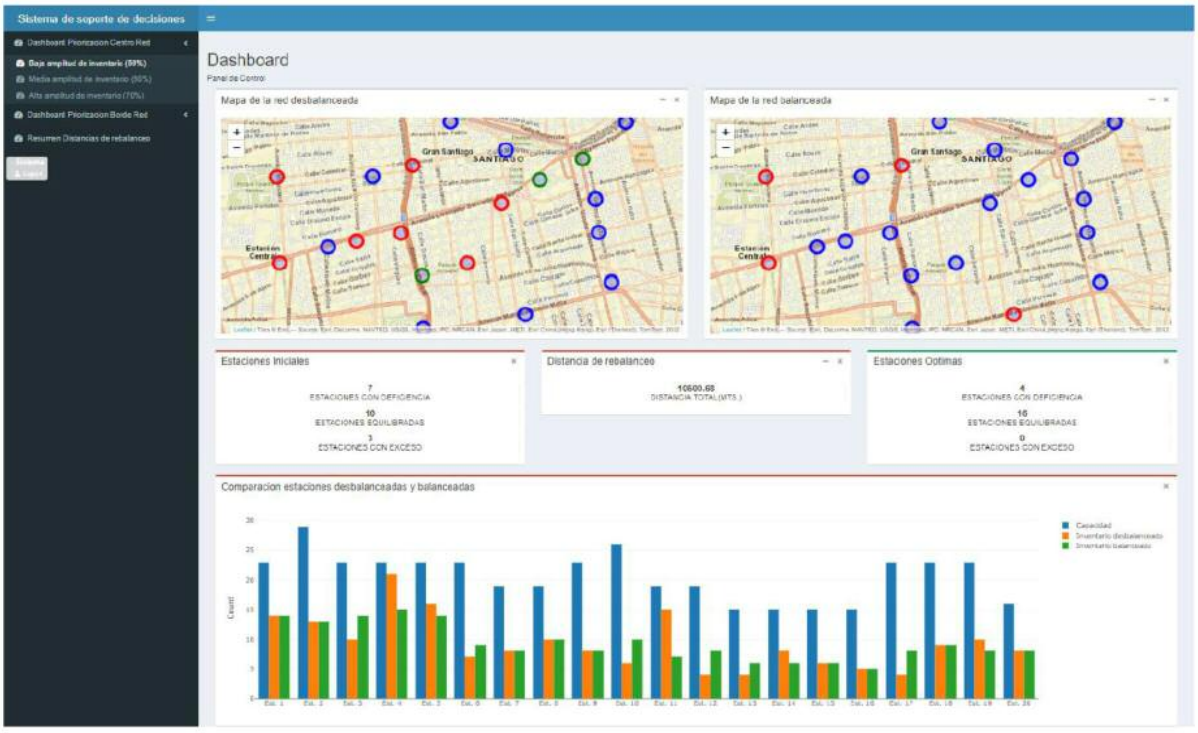


Figura 1. Dashboard del Sistema de soporte de decisiones desarrollado en Shiny

desarrollada en Shiny. Para modelar y resolver el problema BRP, se diseñó un modelo de optimización entera (MIP) y además se desarrolló un DSS en R junto a la librería Shiny. Para la implementación de este framework se contemplaron dos dimensiones de análisis: (1) Flexibilización de la capacidad y (2) Priorización de estaciones. Para la primera dimensión se aumenta la capacidad de bicicletas que se permite en cada estación y para la segunda dimensión se establecen dos tipos de priorizaciones: (1) Red con priorización central y (3) Red con priorización del borde. Dentro de los resultados relevantes, se ha determinado que a medida que disminuye la distancia total de rebalanceo aumenta la heterogeneidad de la solución de la red, disminuyendo así el nivel de servicio para los usuarios de la red, junto a esto el DSS permite visualizar las estaciones con sus capacidades rebalanceadas y la visualización de la distancia optima total de rebalanceo, lo cual comprende una métrica relevante para los tomadores de decisión.

URL: https://app-test-1.shinyapps.io/app_bike_opt/ ; Username: bike ; Password: bike

Keywords: Bike sharing,Bike rebalancing problem,Sistema de soporte de decisiones

En la siguiente imagen se muestra un menú del DSS desarrollado en Shiny.

Luis Eduardo Ferrer Cruz

Evangelización en el uso y manejo de información espacial para servicios educativos en ámbitos de contextos bilingües, en lengua originaria e interculturalidad en el Perú, utilizando R y QGIS

Desde tiempos inmemoriales, empleamos el lenguaje no solo para comunicarnos. Ha sido y es una herramienta que nos ha servido para (re)definir nuestra identidad, expresar nuestro arte o mostrar nuestra cultura a través de la historia de nuestros pueblos. La Organización de las Naciones Unidas decidió dedicar el 2019, como el Año Internacional de las Lenguas Indígenas (a fin de alentar la adopción de medidas urgentes para preservarlas, revitalizarlas y promoverlas). Las herramientas de código abierto, específicamente las referidas al tratamiento de datos espaciales, tales como R, QGIS entre otros, nos permiten identificar a nuestros cincuenta y cinco pueblos originarios en donde se hablan cuarenta y ocho lenguas originarias, cuatro andinas y cuarenta y cuatro amazónicas, que expresan la compleja diversidad que define las múltiples realidades indígenas de más de un millón doscientos mil treinta estudiantes, que reciben una Educación Intercultural Bilingüe en más de veinte seis mil escuelas, en cada uno de los ocho pisos altitudinales del Perú. Frente a ello, tenemos un enorme desafío para la pedagogía de lenguas en contextos donde no solo se aprende hablando, sino observando, escuchando y haciendo. En ese sentido, la Dirección de Educación Intercultural Bilingüe del Ministerio de Educación del Perú ha optado por desarrollar estrategias territoriales, desarrollando análisis espaciales, utilizando R, QGIS entre otros, que nos permitirán manejar situaciones como el drama de las lenguas minoritarias que hoy peligran su existencia y proponer políticas de intervención pedagógica en aquellos espacios donde su revitalización y recuperación social sea una tarea urgente, ante la pérdida de su uso en la población infantil y juvenil. Actualmente contamos con cuarenta alfabetos oficiales de lenguas originarias y dos lenguas en proceso de normalización (nanti y asheninka), así como seis lenguas por iniciar el proceso de normalización (resígaro, taushiro, muniche, omagua, iñapari, chamicuro). La evangelización en el uso y manejo de herramientas de código abierto para el uso, depuración, análisis, sistematización y desarrollo de propuestas de gestión del territorio, permitirán compartir información, no solo con los decisores del Estado, sino con nuestros mismos estudiantes, que se encuentran en la capacidad de manejar los recursos naturales de su territorio particular con estas herramientas.

¿El auge de un "precariado"?: Patrón de inversión en capital humano avanzado en Chile

Pasar de la dependencia de las materias primas a una economía basada en el conocimiento es un desafío crítico para los países en desarrollo. La creación y el fortalecimiento de los programas de educación de posgrado (programas de maestría y doctorado) es clave para la generación de conocimiento. Chile, un país cuya economía se basa principalmente en las exportaciones de materias primas, ha experimentado un fuerte crecimiento macroeconómico en las últimas dos décadas. Este progreso, sin embargo, no ha estimulado un crecimiento significativo en la inversión en investigación y desarrollo. El objetivo de este estudio fue describir los patrones de inversión en la formación de capital humano por parte del estado de Chile durante el período de 2008 a 2018. Además, evaluamos si estos patrones se reflejan en subvenciones de apoyo a la investigación por parte del estado Chileno. También evaluamos la potencial asociación entre producto interno bruto per cápita, un indicador macroeconómico, y los patrones de inversión en becas de formación de postgrado y fondos de apoyo a la investigación.

Finalmente, examinamos el nivel actual de precariedad del capital humano avanzado chileno. Aquí definimos la precariedad como un concepto político y social que describe la ausencia de empleo consistente, la inseguridad en el lugar de trabajo y la falta de recursos financieros y de bienestar. Cuatro hallazgos del estudio actual sugieren

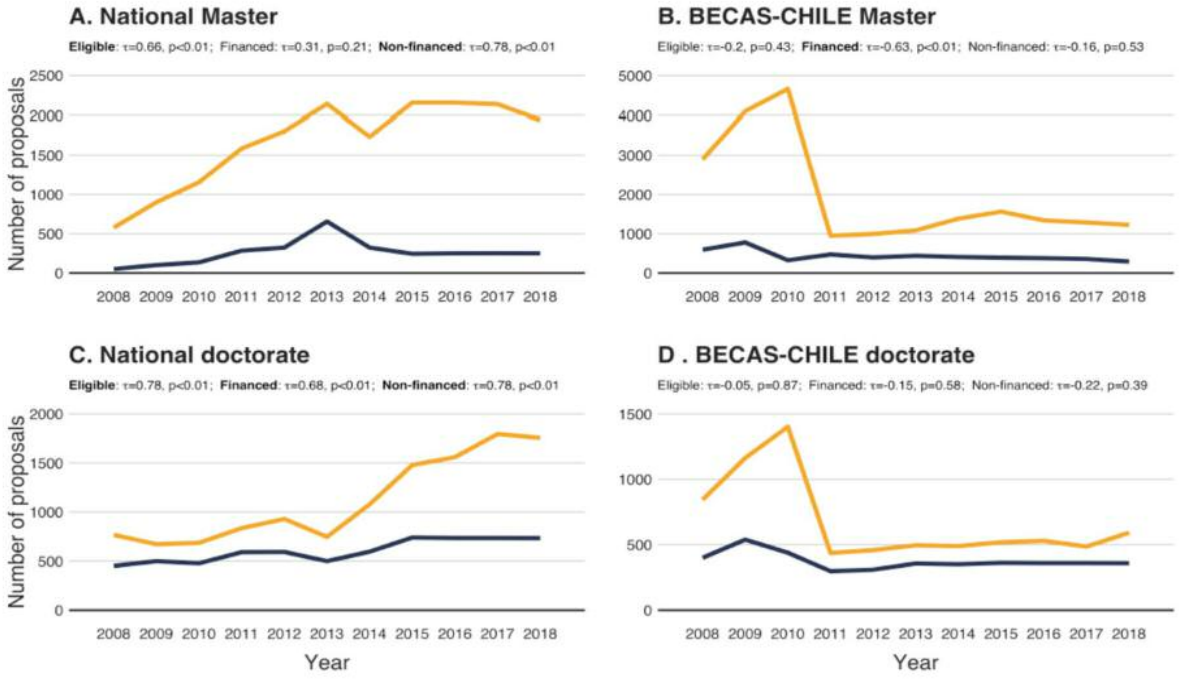


Figura 1. Tendencia anual de las propuestas a becas de formación de capital humano avanzado financiadas y no financiadas por CONICYT en el periodo 2008-2018.

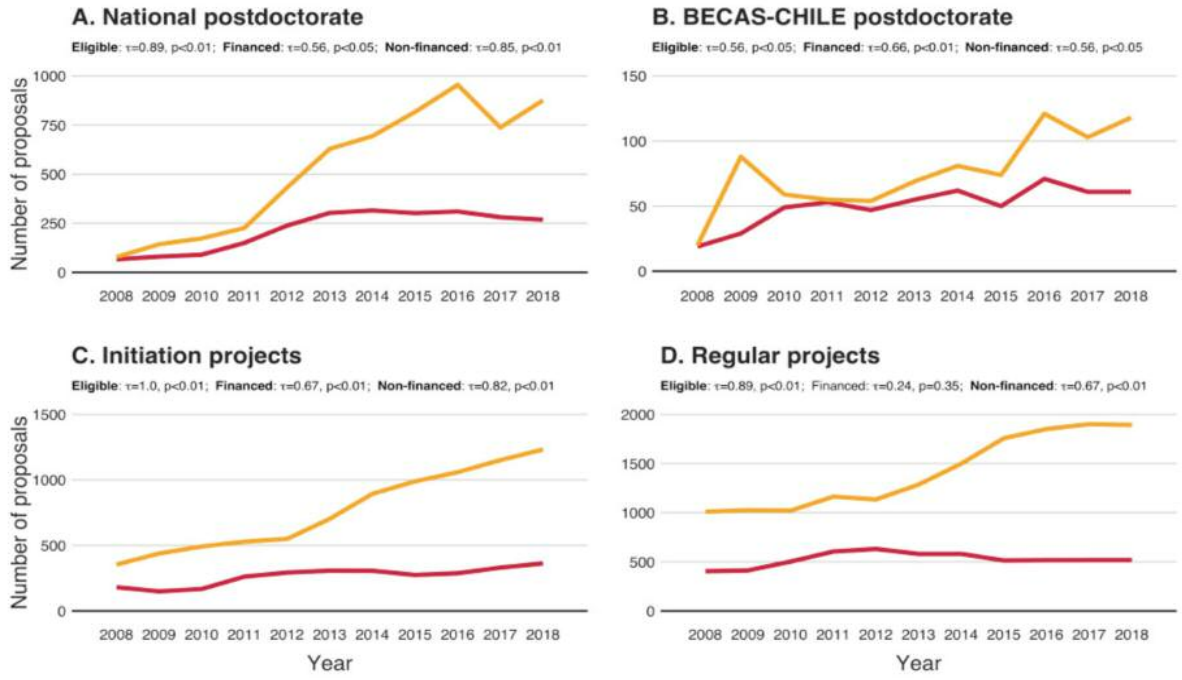


Figura 2. Tendencia anual de las propuestas a fondos de apoyo a la investigación financiadas y no financiadas por FONDECYT en el periodo 2008-2018.

que el estado chileno carece de compromiso con las inversiones avanzadas en capital humano: (1) el número de becas y fondos financiados ha disminuido durante el período de estudio, 2008-2018; (2) La tasa de éxito de las aplicaciones ha disminuido de manera constante desde su peak en 2013; (3) Se encontró una marginal relación estadística entre las inversiones en capital humano avanzado y el producto interno bruto per cápita; (4) El capital humano avanzado en Chile carece de varias formas de seguridad laboral. En general, nuestros resultados sugieren que el capital humano avanzado de Chile se encuentra en una etapa crítica en términos de inversión económica, saturación del mercado laboral y seguridad laboral. En otras palabras, el capital humano avanzado Chileno puede considerarse actualmente como un "precariado".

Matías Alejando Castillo Moine and Mónica Balzarini

Procesamiento eficiente de series de tiempo de raster espacio temporales en R

Introducción

El gran volumen de datos en forma de series de tiempo imágenes satelitales (STIS) plantea desafíos técnicos de almacenamiento y procesamiento. Numerosos paquetes de R ofrecen herramientas para su manejo con soporte para out-of-memory files y procesamiento en paralelo, permitiendo el procesamiento masivo de grandes bases de datos de STIS. Sin embargo la mayoría de los desarrollos estadísticos de R tienen base matricial, y el dato debe ser convertido para poder ser procesado; transformación que generalmente ocurre en copias completas de los datos en memoria. El objetivo de este trabajo es comparar la eficiencia de dos estrategias de manejo y procesamiento en R de datos masivos de STIS convertidas a formato matricial en memory-maped files con soporte para procesamiento en paralelo.

Materiales y métodos

Se usaron los paquetes raster (Hijmans 2019), foreach (Calaway et al. 2018), y bigmemory (Kane et al. 2018) para generar métodos de almacenamiento de datos de STIS en formatos alternativos. raster provee clases para el manejo de datos espaciales y extensiones espacio-temporales en formato imagen; bigmemory provee clases y métodos de creación, indexado y almacenamiento de grandes matrices con memoria compartida (shared memory) y archivos locales mapeados en memoria (memory-maped files); foreach provee construcciones lógicas para la ejecución de código en paralelo. En este trabajo, las STIS fueron almacenadas en objetos de la clase RasterStack del paquete raster. La conversión de datos de STIS a matriz asumió series de tiempo de igual longitud para un área definida S. La clase RasterStack de R almacena STIS en formato espacial explícito: cada capa del RasterStack es una imagen raster con idéntico número de píxeles para diferentes fechas de adquisición (Fig. 1), de longitud igual al número de capas del RasterStack. Si bien la clase RasterStack y otras extensiones permiten el indexado de series de tiempo de pixeles de una STIS, su velocidad de proceso limita su aplicación con STIS extensas. Los RasterStack fueron reacomodados en matrices de la clase big.matrix, de dimensión tantas filas como series de tiempo y tantas columnas como capas hay en el RasterStack, y almacenadas en disco (filebacked) para optimizar el uso de memoria. Para procesar en paralelo las series de tiempo de cada píxel se desarrollaron funciones que primero dividen la matriz por partes de manera secuencial, convierten cada parte a la clase matrix (en memoria) y procesa en paralelo las series de tiempo de los píxeles. El formato de almacenamiento depende del resultado,. El índice espacial fue conservado, y una función accesoria permite enmascarar píxeles (filas de la matriz) que no serán considerados en el análisis. En la Figura 2 se muestra el flujo de trabajo.

Resultados y conclusión

Las funciones desarrolladas permitieron leer en forma secuencial por partes la big.matrix y ejecutar funciones en paralelo en cada parte. Almacenar las series de tiempo de los píxeles de una STIS en una big.matrix mejoró el tiempo de procesamiento de las series de tiempo de los píxeles respecto a cuando se usó formato RasterStack. Por ejemplo el tiempo de procesamiento para filtrado de todas las series de tiempo se redujo de un día completo cuando estaban almacenadas como RasterStack a algunas horas cuando estaban almacenadas como big.matrix. Dividir la big.matrix por partes provee un marco de trabajo

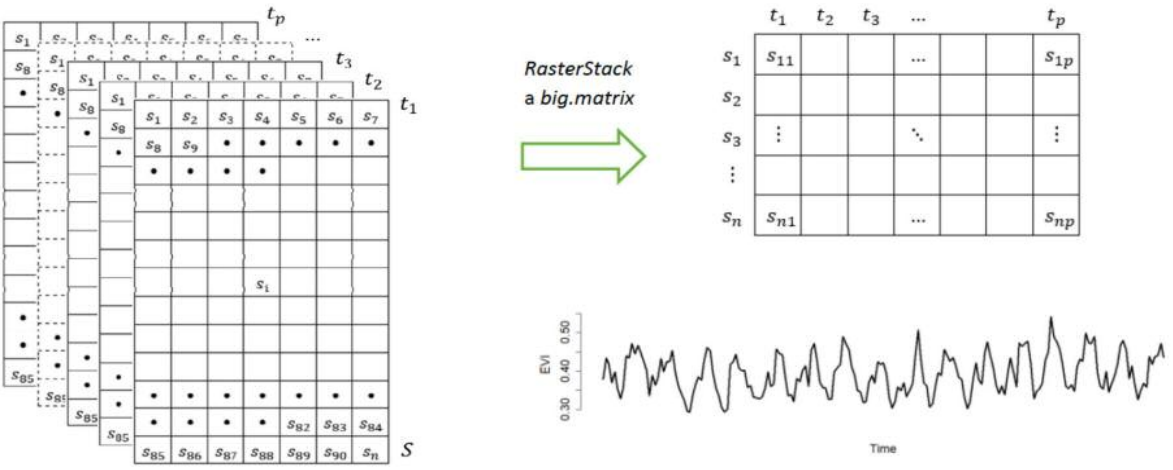


Figura 1. La región S de una imagen satelital queda representada por S={si} con i=1, …,n píxeles. En un RasterStack con imágenes para $t1,t2,…,tp$ hay n series de tiempo de longitud p y pueden ser reacomodadas en una matriz con n filas y p columnas.

```
function ( stack, mascara ) {
    ## Generar índice de píxeles no enmascarados y convertir a bigmatrix
        valid_index <- n - mascara
        matriz <- crear una filebacked big.matrix  n × p con n = número de píxeles y p = número de capas del stack
        results <- crear un objeto de formato adecuado para almacenar los resultados (por ejemplo otra big.matrix)
        for ( i in 1:nlayers ( stack ) ) {
            vectorizar cada capa del RasterStack y almacenar cada vector en columnas de la big.matrix
            }
    ## Crear un índice de partición de la matriz:
        firstRow, lastRow = crear índice ( primeras filas de partición )
    ## Inicializar cluster y procesamiento:
        cl <- abrir y registrar cluster
        for ( i in 1:firstRow ) {
            #Crear un subset de la big.matrix y transformar a clase matrix
            subset <- as.matrix ( sub.big.matrix ( matriz, firstRow, lastRow ) )
            #Procesar en paralelo
            s <- foreach (in subset) %dopar% and combine
            }
            ## exportar los resultados parciales
        }
Compilar resultados y cerrar cluster
}
```

Figura 2. Flujo de trabajo general de una función de procesamiento en paralelo para series de tiempo de imágenes satelitales. En negrita se resalta el bucle for de procesamiento de la big.matrix por partes.

general que permite procesar grandes volúmenes de series de tiempo provenientes de imágenes satelitales usando funciones de R de base matricial.

Bibliografía

- Calaway, Rich, Microsoft Corporation, Steve Weston, y Dan Tenenbaum. 2018. «Foreach Parallel Adaptor for the "parallel" Package». CRAN.
- Hijmans, Robert J. 2019. «raster: Geographic Data Analysis and Modeling». https://cran.r-project.org/package=raster.
- Kane, Michael J., John W. Emerson, Peter Jr. Haverty, y Charles Determan. 2018. «bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped Files». https://cran.r-project.org/web/packages/bigmemory/index.html.

El imaginario social asociado con la Feria Nacional de San Marcos en Aguascalientes

La realización de festivales, eventos y ferias conlleva diversos beneficios, entre ellos permitir a las ciudades sede ofertar una imagen positiva del destino, atraer turistas y provocar la atención de los medios de comunicación (Sola, 1998; Atkinson et al 2008); en su mayoría estos eventos son promovidos por el gobierno local y atraen tanto a turistas nacionales como a extranjeros.

En México, la Feria de San Marcos, celebrada en el estado de Aguascalientes, inicia por vez primera en 1828 y año con año continúa siendo un referente en áreas culturales y de entretenimiento. Durante el periodo de feria varias actividades, como las académicas, se suspenden con la finalidad de que los locales asistan a los eventos ofertados; Vila, Koster y Marin (2016) identificaron, a través de encuestas in situ, una relación significativa entre experiencias emocionales en la feria de san marcos y la actitud hacia la marca-ciudad de Aguascalientes.

Los objetivos de esta investigación son: Identificar el uso de redes sociales durante la feria; Catalogar la experiencia turística ofertada (Al analizar el timeline de @FNSM_Oficial, la página de twitter del patronato de la feria) y la experiencia percibida en la feria de san marcos 2019 (al analizar los tweets relacionados con palabras clave: "Feria Nacional de San Marcos", "Aguascalientes AND Feria", "FNSM").

El explorar el imaginario social a través de las experiencias turísticas ofertadas y percibidas, en torno a la feria puede contribuir a la toma de decisiones relacionadas con la promoción de la misma.

Método: Análisis textual de tweets recolectados del 21 abril al 12 de mayo 2019 (Fechas en las que se lleva a cabo la feria) con apoyo del paquete rtweet, tidyverse (ggplot2 y tokenizer)

Ignacio Castro, Gonzalo Nuñez and María Simian

Desarrollo e implementación de un Observatorio de Susceptibilidad Antimicrobiana en un hospital de alta complejidad en Chile

Introducción

La resistencia antimicrobiana es un importante problema de salud pública. En el mundo se le atribuyen 700.000 muertes por año y se estima que para el año 2050 causará 10 millones de muertes por un costo global acumulado de 100 trillones de dólares (Lesho y Laguio-Vila 2019). Se han identificado distintos factores que influyen en el desarrollo de la resistencia, sin embargo, el principal es el uso indiscriminado de antibióticos, tanto en las atenciones médicas como en la industria alimentaria (Byarugaba 2004). Es por esto que la restricción del uso de antimicrobianos mediante la utilización de protocolos locales de prescripción es una herramienta fundamental para el control de la resistencia (Rice 2018). Para llevar a cabo estos protocolos es necesario conocer la microbiología local: los microorganismo más prevalentes y su respectivo perfil de susceptibilidad antimicrobiana (Masterton 2008).

En Chile, año a año, un grupo de expertos con representación nacional, se reúnen en el mes de abril para consolidar y publicar la susceptibilidad antimicrobiana de las cepas de 40 hospitales aisladas el año anterior en una revista científica de circulación nacional. Sin duda alguna, ha sido un gran aporte para el desarrollo de guías clínicas de los distintos hospitales. Sin embargo, la periodicidad, la forma de recolección y consolidación de datos utilizada ofrece una oportunidad de mejora. La mayoría de los hospitales maneja una base de datos en servidores locales ofrecidos por la misma empresa, donde se respaldan todos los análisis de susceptibilidad realizados en cada centro asistencial.

Objetivos

Es por esto que el objetivo de este trabajo fue optimizar el proceso de recolección y consolidación de los estudios de susceptibilidad antimicrobiana y ofrecer una alternativa de publicación de dicha información.

Métodos

Para esto se realizó un piloto en un hospital de alta complejidad de la ciudad de Santiago de Chile. Se programó un análisis descriptivo utilizando R y se confeccionó un dashboard con Shiny. El producto obtenido se puso a disposición de manera abierta en la web con los estudios de susceptibilidad realizados entre el 01 de enero del 2016 y el 31 de diciembre de 2018. Los estudios de susceptibilidad pueden ser segmentados por el tipo de servicio donde se tomó la muestra (Unidad de Paciente Crítico o no), tipo de muestra y por periodo de tiempo. Por otra parte, se publicaron los microorganismos aislados con mayor frecuencia por tipo de unidad y por tipo de muestra. Todos los análisis fueron realizados siguiendo las directrices del capítulo de Microbiología de la Sociedad Chilena de Infectología (Sociedad Chilena de Infectología 2010).

Resultados

Se analizaron inicialmente 481.390 estudios de susceptibilidad que se consolidaron, analizaron y publicaron en: https://ochisam.shinyapps.io/OCHISA/ Esta herramienta ha sido bien valorada, tanto por los médicos infectólogos (clínicos) como por los microbiólogos (paraclínicos).

Proyecciones

Luego de la puesta en marcha del piloto en dicho hospital, esperamos poder sumar paulatinamente a los distintos hospitales del territorio nacional y programar la actualización permanen-

te de la base de datos. Esto nos permitiría procesar y publicar, al menos con periodicidad mensual, la susceptibilidad antimicrobiana a nivel nacional y segmentada por establecimiento. Esto significa un ahorro sustantivo de recursos y permite diseñar con mayor precisión las guías de tratamiento antibiótico empíricos.

Referencias

- Byarugaba, D. K. (2004). Antimicrobial resistance in developing countries and responsible risk factors.International journal of antimicrobial agents , 24 (2), 105-110.
- Lesho, E. P., y Laguio-Vila, M. (2019, March). The Slow-Motion Catastrophe of Antimicrobial Resistance and Practical Interventions for All Prescribers. In Mayo Clinic Proceedings . Elsevier.
- Masterton, R. (2008). The importance and future of antimicrobial surveillance studies. Clinical infectious diseases , 47 (Supplement_1), S21-S31.
- Sociedad Chilena de Infectología (2010). Recomendaciones para el análisis de datos acumulados de susceptibilidad antimicrobiana en instituciones de salud. Rev Chil Infect , 27 (2), 126-132.
- Rice, L. B. (2018). Antimicrobial Stewardship and Antimicrobial Resistance. Medical Clinics , 102 (5), 805-818.